Interestingness Measures for Association Rules in a KDD Process : PostProcessing of Rules with ARQAT Tool

Xuan-Hiep Huynh

HAL Id: tel-00482649

https://tel.archives-ouvertes.fr/tel-00482649

Submitted on 11 May 2010

UNIVERSITÉ DE NANTES

ÉCOLE DOCTORALE
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DES MATÉRIAUX

Année : 2006

Thèse de Doctorat de l'Université de Nantes

Spécialité : INFORMATIQUE

Présentée et soutenue publiquement par

Xuan-Hiep HUYNH

le 07 décembre 2006
à l'École polytechnique de l'Université de Nantes

INTERESTINGNESS MEASURES FOR ASSOCIATION RULES IN A KDD PROCESS: POSTPROCESSING OF RULES WITH ARQAT TOOL

	Jury	
Président :	Gilles VENTURINI	Professeur à l'École polytechnique de l'Université de Tours
Rapporteurs :	Ludovic LEBART	Directeur de recherche CNRS à l'ENST
	Gilbert RITSCHARD	Professeur à l'Université de Genève
Examinateurs :	Henri BRIAND	Professeur à l'École polytechnique de l'Université de Nantes
	Fabrice GUILLET	MCF à l'École polytechnique de l'Université de Nantes
	Pascale KUNTZ	Professeur à l'École polytechnique de l'Université de Nantes

Directeur de thèse : Henri BRIAND
Co-encadrant de thèse : Fabrice GUILLET
Laboratoire : Laboratoire d'Informatique de Nantes Atlantique (LINA)
Composante de rattachement du directeur de thèse : Polytech'Nantes - LINA

N° ED 0366-279

To my wife and children.

Abstract

This work takes place in the framework of Knowledge Discovery in Databases (KDD), often called "Data Mining". This domain is both a main research topic and an application field in companies. KDD aims at discovering previously unknown and useful knowledge in large databases. In the last decade many researches have been published about association rules, which are frequently used in data mining. Association rules, which are implicative tendencies in data, have the advantage to be an unsupervised model. But, in counter part, they often deliver a large number of rules. As a consequence, a postprocessing task is required by the user to help him understand the results. One way to reduce the number of rules - to validate or to select the most interesting ones - is to use interestingness measures adapted to both his/her goals and the dataset studied. Selecting the right interestingness measures is an open problem in KDD. A lot of measures have been proposed to extract the knowledge from large databases and many authors have introduced the interestingness properties for selecting a suitable measure for a given application. Some measures are adequate for some applications but the others are not.

In our thesis, we propose to study the set of interestingness measure available in the literature, in order to evaluate their behavior according to the nature of data and the preferences of the user. The final objective is to guide the user's choice towards the measures best adapted to its needs and *in fine* to select the most interesting rules.

For this purpose, we propose a new approach implemented in a new tool, ARQAT (Association Rule Quality Analysis Tool), in order to facilitate the analysis of the behavior about 40 interestingness measures. In addition to elementary statistics, the tool allows a thorough analysis of the correlations between measures using correlation graphs based on the coefficients suggested by Pearson, Spearman and Kendall. These graphs are also used to identify the clusters of similar measures.

Moreover, we proposed a series of comparative studies on the correlations between interestingness measures on several datasets. We discovered a set of correlations not very sensitive to the nature of the data used, and which we called stable correlations.

Finally, 14 graphical and complementary views structured on 5 levels of analysis: ruleset analysis, correlation and clustering analysis, most interesting rules analysis, sensitivity analysis, and comparative analysis are illustrated in order to show the interest of both the exploratory approach and the use of complementary views.

Keywords: Knowledge Discovery in Databases (KDD), interestingness measures, postprocessing of association rules, clustering, correlation graph, stability analysis.

Acknowledgements

I would like to thank Mr Henri Briand, professor at Nantes university, my supervisor, for his great encouragement, many suggestions and willing support during this research. I am strongly thankful to Mr Fabrice Guillet, my co-supervisor, for his support, unique enthusiasm, unending optimism, and valuable guidance from the beginning of my thesis.

I am grateful to Mr Tran Thuong Tuan, Mr Vo Van Chin, Mr Le Quyet Thang, Mr Ho Tuong Vinh, and Mr Gilles Richard who helped me to follow these studies of doctorate.

The warmest and most sincere thanks go to Mr Régis Gras, Mr Ivan Kodjadinovic for their advices during these years. Particularly, I owe my gratitude to Mrs Catherine Galais for her corrections of my articles in English.

I thank Mr Ludovic Lebart and Mr Gilbert Ritschard, who have agreed to review my work. I also thank Mrs Pascale Kuntz for being part of the jury. I specially thank Mr Gilles Venturini, the president of the jury.

I want to acknowledge Mrs Jacqueline Boniffay, Mrs Martine Levesque, Mr Yvan Lemeur, Ms Stéphanie Legeay, Ms Sandrine Laillet, Ms Carolie Robert, Ms Leila Necibi, Ms Christine Brunet, Ms Corinne Lorentz, CROUS of Nantes, BU technologies, Informatics service of Polytech'Nantes for their helps.

I should also mention that my Ph.D studies in France were supported by the Vietnamese and French governments. I thank warmly to Ms Nguyen Thi Mai Huong (the French Consulate in Ho Chi Minh city) and Ms Hoang Kim Oanh (the Project 322 of the Vietnamese government).

I wish to thank all my good friends Thanh-Nghi Do, Julien Lorec, Jérôme Mikolajczak, Quang-Khai Pham, Julien Blanchard, Jérôme David, Nicolas Beaume, Emmanuel Blanchard, Bruno Pinaud.

I owe my gratitude to Mrs Vo Le Ngoc Huong for her advice and help to me as well as the other Vietnamese students in Nantes.

Of course, I am grateful to my family, especially my wife Nguyen Thi Anh Thu, my two children Huynh Xuan Anh Tram and Huynh Quoc Dung, my parents, and my parents-in-law, for their patience, confidence and love.

Nantes, France
December 07, 2006

Huynh Xuan Hiep

Table of Contents

Abstract **iii**

Acknowledgements **iv**

Table of Contents **v**

1 Introduction **1**
- 1.1 The KDD process 1
  - 1.1.1 KDD system 2
  - 1.1.2 Who is the user? 2
  - 1.1.3 Postprocessing of association rules 2
- 1.2 Postprocessing models 5
- 1.3 Clustering techniques 5
  - 1.3.1 Partitioning clustering 6
  - 1.3.2 Hierarchical clustering 6
- 1.4 Thesis contribution 6
- 1.5 Thesis structure 7

2 Discovery of association rules **9**
- 2.1 From the data 9
- 2.2 ... To association approach 10
  - 2.2.1 Monotonicity 11
  - 2.2.2 Frequent itemsets 11
  - 2.2.3 Rule generation 12
- 2.3 Algorithm improvement 13
  - 2.3.1 Database pass 13
  - 2.3.2 Computation 15
  - 2.3.3 Frequent itemset compaction 16
  - 2.3.4 Search space 16
  - 2.3.5 Data type 18
- 2.4 Summary 21

3 Measures of interestingness **22**
- 3.1 Subjective measures 22
  - 3.1.1 Actionability 22

3.1.2 Unexpectedness 23
3.2 Objective measures 23
3.2.1 Criteria 24
3.2.2 Classification 26
3.2.3 Mathematical relation 26
3.3 Summary 27

4 Postprocessing approaches 31
4.1 Constraints 31
4.1.1 Boolean expression constraints 31
4.1.2 Constrained association query 32
4.1.3 Minimum improvement constraints 32
4.1.4 Action hierarchy 33
4.1.5 AND-OR taxonomy 33
4.1.6 Minimal set of unexpected patterns 33
4.1.7 Interestingness preprocessing 34
4.1.8 Negative association rules 35
4.1.9 Substitution 35
4.1.10 Cyclic/Calendric 35
4.1.11 Non-actionable 36
4.1.12 Softness 36
4.2 Pruning 36
4.2.1 Rule cover 36
4.2.2 Maximum entropy 37
4.2.3 Relative interestingness 37
4.2.4 Interestingness-based filtering 38
4.2.5 Ranking subset of contingency table 38
4.2.6 Rule template 38
4.2.7 Uninteresting rules 38
4.2.8 Template-based rule filtering 39
4.2.9 Merging 40
4.2.10 Correlation between objective measures and real human interest 40
4.3 Summarization 41
4.3.1 Direction setting 41
4.3.2 GSE pattern 41
4.4 Grouping 42
4.4.1 Differentiating with support values 42
4.4.2 Neighborhood 42
4.4.3 Triple factors 43
4.4.4 Average distance 44
4.4.5 Attribute hierarchy 44
4.4.6 Probability distance 45
4.5 Visualizing 45
4.5.1 Rule visualizer 45
4.5.2 Human-centered rummaging 45
4.6 Representative measures 45

4.6.1 Dissimilarity between measures 46
4.6.2 AHC 48
4.6.3 PAM 48
4.6.4 Correlation graph 49
4.6.5 Correlated versus uncorrelated graphs 51
4.7 Summary 52

5 ARQAT tool **53**
5.1 Principles 53
5.2 Modules 54
5.3 Preprocessing 55
5.3.1 Filtering 55
5.3.2 Interestingness calculation 56
5.3.3 Sampling 57
5.4 Evaluation 57
5.4.1 Basic statistics 57
5.4.2 Correlation analysis 61
5.4.3 Comparative study 68
5.5 Display 71
5.6 Utility 72
5.7 Summary 72

6 Studies: on general evaluation **74**
6.1 Datasets 74
6.2 Rulesets 75
6.3 Used measures 76
6.4 Efficiency of the sample model 76
6.5 Distribution of interestingness values 77
6.6 Joint-distribution matrix 78
6.7 Correlation analysis 79
6.8 Interesting rule analysis 80
6.9 Ranking measures by sensitivity values 81
6.9.1 On a ruleset 81
6.9.2 On a set of rulesets 82
6.10 Summary 82

7 Comparative studies **85**
7.1 Graph of stable clusters 85
7.2 Comparative analysis: Pearson's coefficient 86
7.2.1 Discussion 86
7.2.2 $CG+$ and $CG0$ 86
7.2.3 $\overline{CG0}$ graphs: uncorrelated stability 86
7.2.4 $\overline{CG+}$ graph: correlated stability 87
7.3 Comparative analysis: Spearman's coefficient 89
7.3.1 Discussion 89
7.3.2 $CG+$ and $CG0$ 89
7.3.3 $\overline{CG0}$ graphs: uncorrelated stability 91

7.3.4 $\overline{CG+}$ graph: correlated stability . . . 91

7.4 Comparative analysis: Kendall's coefficient . . . 92

7.4.1 Discussion . . . 93

7.4.2 $CG+$ and $CG0$ . . . 94

7.4.3 $\overline{CG0}$ graphs: uncorrelated stability . . . 94

7.4.4 $\overline{CG+}$ graph: correlated stability . . . 95

7.5 Comparative analysis: all coefficients . . . 95

7.5.1 Discussion . . . 96

7.5.2 $\overline{CG0}$ graphs: uncorrelated stability . . . 96

7.5.3 $\overline{CG+}$ graph: correlated stability . . . 96

7.6 Comparative analysis: with AHC and PAM . . . 97

7.6.1 On a ruleset . . . 97

7.6.2 On two opposite rulesets . . . 98

7.7 Summary . . . 100

8 Conclusions and perspectives 108

8.1 Main results . . . 108

8.1.1 ARQAT tool . . . 109

8.1.2 Correlation graph . . . 110

8.1.3 Representative measure . . . 110

8.1.4 Comparison study . . . 110

8.1.5 Cluster evaluation . . . 111

8.2 Perspectives . . . 111

8.2.1 Improving the sample model . . . 111

8.2.2 Improving the cluster evaluation . . . 111

8.2.3 Hierarchy view . . . 112

8.2.4 Aggregation technique . . . 112

8.3 Publication . . . 112

Bibliography 114

List of Tables

2.1 A normalized database of sale transactions. 9

2.2 Binary representation of transactions. 10

2.3 Ordering items with support count (*minsup* = 2). 17

2.4 Mining frequent itemsets with conditional FP-tree. 18

3.1 A list of objective measures. 28

3.2 Some matched properties of objective measures (IND.: Independency, EQU.: Equilibrium, SYM.: Symmetry, VAR.: Variation, DES.: Descriptive, STA.: Statistical. < ● >: matched, < ○ >: unmatched). 29

3.3 A classification of objective measures. 30

3.4 Some relations between objective measures 30

4.1 Maximum entropy induced. 37

4.2 Rule structure for exceptions. 37

4.3 Rule classifications with a subjective measure. 40

4.4 An example of average ranking. 41

4.5 Dissimilarity values for three interestingness measures and five association rules. . . 48

4.6 Clusters of the measures obtained with PAM. 51

4.7 The supplementary information obtained on the clusters of a ruleset. 52

4.8 Correlation values for three measures and five association rules. 52

5.1 Characteristic types (remind that $n_{XY} = n_X - n_{X\overline{Y}}$). 58

5.2 Frequency and inverse-cumulative in bins. 59

5.3 Some statistical indicators on a measure m. 59

5.4 Sensitivity structure. 60

5.5 Average structure to evaluate sensitivity on a set of rulesets. 60

5.6 Matrix of correlation values. 62

5.7 Pearson's values for three measures and five association rules. 62

5.8 Spearman's values for three measures and five rules. 63
5.9 Kendall's values for three measures and five rules. 64
5.10 An example of concordant and discordant between the pairs of interestingness values computed from two objective measures m_1 and m_2 ($< \bullet >$: concordant, $< \circ >$: discordant). 64

6.1 Dataset description. 74
6.2 Rulesets extracted. 75
6.3 Trends on interestingness values 78

7.1 Comparison of correlations with Pearson's coefficient. 86
7.2 Comparison of correlations with Spearman's coefficient. 92
7.3 Comparison of correlations with Kendall's coefficient. 98
7.4 Comparison of correlation with all coefficients on the $\mathcal{R}_1$ ruleset. 101
7.5 Clusters of measures with AHC on $\mathcal{R}_1$ (dissimilarity interval = 0.15). 102
7.6 Clusters of measures with PAM on the $\mathcal{R}_1$ ruleset. 104
7.7 Clusters agreed with both AHC and PAM on the $\mathcal{R}_1$ ruleset. 105
7.8 Clusters of measures with AHC (measures are represented by their orders). 105
7.9 Cluster comparison from AHC (measures are represented by their orders). 106
7.10 Clusters of measures with PAM (measures are represented by their orders). 106
7.11 Cluster comparison from PAM (measures are represented by their orders). 107
7.12 Clusters agreed with both AHC and PAM on the $\mathcal{R}_1$, $\mathcal{R}'_1$, $\mathcal{R}_2$, and $\mathcal{R}'_2$ rulesets. . . 107

List of Figures

1.1 The steps in a KDD process. 2

1.2 Postprocessing models. 5

2.1 Lattice structure. 11

2.2 An example of finding all frequent itemsets with three database passes. 13

2.3 An FP-tree constituted from a set of transactions. 17

2.4 Lexicographic tree. 18

2.5 P-tree in the form of Rymon's set enumeration tree. 19

2.6 Quantitative and categorical items. 19

2.7 Mapping to the boolean problem. 20

2.8 A taxonomy of items. 20

2.9 Adding taxonomy items to find frequent itemsets. 20

3.1 The cardinalities of a rule $X \to Y$ and the "surprisingness" of negative examples. . . 23

4.1 Rymon's set-enumeration tree for $U = \{A, B, C, D\}$. 32

4.2 Fragment of an action tree for the supermarket management. 33

4.3 An AND-OR taxonomy. 34

4.4 IPS in the interestingness process. 34

4.5 Softness interval. 36

4.6 Ranking subset of contingency tables. 39

4.7 The profile building process. 39

4.8 A fragment of the template specification language. 40

4.9 The leaf level of the M-Tree. 41

4.10 The process of finding DS rules. 42

4.11 Interesting rules with r-neighborhood. 43

4.12 Overlapping of an attribute. 44

4.13 An attribute hierarchy. 45

4.14 RuleVisualizer prototype. 46
4.15 The neighborhood relation for rummaging. 47
4.16 Visualization with the ARVis platform. 47
4.17 Clusters of measures with AHC on a ruleset. 49
4.18 Measures and clusters projected on the two principal axes of a PCA. The measures are presented by the symbols and the clusters are represented by the numbers. . . . 50
4.19 An illustration of the correlation graph. 51

5.1 ARQAT structure. 54
5.2 ARQAT modules. 55
5.3 Preprocessing of rulesets. 55
5.4 The filtering module. 56
5.5 The calculation module. 56
5.6 The sampling module. 57
5.7 Evaluating some characteristics of a ruleset. 58
5.8 Frequency and inverse-cumulative histograms of the Lift measure. 59
5.9 Drawing scatterplot view. 61
5.10 An example of two scatterplots on the ruleset Mushroom. 61
5.11 Computing correlation values. 62
5.12 An example of a line chart on Mushroom ruleset. 65
5.13 Summary scheme. 65
5.14 A summary structure (P: Pearson, S: Spearman, K: Kendall). 65
5.15 Stable combinations with correlation values (P: Pearson, S: Spearman, K: Kendall). 66
5.16 Summary scheme extracted from a ruleset (A lighter color box shows a τ-cluster, the otherwise a black box is a θ-cluster). 66
5.17 Clustering measures. 67
5.18 Combination of cluster type for a cell in a matrix ([**b**] and [**c**] are only visualized for τ-cluster but it is the same for θ-cluster). [**a**] is visualized for the matrix with only one coefficient such as P, S or K. [**b**] is visualized for the matrix with two coefficients PS, PK or SK. [**c**] is visualized for the matrix with three coefficients PSK. 67
5.19 A process to compute a sample. 68
5.20 A stable hierarchy on a set of rulesets. 69
5.21 A summary computed from all the rulesets (original and sample). 69
5.22 Union of the ten most interesting rules from all the measures on the LBD ruleset. . . 71
5.23 ARQAT Display. 71
5.24 Principal menu (extracted). 72

5.25 A set of complementary views on a cluster. 73

6.1 Characteristics of the rulesets. 76
6.2 Characteristics of the rulesets (cont.). 77
6.3 Variation of the measure IPEE on the $\mathcal{R}_1$ ruleset. 77
6.4 Distribution of some measures on the $\mathcal{R}_1$ ruleset (extracted). 78
6.5 Some agreement and disagreement shapes. 79
6.6 Scatterplot matrix of joint-distributions on the $\mathcal{R}_1$ ruleset (extracted). 80
6.7 Coefficient matrix on the $\mathcal{R}_1$ ruleset. 81
6.8 CG0 and CG+ graphs on the $\mathcal{R}_1$ ruleset (clusters are highlighted with a gray background). 82
6.9 Union of the ten most interesting rules for each measure of cluster $C1$ on the $\mathcal{R}_1$ ruleset (extracted). 83
6.10 Plot of the union of the ten most interesting rules for each measure of cluster $C1$ on the $\mathcal{R}_1$ ruleset. 83
6.11 Sensitivity rank on the $\mathcal{R}_1$ ruleset (extracted). 84
6.12 Comparison of sensitivity values on a couple of measures of the $\mathcal{R}_1$ ruleset. 84
6.13 Sensitivity rank on all the original rulesets (extracted). 84

7.1 τ-cluster on $\mathcal{R}_1$ with Pearson's coefficient. 87
7.2 CG+ graphs with Pearson's coefficient. 88
7.3 CG+ graphs with Pearson's coefficient (cont.). 89
7.4 CG0 graphs with Pearson's coefficient. 90
7.5 CG0 graphs with Pearson's coefficient (cont.). 91
7.6 $\overline{CG+}$ graph with Pearson's coefficient. 92
7.7 τ-cluster on $\mathcal{R}_1$ with Spearman's coefficient. 93
7.8 CG+ graphs with Spearman's coefficient. 94
7.9 CG+ graphs with Spearman's coefficient (cont.). 95
7.10 CG0 graphs with Spearman's coefficient. 96
7.11 CG0 graphs with Spearman's coefficient (cont.). 97
7.12 $\overline{CG+}$ graph with Spearman's coefficient. 98
7.13 τ-cluster on $\mathcal{R}_1$ with Kendall's coefficient. 99
7.14 CG+ graphs with Kendall's coefficient. 100
7.15 CG+ graphs with Kendall's coefficient (cont.). 101
7.16 CG0 graphs with Kendall's coefficient. 102
7.17 CG0 graphs with Kendall's coefficient (cont.). 103

7.18 $\overline{CG+}$ graph with Kendall's coefficient. 103

7.19 $\overline{CG+}$ graphs with PS, PK, SK and PSK summaries on $\mathcal{R}_1$. 104

Chapter 1

Introduction

This work takes place in the framework of knowledge discovery in databases (KDD), often called data mining. In the context of KDD, the extraction of rules in forms of association rules is a technique that is used frequently. The technique has the advantages to offer a simple model and unsupervised algorithms but it delivers a huge number of rules. So it is necessary to implement a postprocessing step to help the user to reduce the number of rules discovered.

One of the most attractive area in the knowledge discovery research is the automatic analysis of changes and deviation [PSM94] or the development of good measures of interestingness of discovered patterns [ST96]. As the number of discovered rules increases, end-users, such as data analysts and decision makers, are frequently confronted with a major challenge: how to validate and select the most interesting ones of those rules. More precisely, we are interested in the postprocessing of association rules with a set of interestingness measures. The main goal of our work relies on studying the behavior – theoretical and experiment – of interestingness measures for association rules.

1.1 The KDD process

Knowledge Discovery in Databases (KDD) is a research framework first introduced by Frawley *et al.* [FPSM91]. A general definition of KDD is given in [FPSS96]: "KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data". The KDD process is well examined in the literature [BA96] [FPSS96] [HMS01]. Interaction and iteration in many steps with user's decisions are the principal features of the process. The KDD process is illustrated in Fig. 1.1 [FPSS96]. The user interacts with the process by making decisions. The process operates on the following basic steps: (i) identifying the goal from the user's point of view – based on the relevant knowledge about the domain –, (ii) creating a target data, (iii) data preprocessing, (iv) data reduction and projection, (v) matching the goals of the KDD process, (vi) exploratory analysis, (vii) data mining, (viii) interpreting mined patterns, (ix) acting on the discovered knowledge. These steps can be divided into three tasks: the **preprocessing** of data (steps i → vi), the **mining** of data (steps vii) and the **postprocessing** of data (steps viii → ix). The principal notions of the KDD process can be found in [KZ96].

The domain knowledge or background knowledge is the supplementary knowledge on the form, the data content, the data domain, the special situations and the process goal. The domain knowledge helps the process to focus on the research content. Most of the knowledge comes from the domain expert. The basic knowledge contains the information on the knowledge that is already

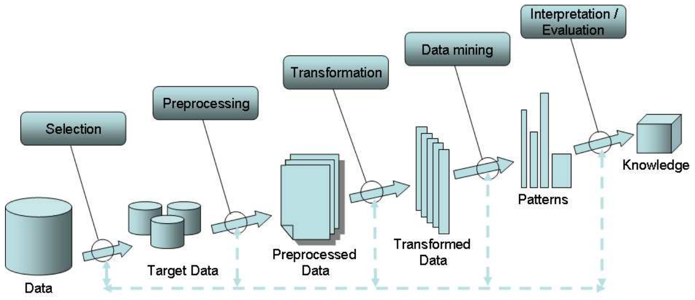


Figure 1.1: The steps in a KDD process.

saved in the system. A dictionary[1], a taxonomy, or a constraint of domain knowledge is an example of this kind of knowledge [KZ02]. The structures such as decision tree or rules are used frequently. The procedures to extract the knowledge from data are called the discovery algorithms. Contingency tables, subgroup patterns, rules, decision trees, functional relations, clusters, taxonomies and concept hierarchies, probabilistic, causal networks, and neural networks are some forms of knowledge (i.e. pattern) [KZ02].

1.1.1 KDD system

A system implementing the steps of the KDD process is considered as a Knowledge Discovery System (KDS). It finds the knowledge that it has not before implicitly in its algorithms or explicitly in its domain knowledge. In this case, a knowledge (i.e. interesting and utility) is a relation or a pattern in the data [FPSM91]. A KDS includes a collection of components to identify or to extract the new patterns from the real data [MCPS93]. Interest and utility are considered as two important aspects.

The components in a KDS can differ from each others but we can determine some principle functions such as the control, the data interface, the focus, the pattern extraction, the evaluation and the knowledge base. The *control function* manages the user's demands and the parameters of other components. The *data interface* creates the questions on the data and then treats them. The *focus* determines what portion of the data to analyze. The *pattern extraction* collects the algorithms to extract the patterns. The *evaluation* evaluates the interests and utilities of extracted patterns. The *knowledge base* stores the specific information about the domain.

1.1.2 Who is the user?

Application is the final goal of any KDS. This application is used in a society, a company, etc. to supply the recommended analysis and actions. Its users can be a commercial person to see the important events in the business [BA96] (e.g. a decider, a decision-maker, a manager). A data analyst searches exploratory the basic tendencies or the patterns in a domain can become a user.

1.1.3 Postprocessing of association rules

In the framework of data mining [FPSS96], association rules [AIS93a] are a key tool aiming at discovering interesting implicative patterns in data. Since the precursor works about finding interesting patterns [PS91] [AIS93b] [AS94] in knowledge discovery in databases, the validation of association rules [AIS93a] [AMS$^+$96] still remains a research challenge.

[1] A dictionary contains the relations between items and values.

Association rules

Association rule is one of the most important forms to represent the discovered knowledge in the mining task [AIS93a] [AS94] [AMS$^+$96]. The first representation is of the form $X_1 \wedge X_2 \wedge ...X_k \rightarrow Y_1$ in which the antecedent is composed of many elements and the consequence is of only one element. Each element in the antecedent or consequent represents an item in a dataset. The transactions containing the bought items of the clients (e.g. in a supermarket) are often used as the datasets to analyze. The standard form $X_1 \wedge X_2 \wedge ... \wedge X_k \rightarrow Y_1 \wedge Y_2 \wedge ... \wedge Y_l$ is well developed with the APRIORI algorithm [AS94] [AMS$^+$96]. In this form, both of the two parts of a rule (i.e. the antecedent and the consequence) are composed with many items (i.e. a set of items or *itemset*), taking an important role in KDD.

Measures of interestingness

To evaluate the patterns issued from the second task (the mining task) [Sec. 1.1], the notion of interestingness is introduced [PSM94]. The patterns are transformed in value by interestingness measures. The interestingness value of a pattern can be determined explicitly or implicitly in a KDS. The patterns may have different ranks because their ranks depend strongly on the choice of interestingness measure. The interestingness measures are classified into two categories [ST96]: subjective measures and objective measures. The subjective measures rely on the goals, the knowledge, and the belief of the user [PT98] [LHML99]. The objective measures are statistical indexes [AIS93a] [BA99] [BGGB05a] [BGGB05b]. The properties of the objective measures are briefly studied in [PS91] [MM95] [BA99] [HH01] [TKS04] [GCB$^+$04] [Fre99] [Gui04] [VLL04].

In the literature, many surveys deal with the interestingness measures according to two different aspects : the definition of a set of principles to select a suitable interestingness measure, and their comparison with theoretic criteria or experiments on data.

In the perspective to establish the principles of a best interestingness measure, Piatetsky-Shapiro [PS91] presented a new interestingness measure, called Rule-Interest (RI), and proposed three fundamental principles for a measure on a rule $X \rightarrow Y$: $RI = 0$ when X and Y are independent, RI monotonically increases with $X \wedge Y$, RI monotonically decreases with X or Y. Hilderman and Hamilton [HH01] proposed five principles : minimum value, maximum value, skewness, permutation invariance, transfer. Tan et al. [TKS02] [TKS04] defined five interestingness principles : symmetry under variable permutation, row/column scaling invariance, anti-symmetry under row/column permutation, inversion invariance, null invariance. Freitas [Fre99] proposed an "attribute surprising" principle. Gras et al. [GCB$^+$04] [Gui04] proposed a set of ten criteria to design a good interestingness measure.

Among these reviews, some examine the comparison of the interestingness measures from their classification.

- Bayardo and Agrawal [BA99] concluded that the best rules according to all the interestingness measures must reside along a support/confidence border.

- Kononenco [Kon95] analyzed the biases of eleven measures for estimating the quality of multi-valued attributes and showed that the values of information gain, j-measure, gini-index, and relevance tend to linearly increase with the number of values of an attribute.

- Gavrilov et al. [GAIM00] studied the similarity between the measures for classifying them.

- Hilderman and Hamilton [HH01] proposed five principles for ranking summaries generated from databases by using sixteen diversity measures and illustrate that : (1) six measures satisfied the five principles proposed, (2) nine remaining measures satisfied at least one principle.

- By studying twenty-one measures, Tan et al. [TKS02] [TKS04] showed that none of the measures is adapted with all the cases and that the correlation of measures increases when support decreases.

- By using a method of multi-criteria decision-making aid integrating eight criteria, Vaillant et al. [VLL04] [LMP$^+$04] extracted a pre-order on twenty measures and identify four clusters of measures.

- Carvalho et al. [CFE05] [CFE03] evaluated eleven objective interestingness measures in order to rank them according to their effective interest for a decision maker.

- Choi et al. [CAK05] used an approach of multi-criteria decision-making to find the best association rules.

- Blanchard et al. [BGGB05a] [BGGB05b] classified eighteen objective measures into four clusters according to three criteria : independence, equilibrium, and descriptive or statistical characteristics.

- Huynh et al. [HGB05b] proposed a clustering approach by correlation graph which allows to identify eleven clusters on thirty-four interestingness measures.

Finally, two experimental tools are available : HERBS [VPL03] and ARQAT [HGB05a].

Postprocessing

Although the association rule model has the advantage of allowing an unsupervised extraction of rules and of illustrating implicative tendency in data, it has the disadvantage of producing a prohibitive number of rules. The final stage of the rule validation will let the user facing a main difficulty: how he/she can extract the most interesting rules among the large amount of discovered rules.

It is necessary to help the user in his/her validation task by implementing a preliminary stage of postprocessing of discovered rules. The postprocessing task aims at reducing the amount of rules by preselecting a reduced number of rules potentially interesting for the user. This task must take into account both his/her preferences and the data structure.

To solve this problem, five complementary postprocessing approaches are proposed in the literature: constraints, pruning, grouping, summarizing, and visualizing. The first approach considers the whole set of rules as a database on which the user can extract the subsets of rules by requesting some integrated constraints [BBJ00] [SVA97] [WHP06]. The second one allows a significant reduction by pruning the redundant or uninteresting rules [LGB04] [TKR$^+$95] [BKGG03] [BGGB05a] [BGGB05b]. The third one finds a special subset of rules that represents the underlying relationships [LHM99] [LHH00]. The fourth one groups the rules having the same properties into a meta-rule (e.g. a cover) that is more understandable [TKR$^+$95] [DL98]. The last one uses graphical representations to improve the readability of the results [BGB03b].

Since the introduction of *support* and *confidence* measures [AS94], many interestingness measures have been proposed in the literature [BA99] [HH01] [TKS04]. This abundance of interestingness measures leads to a second problem: how to help the user to choose the interestingness measures that are the best adapted to its goals and its data, in order to detect the most interesting rules.

We proposed a data-analysis technique for calculating the most suitable objective interestingness measures on a ruleset or a set of rulesets. For this purpose, some clustering techniques such as AHC [KR90], PAM [KR90] or correlation graph [HGB06d] are used with the correlation values, in order to partition a set of interestingness measures into k_c clusters. The subset of representative measures on the studied data is then constituted by the k_c "central of gravity" obtained.

1.2 Postprocessing models

Together with the huge patterns discovered by the KDS at the mining step, there are many models have been proposed to evaluate the discovered patterns [ST95] [LHML99] and to identify the most interesting ones [ST96]. Fig. 1.2 [ST95] [LHML99] shows the four main models that are widely used in the literature. In Fig. 1.2 (a), the patterns mined by the KDS are immediately considered as the most interesting ones. In Fig. 1.2 (b), the uninteresting patterns are filtered out by an interestingness filter. In Fig. 1.2 (c), the KDS interacts with an interestingness engine. The last model in Fig. 1.2 (d) [LHML99] uses a post-analysis component IAS (Interestingness Analysis System) to help the user identify interesting patterns.

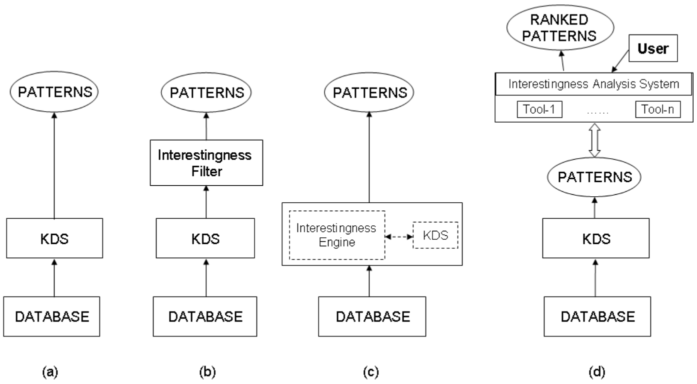


Figure 1.2: Postprocessing models.

The fist model (a) in these four main models is the simplest approach. The KDS outputs the patterns directly to the user. The user can not interact with the KDS because the KDS works independently. When the KDS delivers a considerable number of patterns, it is difficult to the user to realize which ones are the interesting or uninteresting to learn about. The second model (b) depends on the efficiency of the interestingness filter. A lot of number of uninteresting patterns can be cut out by the interestingness filter but like the first model, no user interaction that are integrated into the filter process. The third model (c) is more flexible than the two previous models (a) and (b). It implements an interestingness engine to communicate with the KDS. The domain knowledge can be used efficiently in this model and many users can interact separately or parallel to determine his/her own interests. The fourth model (d) evaluates the patterns issued from the KDS by an independent system. The system includes several tools to rank the patterns with user's interactions. One important feature of the last model is that it allows to use interestingness measures to examine the patterns. Each tool can work independently.

1.3 Clustering techniques

Clustering aims at finding the similar elements in data and to group them into the same group or cluster [KR90] [DHS01]. The number of cluster has to be predetermined [HBV01] [KR90]. The notion of similarity can be understood according to the domain knowledge, the context or the user's point of view. The clustering process is used with no prior knowledge so it is also called unsupervised leaning.

The clustering methods can be divided into 4 groups [JMF99] [HBV01] such as partitioning clustering, hierarchical clustering, density-based clustering and grid-based clustering. Partitioning clustering and hierarchical clustering are two methods that are used frequently. The datasets can be available in one of the following matrices [KR90]:

- an $q \times p$ attribute-value matrix. Columns are attributes and rows are the values of each elements according to the corresponding attributes,
- an $q \times q$ dissimilarity matrix. In this matrix, the difference between each pair of attributes are calculated. This value is considered as a dissimilarity value or a distance value. Noted that $d(i, j) = d(j, i)$ and $d(i, i) = 0$ where $d(,)$ is the dissimilarity or distance between any two attributes i and j.

In the following, we will describe the two techniques that are used in our work.

1.3.1 Partitioning clustering

The elements in a dataset is divided into k_c clusters. The number of clusters k_c is given by the user. How we can determine the "best" partition with a value of k_c is always an attractive problem [HBV01] [Sap90]. The best partition depends on the point of views of the user or it can be determined automatically by a "quality index". Some methods in this clustering type are PAM (Partitioning Around Medoids) – more robust than the k-means [McQ67] –, CLARA (Clustering Large Applications), CLARANS (Clustering Large Application Based on Randomized Search) [KR90].

1.3.2 Hierarchical clustering

The elements in a dataset is firstly grouped into small clusters. These small clusters are then merged into bigger ones. The result is represented by a tree called dendrogram [KR90]. The method that uses the small clusters to merge into bigger clusters is called agglomerative method (i.e. Agglomerative Hierarchical Clustering – AHC). Inversely, the method that processes, from bigger cluster into smaller cluster, is called divisive method.

1.4 Thesis contribution

In recent years, the problem of finding interestingness measures to evaluate association rules has become an important issue in the postprocessing stage in KDD. Many interestingness measures may be found in the literature, and many authors have discussed and compared interestingness properties in order to improve the choice of the most suitable measures for a given application. As interestingness depends both on the data structure and on the decision-maker's goals, some measures may be relevant in some context, but not in others. Therefore, it is necessary to design new contextual approaches in order to help the decision-maker select the most suitable interestingness measures and as a final goal, select the most interesting association rules.

In our thesis, we focus on the interestingness measures technique in the pruning approach, especially on the objective interestingness measures. Postprocessing of association rules with interestingness measures is our main research. Our thesis contributions are:

- A dedicated **ARQAT tool** (Association Rule Quality Analysis Tool) for the postprocessing of association rules [HGB05a]. ARQAT is used to study the specific behavior of a set of interestingness measures in the context of specific datasets and in an exploratory analysis perspective. More precisely, ARQAT is a toolbox designed to help a data-analyst to capture the most suitable measures and as a final purpose, the most interesting rules within a specific ruleset.

Similarly with the model (d) in Fig. 1.2, ARQAT tool has some strong features in comparison with IAS tool: (i) allows ranking the discovered patterns with a set of interestingness measures, (ii) easily determines the most interesting patterns with complementary views, (iii) the comparisons can be drawn on many datasets.

- A **representative measure approach**, especially with **correlation graph** (CG), agglomerative hierarchical clustering (AHC), and partitioning around medoids (PAM) techniques. This approach reflects the postprocessing of association rules, implemented by the ARQAT tool[2]. The approach is based on the analysis the dissimilarities computed from interestingness measures on the data to find the most suitable measures.

- A **comparative study** on the stable clusters between interestingness measures. The result is used to compare and discuss the behavior of 40 interestingness measures on two prototypical and opposite datasets (i.e. a correlated one and a weakly correlated one) and on other two real-life datasets. We focus on the discovery of the stable clusters obtained from the data analyzed between these 40 measures, showing unexpected stabilities.

1.5 Thesis structure

The remaining of the document is structured in seven chapters as follows.

In the first chapter, the association rules discovery process is introduced. We start from the problem of basket market transactions to the association approach with binary representation. Finding frequent itemsets as well as generating rules with anti-monotonicity property is also examined. We classify the improvements from the Apriori algorithm into five principal groups: database pass, computation, itemset compaction, search space and data type.

Chapter 3 gives an overview on the measures of interestingness with two types: subjective measures and objective measures. Two important aspects of the subjective measures such as actionability and unexpectedness are introduced. In this chapter, we analyze an association rule $X \rightarrow Y$ objectively by a function $f(n, n_X, n_Y, n_{X\overline{Y}})$ with four parameters. We also conduct a detail survey on several important properties of an objective measure studied in the literature. As a result, this survey gives us a classification of 40 objective measures. Some mathematical relations between these objective measures are also given in this chapter.

Chapter 4 examines five principal approaches in the postprocessing step of association rules that are well studied in the literature: constraints, pruning, summarization, grouping and visualizing. We proposed our new technique *representative measures* as a pruning approach. This new technique uses AHC, PAM or correlation graph as a means to achieve the most interesting rules.

Chapter 5 represents the ARQAT tool, a new tool for the postprocessing step of association rules written in Java. The tool provides five principal tasks with 14 views. The tool is structured into three analyzed phases: preprocessing, evaluation and display.

Chapter 6 analyzes some principal opposite types of rulesets: correlated vs weakly correlated, real-life vs synthetic. To have a general analysis on these rulesets, we extracted a sample rulesets from each original rulesets. The sample rulesets contains only the most interesting rules due to some objective measures. We also introduce *couple* and *multiple* as two abstract types of rulesets. The couple rulesets is used to evaluate the behaviors on objective measures on an original ruleset with its sample ruleset. The multiple ruleset is used to evaluate on many rulesets. These two types of rulesets are also named commonly as *complement* ruleset. This chapter gives some results on evaluating the

[2] Some other representations with AHC and PAM are illustrated with the R tool – http://www.r-project.org/.

efficiency of the sample models, distributions of interestingness values, joint-distribution matrix, correlation analysis, interesting rules analysis, and ranking of measures by sensitivity values.

Chapter 7 examines the stable clusters of measures between 40 objective measures on three well-known correlation coefficients: Pearson, Spearman and Kendall. A comparative study on the stable behaviors of measures in a rulesets over all these three coefficients is also evaluated.

Finally, chapter 8 gives a summarization of our thesis contribution on the postprocessing step in a KDD process with association rules. We open some future research topics such as improving the sample model, improving the cluster evaluation, hierarchy view, and the aggregated measures by Choquet's or Sugeno's integral.

Chapter 2

Discovery of association rules

Discovering association rules between items in large databases is a frequent task in KDD. The purpose of this task is to discover hidden relations between items of sale transactions. This later is also known as the market basket database[1]. An example of such a relation might be that 90% of customers that purchase bread and butter also purchase milk [AIS93a].

2.1 From the data ...

We consider a database of sale transactions in Tab. 2.1 as a basket data. Each record in this database consists of items bought in a transaction. The problem is how we can find some interesting (i.e. hidden) relations existing between the items in these transactions or some interesting rules that a manager (a user, a decider or a decision-maker) who owns this database can take some valuable decisions. Some rules derived from this database can be {Wine} → {Cheese} or {Bread, Chocolate} → {Milk}.

To facilitate the process of finding the hidden relations mentioned above, the database can be normalized [AS94] [AMS+96]. For instance, the database in Tab. 2.1 is a normalized database because all of its items in the four records are in lexicographic orders.

TID	ITEMS
100	{Bread, Chocolate, Milk}
200	{Cheese, Chocolate, Wine}
300	{Bread, Cheese, Chocolate, Wine}
400	{Cheese, Wine}

Table 2.1: A normalized database of sale transactions.

Definition 2.1.1 (Normalized database [AS94] [AMS+96])**.** A database $\mathcal{D}$ is said normalized if items in each record of $\mathcal{D}$ are kept sorted in their lexicographic order. Each database record is a

[1] In our work, the words *database* and *dataset* can be used interchangeable.

transaction represented by a <TID, item> pair. The transaction is determined by its transaction identifier TID.

2.2 ... To association approach

Association rules are rules of the form *If X then Y* with a percentage of trust (e.g., 2%, 5%, 90%, ...). The formal statement of the problem is given in [AMS+96]. Let $\mathcal{I} = \{i_1, i_2, ..., i_m\}$ be a set of literals, called items. Let $\mathcal{D}$ be a set of transactions, where each transaction T is a set of items such that $T \subseteq \mathcal{I}$. Associated with each transaction is a unique identifer, called its TID. A set of items $X \subset \mathcal{I}$ is called *itemset*. A transaction T is said *contains* X, if $X \subseteq T$. The approach presented in this section is also known as the Apriori algorithm [AS94][AMS+96].

Definition 2.2.1 (Association rule)**.** An *association rule* is an implication of the form $X \rightarrow Y$, where $X \subset \mathcal{I}$, $Y \subset \mathcal{I}$, and $X \cap Y = \emptyset$.

More precisely [AMS+96], $\mathcal{I} = \{i_1, i_2, ..., i_m\}$ is set of attributes over the binary domain $\{0, 1\}$. A tuple T of the database $\mathcal{D}$ is represented by identifying the attributes with value 1. We also called this approach is a binary association rule approach. An example of this transformation is given in Tab. 2.2.

TID	Bread	Cheese	Chocolate	Milk	Wine
100	1	0	1	1	0
200	0	1	1	0	1
300	1	1	1	0	1
400	0	1	0	0	1

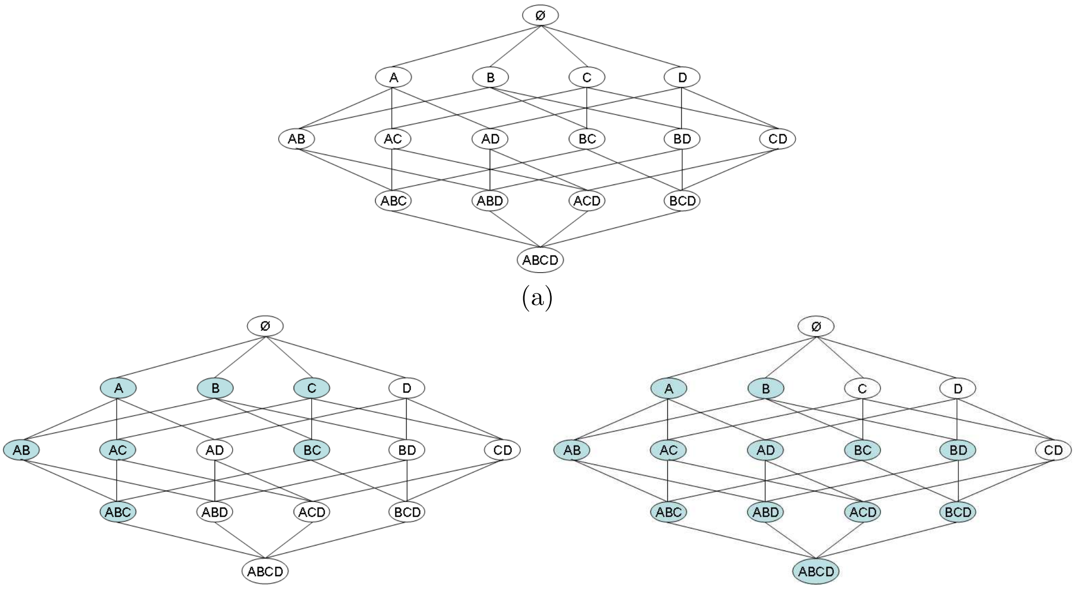
Table 2.2: Binary representation of transactions.

Definition 2.2.2 (Support)**.** The rule $X \rightarrow Y$ has *support* s in the transaction set $\mathcal{D}$ if $s\%$ of transactions in $\mathcal{D}$ contain $X \cup Y$.

Definition 2.2.3 (Confidence)**.** The rule $X \rightarrow Y$ has *confidence* c if $c\%$ of transactions in $\mathcal{D}$ that contain X also contain Y.

The rule $X \rightarrow Y$ holds in the transaction set $\mathcal{D}$ with confidence c and support s. It will be retained when its support and confidence greater than the user-specified minimum support (*minsup*) and minimum confidence (*minconf*).

The process of discovering all association is performed via two steps [AIS93a] [AS94] [AMS+96]: (i) finding all frequent itemsets with *minsup* and *minconf* (all others are *infrequent* itemsets), (ii) generating rules with the frequent itemsets found.



(a)

(b) Monotonicity

(c) Anti-monotonicity

Figure 2.1: Lattice structure.

2.2.1 Monotonicity

A lattice structure is proposed to hold the search space containing all the possibly combinatory cases between the items in the transaction set in the process of finding association rules. The use of this lattice structure leads to an combinatorial explosion of itemsets in the search space 2^n, including the empty set, with n is the number of items. For example, for a search space of 5 items, the number of search elements is $2^5 = 32$. If we note the five items in Tab. 2.1 as A, B, C, D, and E for short, all the combinations between the first four items in form of a lattice is represented in Fig. 2.1 (a).

To overcome the combinatory problem, the monotonicity (anti-monotonicity respectively) property is used to prune unnecessary cases efficiently.

Definition 2.2.4 (Monotonicity/Anti-monotonicity)**.** If a set X has a property t then all its subsets/supersets also have the property t.

Reconsider the lattice constructed from four items {A, B, C, D} in Fig. 2.1 (a). If {A, B, C} are in grey then all its subsets {AB, AC, BC, ABC} are in grey with the monotonicity property (see Fig. 2.1 (b)). For the anti-monotonicity property (see Fig. 2.1 (c)), if {A, B} are in grey then all its supersets {AB, AC, AD, BC, BD, ABC, ABD, ACD, BCD, ABCD} are also in grey.

2.2.2 Frequent itemsets

The itemsets that have transaction support above *minsup* are called frequent itemsets[2]. A *k-itemset* is an itemset has itself k items. To discover frequent itemsets, multiple passes over the database are

[2] The term *frequent* itemset is also known as *large* or *covering* itemset in [AS94] [AMS+96]. Infrequent itemset is called *small* itemsets.

performed. Firstly, the support of each item is counted and the item type (frequent/infrequent) is determined by using *minsup* value.

Candidate itemsets

In the next passes, the frequent itemsets discovered in the previous pass is held as a seed set to generate new potentially frequent itemsets or *candidate* itemsets using the monotonicity property. The support counts for these candidate itemsets are also determined during this process. Among these candidate itemsets at the end of a database pass, those ones have the support count above the *minsup* threshold are held, the others are cut by using the anti-monotonicity property. The process will finish when no new frequent itemsets appear.

Algorithm

Let L_k be the set of frequent k-itemsets. Let C_k be the set of candidate k-itemsets. The algorithm to find all frequent itemsets can be represented as follows [AMS+96].

```
Algorithm FindFrequentItemset
$L_1$ = frequent 1-itemsets;
for (k = 2; $L_{k-1} \neq \emptyset$; k++) do begin
   $C_k$ = apriori-gen($L_{k-1}$);
   forall transactions $t \in \mathcal{D}$ do begin
      $C_t$ = subset($C_k, t$);
      forall candidates $c \in C_t$ do
         c.count++;
   end;
   $L_k = c \in C_k | c.count \geq minsup$
end
return $\bigcup_k L_k$
```

The **apriori-gen** function returns a superset of the set of all frequent $(k-1)$-itemsets from the previous pass L_{k-1}. It works in two steps: (i) joints L_{k-1} with L_{k-1} and (ii) prunes all itemsets $c \in C_k$ in which (k-1)-subset of c is not in L_{k-1}. For instance, let L_3 be {ABC, ABD, ACD, ACE, BCD}. The candidate C_4 will be {ABCD, ACDE}. Because the itemset {ADE} is not in L_3 so the itemset {ACDE} will be deleted and only the itemset {ABCD} exists in C_4. Fig. 2.2 illustrates a complete example with three database passes for the transactions in Tab. 2.1.

2.2.3 Rule generation

The rule generated in this process has the form $X \rightarrow Y$ in which X and Y are two itemsets, $|X| > 0$, $|Y| > 0$, $\frac{support(X \cup Y)}{support(X)} \geq minconf$ [AS94] [AMS+96]. X and Y are called the antecedent and the consequent of the rule $X \rightarrow Y$ respectively. Given a frequent itemset L, all the rules $X \rightarrow (L - X)$ are outputted where $X \subseteq L$ and the previous conditions are held. The RuleGeneration algorithm creates firstly all the rules with one item in the consequent. The consequents of these rules are used as a seed set (as the function **apriori-gen** in Sec. 2.2.2) to generate all the rules with two items in the consequent and so on.

Database $\mathcal{D}$

TID	Items
100	A C D
200	B C E
300	A B C E
400	B E

$\xrightarrow{\text{Counting}}$

C_1

Itemset	Support
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

$\xrightarrow{\text{Pruning}}$

L_1

Itemset	Support
{A}	2
{B}	3
{C}	3
{E}	3

C_2

Itemset
{A B}
{A C}
{A E}
{B C}
{B E}
{C E}

$\xrightarrow{\text{Counting}}$

C_2

Itemset	Support
{A B}	1
{A C}	2
{A E}	1
{B C}	2
{B E}	3
{C E}	2

$\xrightarrow{\text{Pruning}}$

L_2

Itemset	Support
{A C}	2
{B C}	2
{B E}	3
{C E}	2

C_3

Itemset
{B C E}

$\xrightarrow{\text{Counting}}$

C_3

Itemset	Support
{B C E}	2

$\xrightarrow{\text{Pruning}}$

L_3

Itemset	Support
{B C E}	2

Figure 2.2: An example of finding all frequent itemsets with three database passes.

A remark for the RuleGeneration algorithm can be drawn. If the rule $X \rightarrow Y$ are held then all the rules $X \cup (Y - Y') \rightarrow Y'$ are also held where $Y' \subseteq Y$. For example, if the rule $AB \rightarrow CD$ is held, then the other rules such as $ABC \rightarrow D$ and $ABD \rightarrow C$ are also held. D in $ABC \rightarrow D$ and C in $ABD \rightarrow C$ hold the role of Y' respectively.

2.3 Algorithm improvement

The volume of association rules tends to be huge and the time execution to be expensive. To improve the efficiency of finding association rules, several strategies are proposed. We classify these strategies into five subgroups: database pass, computation, itemset compaction, search space and data type. Excellent surveys can be found in [CHY96] [ZB03] [TSK06] [CR06] [Zak99].

2.3.1 Database pass

The principal Apriori algorithm requires multiple passes over the database: for finding the candidates of k-itemsets, k passes are executed (see Fig. 2.2). The approaches proposed in this strategy aims at reducing a significance number of scans, performing efficiently in I/O times.

Partition

Partition [SON95] uses at most two passes on the database. It generates a set of all potentially frequent itemsets in the first scan. The counters for each of the itemsets are set up, together with their actual support is measured, in the second one.

```
Algorithm RuleGeneration
forall large $k$-itemsets $l_k, k \geq 2$ do begin
  $H_1 = \{$consequents of rules from $l_k$ with one item in the consequent$\}$
  call ap-genrules($l_k$, $H_1$);
end

procedure ap-genrules($l_k$: large $k$-itemset, $H_m$: set of $m$-item consequents)
  if (k < m + 1) then begin
    $H_{m+1}$ = apriori-gen($H_m$);
    forall $h_{m+1} \in H_{m+1}$ do begin
      $conf$ = support($l_k$)/support($l_k - h_{m+1}$);
      if (conf $\geq$ minconf) then
        output the rule ($l_k - h_{m+1} \rightarrow h_{m+1}$)
          with confidence = $conf$ and support = support($l_k$)
      else
        delete $h_{m+1}$ from $H_{m+1}$
      end
      call ap-genrules($l_k$, $H_{m+1}$);
end
```

The database in the first step is divided into a number of non-overlapping partitions. Each partition is considered one at a time and all frequent itemsets for that partition are generated. Then, these frequent itemsets are merged to generate a set of all potentially frequent itemsets. The supports for these merged itemsets are generated and the frequent itemsets are identified in the second step. Each partition is read into the memory only one time according to its appropriate size.

DIC

Dic [BMUT97] finds frequent itemsets with fewer database passes based on the idea of item reordering. By keeping the number of itemsets which are counted in any pass, the number of passes is then reduced efficiently. It means that the support of an itemset can be counted whenever it may be necessary to count instead of waiting until the end of the previous pass. For instance, Apriori can produce 3 passes for counting 3-itemsets while Dic produces 1.5 passes. And in the first pass with 1-itemset, Dic can count some itemsets that are in 2-itemsets or 3-itemsets. Four possible states (confirmed frequent, confirmed infrequent, suspected frequent, suspected infrequent) for an itemset are used during the time that the itemsets are counted in one pass.

Sampling

Sampling [Toi96] can find association rules very efficiently in only one database pass and two passes in the worst case. A random sample S is picked and used to find frequent itemsets with the concept of negative border.

Let S be a set of itemsets. The *border* Bd(S) of S is defined as a set of itemsets such that all subsets of Bd(S) are in S and none of the supersets of Bd(S) is in S [MT96]. The positive border $Bd^+(S)$ (the negative border $Bd^-(S)$ respectively) contains the itemsets such that $Bd^+(S) = \{x | x \in Bd(S), x \in S\}$ ($Bd^-(S) = \{x | x \in Bd(S), x \notin S\}$ respectively). We have $Bd(S) = Bd^+(S) \cup Bd^-(S)$ and $Bd(S)$ can be very small even for large S. For instance, consider Fig. 2.1 (b) where S = {A, B, C, AB, AC, BC, ABC}. In S, only ABC has all its supersets are not in S so $Bd^+(S) = \{ABC\}$. We can see the same situation with the negative border when only D has all its subsets are in S (e.g. $\{\emptyset\}$) and D is not in S, so we have $Bd^-(S) = \{D\}$. The border set of S can be established as $Bd(S) = Bd^+(S) \cup Bd^-(S) = \{ABC\} \cup \{D\} = \{D, ABC\}$.

2.3.2 Computation

To resolve the problem of expensive cost of time calculations, some approaches based on the parallel and distributed techniques are proposed [AS96] [PCY95] [Zak99]. The task can be parallel with the whole or partitioned database. The database can be distributed across the multiprocessors. Overall, four techniques are introduced: (i) count distribution, (ii) data distribution, (iii) candidate distribution, and (iv) rule distribution.

Assuming a shared-nothing architecture with n processors in which each processor has a private memory and a private disk. The processors are connected by a communication network and can communicate only by passing messages. Data is distributed on the disks attached to the processors and the disk of each processor has approximately an equal number of transactions.

Count distribution

In this approach, the support counts for L_k in the pass k are distributed. Except the first pass is special, from the second pass we have five tasks in each pass: (i) a processor $\mathcal{P}^i$ generates the complete C_k taking as argument the frequent itemset L_{k-1} of the pass $k-1$. The same C_k will be generated overall processors $\mathcal{P}^i$ because of the same L_{k-1}, (ii) the local support for each candidate is counted with a pass from the processor $\mathcal{P}^i$ on its data $\mathcal{D}^i$, (iii) the global count for C_k are determined by exchange with the local C_k counted by each processor P^i, (iv) L_k is computed from C_k in each processor P^i, (v) the continuation is depended on each processor P^i independently. The most important feature of the count distribution is that no data are exchanged between processors except the counts.

Data distribution

This approach is used to better exploit the total memory in a system with n processors. The purpose is to count in a single pass a candidate set that would require n passes in the count distribution approach. So that every processor must broadcast its local data to all other processors in every pass. On a machine with very fast communication, this approach will work viable.

Candidate distribution

Each processor in this approach may proceed independently with both the data and the candidates partitioned. In a pass k, the algorithm divides the frequent itemset L_{k-1} between processors in such a way that a processor P^i can generate a unique C^i_m $(m \geq k)$ independent of all other processors $(C^i_m \cap C^j_m = \emptyset, i \neq j)$. Because the data is partitioned so that a processor can count candidates in C^i_m independent of all other processors.

Rule distribution

The above three techniques are about generating frequent itemsets. This technique is to generate rule by parallel implementation. By partitioning the set of all frequent itemsets into each processors, then the proportions of rules are generated conformably.

2.3.3 Frequent itemset compaction

Because of the combinatory cases between items in a set of transactions, the problem of compacting a lattice structure is introduced. Two useful representations of the frequent itemsets are *frequent closed itemsets* and *maximal frequent itemsets*. We have the following relation between these three sets: $\{$maximal frequent itemsets$\} \subseteq \{$frequent closed itemsets$\} \subseteq \{$frequent itemsets$\}$.

Frequent closed itemsets

A closed itemset is a maximal set of items common to a restricted set of transactions [PBTL99]. It will be called a frequent closed itemset if its support count is greater or equal to *minsup* threshold. For instance, the database $\mathcal{D}$ in Tab. 2.1 or in Fig. 2.2 has the itemset BCE is a closed itemset because it satisfies the closed condition with two transactions TID=200 and TID=300. It is also a frequent closed itemset when its support count is $2 \geq minsup$. The useful meaning of this technique is that 50% of customers purchase *at most* three items Cheese, Chocolate, and Wine. The total set of frequent closed itemsets for the transactions in the database $\mathcal{D}$ is {AC, BE, C, BCE}.

The closed itemset lattice is often much smaller than the itemset lattice. By using a closure mechanism based on the Galois connection and two properties that (i) the support of an itemset T is equal to the support of its closure and (ii) the set of maximal frequent itemsets[3] is identical to the set of maximal frequent closed itemsets. With a reduced set of frequent closed itemsets instead of a larger frequent itemsets then the set of association rules can be reduced without the loss of information.

Maximal frequent itemset

A maximal frequent itemset [BCF+05] is a frequent itemset and all its superset are infrequent. This technique is especially efficient when the itemsets are very long (more than 15 to 20 items). By using a cut through the lattice structures so all itemsets above the cut are frequent itemsets, and all their subsets below are infrequent. All the combinations above the cut (i.e. frequent itemsets) form the positive border when all the combinations below the cut (i.e. infrequent itemsets) form the negative border.

In Fig. 2.1 (b), if we consider all the nodes in grey are frequent and not in grey are infrequent, then only one itemset ABC is a maximal frequent itemsets because all its supersets are infrequent.

2.3.4 Search space

Due to the explosive combination of the candidate itemsets while searching the frequent itemsets, many authors have proposed some efficient search space techniques in which we can ignore the step of finding candidate itemsets and we can go directly to find the frequent itemsets. Most of them

[3] The meaning of maximal set in this case is used in a normal way. This is not understandable like the next approach with maximal frequent itemset.

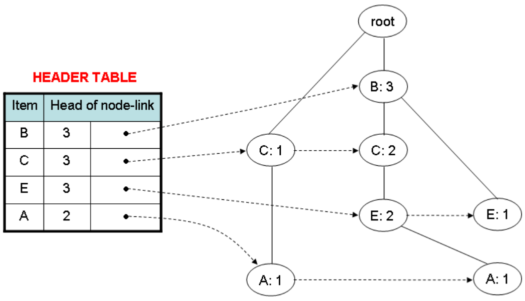
are based on tree structures. We will present such three attractive structures: (i) FP-tree, (ii) lexicographic tree, and (iii) T-tree and P-tree.

FP-tree

FP-tree (Frequent-Pattern Tree) [HPYM04] is a tree structure for mining frequent itemsets efficiently without candidate generation. It is defined as: (i) one root labeled as "null", a set of item-prefix subtrees as the children of the root, and a frequent-item-header table (ii) each node of the item-prefix subtree consists of three fields: item-name, count, and node-link where item-name registers which item this node represents, count registers the number of transactions represented by the portion of the part reaching this node, and node-link links to the next node in the FP-tree carrying the same item-name, or null if there is none (iii) each entry in the frequent-item-header consists of two fields: (iii-1) item-name and (iii-2) head of node-link (a pointer pointing to the first node in the FP-tree carrying the item-name). For instance, in Fig. 2.3, an FP-tree is constituted from a set of transactions in Tab. 2.1, ordered by support count in Tab. 2.3.

TID	Items bought	(Ordered) frequent items
100	A C D	C:3 A:2
200	B C E	B:3 C:3 E:3
300	A B C E	B:3 C:3 E:3 A:2
400	B E	B:3 E:3

Table 2.3: Ordering items with support count (*minsup* = 2).


HEADER TABLE

Item	Head of node-link	
B	3	•
C	3	•
E	3	•
A	2	•

Figure 2.3: An FP-tree constituted from a set of transactions.

FP-tree mines completely the set of frequent itemsets by itemset fragment growth. The FP-growth algorithm uses this tree to find frequent itemset efficiently with both long and short itemsets. It can be considered as an order of magnitude faster than the Apriori algorithm. Tab. 2.4 shows the conditional FP-tree constituted from the FP-tree in Fig. 2.3 for extracting frequent itemsets.

COFI-tree [EHZ03] and CATS-tree [CZ03] are two tree structures that are inspired from the FP-tree to reduce the memory space needed.

Item	Conditional pattern-base	Conditional FP-tree
A	{(C:1), (BCE:1)}	{(C:2)}|A
E	{(BC:2), (B:1)}	{(B:3)}|E
C	{(B:2)}	{(B:2)}|C
B	∅	∅

Table 2.4: Mining frequent itemsets with conditional FP-tree.

Lexicographic tree

By constructing successively the nodes of a lexicographic tree of itemsets, the set of frequent itemsets are constituted [AAP01]. Fig. 2.4 show a lexicographic tree established from 5 items. The frequent itemsets are represented as nodes of the lexicographic tree to reduce the CPU time for counting frequent itemsets. The support counts of the frequent itemsets, in a top-down manner, are determined by projecting the transactions onto the nodes of the tree. It illustrates an advantage of visualizing itemset generation with the flexibility of picking the correct strategy during the tree generation phase as well as transaction projection phase.

The lexicographic tree is defined as: (i) a vertex exists in the tree corresponding to each frequent itemset (the root corresponds to the null itemset), (ii) let $I = \{i_1, i_2, ..., i_k\}$ be a frequent itemset, where $i_1, i_2, ..., i_k$ are listed in lexicographic order and the parent of the node I is the itemset $\{i_1, i_2, ..., i_{k-1}\}$. The levels in a lexicographic tree correspond the itemset sizes. The possible states for a node are: generated, examined, null, active or inactive.

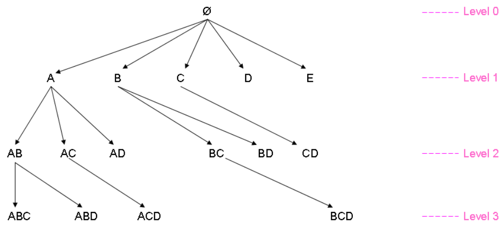


Figure 2.4: Lexicographic tree.

T-tree and P-tree

P-tree and T-tree that are two tree structures quite similarly for counting the support of itemsets based on Rymon's set enumeration tree [CGL04]. T-tree (Total Support Tree) is implemented to reduce number of links needed and providing direct indexing. P-tree (Partial Support Tree) is used with the concept of *partial support*. Fig. 2.5 illustrates an example with 5 items {A, B, C, D, E}.

2.3.5 Data type

Due to the different types of data, many authors have conducted their researches to discovery association rules adapted to different data types such as quantitative and categorical [SA96], interval [MY97], spatial [KH95], temporal [CP00], ordinal [Gui02], multimedia [ZHZ00], text [AAFF05],

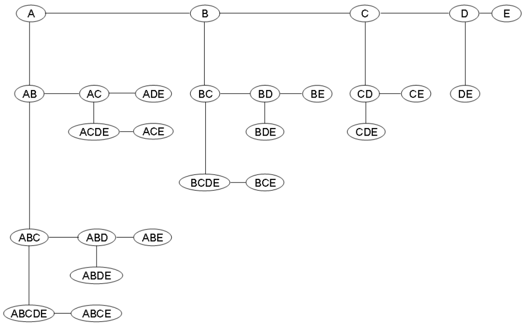


Figure 2.5: P-tree in the form of Rymon's set enumeration tree.

multidimensional [LHF98], and taxonomy-based [SA95]. We will present in this part the two most important data types are recently developed and widely used: quantitative/categorical and taxonomy-based.

Quantitative and categorical

When an attribute (i.e. an item) is quantitative or categorical then boolean attributes can be considered a special case of categorical attributes [SA96]. Fig. 2.6 shows a People table with three non-key attributes: *Age*, *NumCars* (quantitative), and *Married* (categorical). A possible quantitative association rule is: $< Age : 30..39 > and < Married : Yes > \rightarrow < NumCars : 2 >$.

TID	Age	Married	NumCars
100	23	No	1
200	25	Yes	1
300	29	No	0
400	34	Yes	2
500	38	Yes	2

Figure 2.6: Quantitative and categorical items.

Many fields as the number of attribute values are established instead of just one field in the table. For a quantitative attribute, its values will be partitioned into intervals and then map each $< attribute, interval >$ pair to a boolean attribute as in the boolean field. The number of intervals can be calculated as: $\frac{2n}{m(K-1)}$ where n is the number of quantitative attributes, m is the minimum support (as a fraction) and K is the partial completeness level (see how we can find K in [SA96]). Fig. 2.7 shows this mapping for the non-key attributes of the People table given in Fig. 2.6.

Taxonomy-based

When there is a taxonomy on the items in the transactions of a database (see Fig. 2.8), the association rules can be generated for all levels (e.g. generalized – [SA95] or at different levels (e.g. multiple-level association rules – [HF99]). For example, a generalized association rule is an association rule $X \rightarrow Y$ that is no item in Y is an ancestor of any item in X. The reason for

TID	Age:20..29	Age:30..39	Married:Yes	Married:No	NumCars:0	NumCars:1	NumCars:2
100	1	0	0	1	0	1	0
200	1	0	1	0	0	1	0
300	1	0	0	1	1	0	0
400	0	1	1	0	0	0	1
500	0	1	1	0	0	0	1

Figure 2.7: Mapping to the boolean problem.

the condition that no item in Y should be an ancestor of any item in X is that a rule of the form ”x → ancestor(x)” is trivially true with 100% confidence, and hence redundant. The rules are called generalized association rules because both X and Y can contain items from any level of the taxonomy, a possibility not entertained by the formalism introduced in [AIS93a].

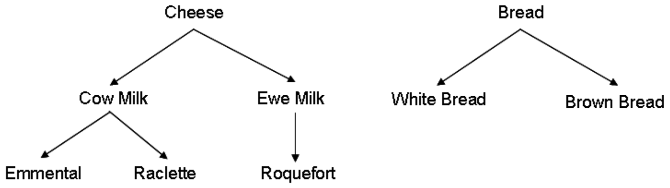


Figure 2.8: A taxonomy of items.

Inspired from a similar example illustrated in [SA95], given a set of items {Cheese, Cow Milk, Emmental, Raclette, Ewe Milk, Roquefort, Bread, White Bread, Brown Bread}, Fig. 2.8 shows the corresponding taxonomy being composed of these items. Consider the set of transactions shown in Fig. 2.9 (a) with *minsup* = 2. Then the frequent itemsets corresponding to these itemsets are illustrated in Fig. 2.9 (b).

TID	Items bought
100	Roquefort
200	Emmental, Brown Bread
300	Raclette, Brown Bread
400	White Bread
500	White Bread
600	Emmental

(a) $\mathcal{D}$

$\xrightarrow{\text{Adding taxonomy items}}$

Itemset	Support
{Emmental}	2
{Cow Milk}	3
{Cheese}	4
{White Bread}	2
{Brown Bread}	2
{Bread}	4
{Cow Milk, Brown Bread}	2
{Cheese, Brown Bread}	2
{Cow Milk, Bread}	2
{Cheese, Bread}	2

(b) Frequent itemsets

Figure 2.9: Adding taxonomy items to find frequent itemsets.

2.4 Summary

Mining association rules is one of the most attractive problems in KDD. In this chapter, we have presented the principal keys in the Apriori algorithm to resolve this challenge. We have classified the recent techniques in this domain into five different subgroups: database pass, computation, itemset compaction, search space, data type. Some concrete examples are also provided.

Chapter 3

Measures of interestingness

Over the last decade the KDD community has recognized this challenge – often referred to as interestingness – as an important and difficult component of the KDD process [**?**] [ST96] [LHML99] [HH01] [TKS04]. To tackle this problem, the most commonly used approach is based on the construction of interestingness measures (called measure for short).

In defining association rules, Agrawal *et al.* [AIS93a] [AS94] [AMS$^+$96], introduced two measures: support and confidence. These are well adapted to APRIORI algorithm constraints, but are not sufficient to capture the whole aspects of the rule interestingness. To push back this limit, many complementary measures have been then proposed in the literature (see [PSM94] [ST96] [SM99] [BA99] [HH01] [TKS04] [TS06] [McG05] [GH06] [Omi03] for a survey). They can be classified into two categories [ST96]: *subjective* and *objective.* Subjective measures explicitly depend on the user's goals and his/her knowledge or beliefs. They are combined with specific supervised algorithms in order to compare the extracted rules with the user's expectations [ST96] [PT98] [LHML99]. Consequently, subjective measures allow the capture of rule novelty and unexpectedness in relation to the user's knowledge or beliefs. Objective measures are numerical indexes that only rely on the data distribution. Interestingness refers to the degree to which a discovered pattern is of interest to the user and is driven by factors such as novelty, utility, relevance and statistical significance [FPSM91] [PSM94].

3.1 Subjective measures

Subjective measures [PSM94] [ST95] [ST96] are studied in a domain-independent context. The interestingness of a pattern is evaluated subjectively from the user point of view. How a pattern can be interesting is determined by the two following approaches [ST96]: (i) a pattern is *unexpectedness* if it is "surprising" to the user [ST95], and (ii) a pattern is *actionability* if the user can act on it to his advantage [PSM94].

3.1.1 Actionability

Actionability is a subjective measure that allows the user taking some specific actions in response to the newly discovered knowledge [ST96]. How we can capture actionable patterns is a difficult issue because the actions from a point of view can change with the time and is is not easy to maintain them.

The actionable patterns can be found by a system [PSM94] via the discovery of deviation of patterns, an action hierarchy [AT97] or mining patterns that respond to actions [JWTF05].

3.1.2 Unexpectedness

Unexpectedness is a subjective measure that provides unexpected patterns contradicting the user's expectations which depend strongly on his/her beliefs [ST96] [PT99]. Beliefs can be classified into categories: (i) hard beliefs (i.e., unchanged constraints, depending strongly on the user's point of view), and (ii) soft beliefs (i.e., the user is willing to change with a *degree* of believe). Degree of soft beliefs can be assigned in different approaches: Bayesian, Dempster-Shafer, frequency, Cyc, or statistic.

A pattern is always interesting if it contradicts the set of hard belief. For soft belief, the interestingness of pattern p is computed as:

$$I(p,B,\xi) = \sum_{\alpha_i \in B} w_i |d(\alpha_i|p,\xi) - d(\alpha_i|\xi)|$$

where w_i is a weight function associated with each soft belief α_i in the soft belief system B, $\sum_{\alpha_i \in B} w_i = 1$, and ξ is the previous evidence.

3.2 Objective measures

In the following, we consider a finite set $\mathcal{T}$ of transactions. We denote an association rule by $X \rightarrow Y$ where X and Y are two disjoint itemsets. The itemset X (respectively Y) is associated with a transaction subset $t_X = \mathcal{T}(X) = \{T \in \mathcal{T}, X \subseteq T\}$ (respectively $t_Y = \mathcal{T}(Y)$). The itemset $\overline{X}$ (respectively $\overline{Y}$) is associated with $t_{\overline{X}} = \mathcal{T}(\overline{X}) = \mathcal{T} - \mathcal{T}(X) = \{T \in \mathcal{T}, X \nsubseteq T\}$ (respectively $t_{\overline{Y}} = \mathcal{T}(\overline{Y})$). In order to accept or reject the general trend to have Y when X is present, it is quite common to consider the number $n_{X\overline{Y}}$ of negative examples (contra-examples, counter-examples) of the rule $X \rightarrow Y$. Each rule is described by the four parameters : $n = |\mathcal{T}|$, $n_X = |t_X|$, $n_Y = |t_Y|$, $n_{\overline{X}} = |t_{\overline{X}}|$, $n_{\overline{Y}} = |t_{\overline{Y}}|$.

Let us denote that, for clarity, we also keep the probabilistic notations $p(X)$ (respectively $p(Y)$, $p(X \cap Y)$, $p(X \cap \overline{Y})$) as the probability of X (respectively Y, $X \cap Y$, $X \cap \overline{Y}$). This probability is estimated by the frequency of X: $p(X) = \frac{n_X}{n}$ (respectively $p(Y) = \frac{n_Y}{n}$, $p(X \cap Y) = \frac{n_{XY}}{n}$, $p(X \cap \overline{Y}) = \frac{n_{X\overline{Y}}}{n}$).

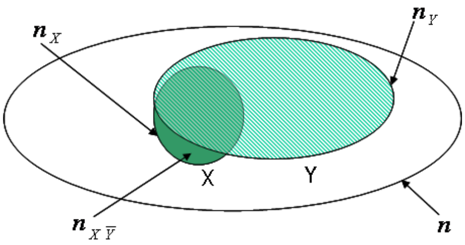


Figure 3.1: The cardinalities of a rule $X \rightarrow Y$ and the "surprisingness" of negative examples.

An interestingness value is then calculated by a numerical function on the cardinalities of a rule: $m(X \rightarrow Y) = f(n, n_X, n_Y, n_{X\overline{Y}}) \in \mathbb{R}$. The higher the interestingness value, the more interesting the rule.

Example. Given two disjoint itemsets X and Y as follows. The itemset X has one item and the itemset Y has three items. The association rule is of the form $X \rightarrow Y$. The item of the premise presents a logical expression and the items in the conclusion differentiate each others by the symbol $\cap$.

$X = \{stalk\_surf\_above = SMOOTH\}$

$Y = \{BROAD \wedge BRUISES \wedge EDIBLE\}$

where $n = 100$, $n_X = 50$, $n_Y = 80$, $n_{X\overline{Y}} = 10$ respectively.

Given an interestingness measure, namely Pavillon, determined by the negative examples $m(X \rightarrow Y) = f(n, n_X, n_Y, n_{X\overline{Y}}) = \frac{n_{\overline{Y}}}{n} - \frac{n_{X\overline{Y}}}{n_X}$.

For a coherent representation of the measure formula, some intermediatory parameters can be easily established:

$n_{XY} = n_X - n_{X\overline{Y}}$,

$n_{\overline{X}} = n_X - n_{X\overline{Y}}$,

$n_{\overline{Y}} = n_Y - n_{X\overline{Y}}$,

$n_{\overline{X}Y} = n_Y - n_X + n_{X\overline{Y}}$,

$n_{\overline{XY}} = n - n_Y - n_{X\overline{Y}}$.

Therefore, the interestingness value of the association rule $X \rightarrow Y$ with respect to the interestingness measure m is then computed : $m(X \rightarrow Y) = \frac{80-10}{100} - \frac{10}{50} = 0.5$.

3.2.1 Criteria

To quantify the interestingness of an objective measure, also called quality measure, some criteria are proposed to understand their behaviors [Bla68] [PS91] [MM95] [Fre99] [HH01] [TKS04] [GCB$^+$04] [VLL04] [Gui04] [LT04] [LVL05] [LLV05] [GH06] [RZ06] [GK06]. Several important criteria discuss below are evaluated in Tab. 3.2. To understand the behaviors of a set of objective measures, a special study on this problem is examined in Chap. 7.

Variation

The interestingness value monotonically increases with n_{XY} and monotonically decreases with $n_{X\overline{Y}}$ or $n_{\overline{X}Y}$. Notice that n_α varies when the other parameters are fixed. A rule may have the same order with two measures satisfied the above properties (i.e. conjuncture).

The decrease of value must be slowly with the first appearance of negative examples by chance, noisy, or error of observation [GCB$^+$04]. Then it decreases quickly when the observation of negative examples confirms the strong absence of the rule. The value of a measure must also decrease when the situation of trivial observations appeared (i.e. not containing any information in the sense of Shannon entropy).

In addition, the measure has not to vary linearly with the number of negative examples.

Particular situation

Two particular cases: independence and equilibrium are examined. They are also called the "subjective" (i.e. "subject") aspect of an objective measure.

Independence is a situation in which the antecedent and the consequent of a rule are statistically independent $n_{XY} = \frac{n_X n_Y}{n}$ or $n_{X\overline{Y}} = \frac{n_X n_{\overline{Y}}}{n}$, so we have $m(X \rightarrow Y) = f(n, n_X, n_Y, \frac{n_X n_{\overline{Y}}}{n}) = constant$.

Equilibrium is a situation in which the numbers of examples and negative examples of a rule are the same: $n_{XY} = n_{X\overline{Y}} = \frac{n_X}{2}$, so we have $m(X \rightarrow Y) = f(n, n_X, n_Y, \frac{n_X}{2}) = constant$.

By examining the interestingness value changes from the independence or equilibrium value, the measure are evaluated like the deviation from independence or equilibrium.

Fixing a *threshold* of interestingness is necessary when one would like to observe a strict range of interestingness values. When $n_{X\overline{Y}} = 0$ the implicative rule becomes a "logical" rule that has not the implicative tendency.

Paradoxical phenomenon

The value of a measure must not be the same when two paradoxical situations appear. For example, the symmetric situation $m(X \rightarrow Y) = m(Y \rightarrow X)$; or the contrary situation $m(X \rightarrow Y) = m(X \rightarrow \overline{Y})$.

Countable

The analytical property of a measure are countable to give an order or preorder structure. It allows to evaluate the structure induced on the set of principal variables.

Diversification

A measure must be sufficiently flexible and general analysis to process on different types of variables.

Discriminative ability

The criminative ability of a measure is not influenced with noisy or big volume of data (i.e. $n \nearrow$). The value of a measures does not vary when its cardinalities vary with a coefficient α (i.e. $m(X \rightarrow Y) = f(n, n_X, n_Y, n_{X\overline{Y}}) = f(\alpha \cdot n, \alpha \cdot n_X, \alpha \cdot n_Y, \alpha \cdot n_{X\overline{Y}})$) is called *descriptive measure*, *statistical measure* vice versa [LT04]. Descriptive or statistical aspect of a measure also called the "nature" of a measure.

Interpretable

The formula and the algorithms to measure the rule interestingness have shortly execution time. Their definition are intuitively appreciable and the obtained value holds a signification to interpret.

Imbalance

Taking into account with the small number of examples (i.e. $n_{XY} \ll n$) because they can be the nuggets of knowledge.

Attribute interestingness

When a rule is considered as a whole, it may lead to the situation in which two rules will have the same value of interestingness. In fact, these two rules can have different degrees of interestingness for the user, depending on which attributes occur in the rule antecedent. To resolve this problem, one has to consider the interestingness of individual attributes occurring in the rule antecedent.

Quasi-

The problem of quasi-implication, quasi-conjunction and quasi-equivalence are considered to determine some special cases between objective measures [Bla05].

- A *quasi-implication* measure is a measure which satisfies the condition $m(X \rightarrow Y) = m(\overline{Y} \rightarrow \overline{X})$ where $f(n, n_X, n_Y, n_{X\overline{Y}}) = f(n, n - n_Y, n - n_X, n_{X\overline{Y}})$.
- A *quasi-conjunction* measure is a measure which satisfies the condition $m(X \rightarrow Y) = m(Y \rightarrow X)$ where $f(n, n_X, n_Y, n_{X\overline{Y}}) = f(n, n_Y, n_X, n_{X\overline{Y}})$.
- A *quasi-equivalence* measure is a measure which satisfies the condition $m(X \rightarrow Y) = m(Y \rightarrow X) = m(\overline{Y} \rightarrow \overline{X}) = m(\overline{X} \rightarrow \overline{Y})$ where $f(n, n_X, n_Y, n_{X\overline{Y}}) = f(n, n_Y, n_X, n_{\overline{X}Y}) = f(n, n_{\overline{Y}}, n_{\overline{X}}, n_{X\overline{Y}})$ $= f(n, n_{\overline{X}}, n_{\overline{Y}}, n_{\overline{X}Y})$.

We have: {quasi-equivalence} = {quasi-implication} $\cap$ {quasi-conjunction}.

3.2.2 Classification

A classification according to the "nature" and the "subject" of the objective measures, firstly proposed by Blanchard et al. [HGB+06a], is given in Tab. 3.3. On the column, we can see that most of the measures are descriptive. Another observation shows that IPEE is the only statistical measure computing the deviation from equilibrium.

The classification also gives a quick view on the mutual relations between objective measures. It will help us to understand why the objective measures can be clustered. For instance, most of the measures issued from the Confidence measure are descriptive and deviation from equilibrium: Confidence, Descriptive Confirmed-Confidence, Example & Contra-Example, and Laplace.

3.2.3 Mathematical relation

Observing the relations between measures mathematically are useful when one desires to discover the constraints on interestingness values between them. Tab. 3.4 gives some relationships discovered.

This work is useful and interesting for reducing the quantity of measures. If one measure strongly depends the other measures, we will not consider it any more. For instance, if we have two measures *TauxDeLiaison* and *Lift* that have a mathematical relation $TauxDeLiaison = Lift - 1$, we will only select *Lift* for both of them. From line 15 and 16 (Tab. 3.4), we can see that the two measures Yule's Q and Yule's Y have a close relation with the Odds Ratio measure. If the value of the Conviction measure increases, the value of the Loevinger measure will also increase (line 9, Tab. 3.4).

In addition, we can calculate the interestingness value of a measure by using the interestingness values from the measures participating in the corresponding formula. For example, the value of Rule Interest measure (line 2, Tab. 3.4) can be calculated by the other measures such as Support, Confidence and Pavillon.

3.3 Summary

Ranking association rules by interestingness measures is a research domain that attracts many authors in the literature. Two classes of measures are identified: subjective and objective measures. We have examined some important properties that are widely discussed to give a general panorama on this problem. Some mathematical between objective measures are illustrated to show the constraints on the interestingness values.

N°	Interestingness Measure	$f(n, n_X, n_Y, n_{X\overline{Y}})$	Reference
1	Causal Confidence	$1 - \frac{1}{2}(\frac{1}{n_X} + \frac{1}{n_{\overline{Y}}})n_{X\overline{Y}}$	[Kod01]
2	Causal Confirm	$\frac{n_X+n_{\overline{Y}}-4n_{X\overline{Y}}}{n}$	[Kod01]
3	Causal Confirmed-Confidence	$1 - \frac{1}{2}(\frac{3}{n_X} + \frac{1}{n_{\overline{Y}}})n_{X\overline{Y}}$	[Kod01]
4	Causal Support	$\frac{n_X+n_{\overline{Y}}-2n_{X\overline{Y}}}{n}$	[Kod01]
5	Collective Strength	$\frac{(n_X+n_{\overline{Y}}-2n_{X\overline{Y}})(n_X n_{\overline{Y}}+n_{\overline{X}} n_Y)}{(n_X n_Y+n_{\overline{X}} n_{\overline{Y}})(n_{\overline{X}Y}+n_{X\overline{Y}})}$	[TKS04]
6	Confidence	$1 - \frac{n_{X\overline{Y}}}{n_X}$	[AS94]
7	Conviction	$\frac{n_X n_{\overline{Y}}}{n n_{X\overline{Y}}}$	[TKS04]
8	Cosine	$\frac{n_X - n_{X\overline{Y}}}{\sqrt{n_X n_Y}}$	[Kod01]
9	Dependency	$\lvert\frac{n_{\overline{Y}}}{n} - \frac{n_{X\overline{Y}}}{n_X}\rvert$	[Kod01]
10	Descriptive Confirm	$\frac{n_X - 2n_{X\overline{Y}}}{n}$	[Kod01]
11	Descriptive Confirmed-Confidence/ Ganascia	$1 - 2\frac{n_{X\overline{Y}}}{n_X}$	[Kod01]
12	EII ($\alpha = 1$)	$\sqrt{\varphi \times I^{\frac{1}{2\alpha}}}$	[BKGG03]
13	EII ($\alpha = 2$)	$\sqrt{\varphi \times I^{\frac{1}{2\alpha}}}$	[BKGG03]
14	Example & Contra-Example	$1 - \frac{n_{X\overline{Y}}}{n_X - n_{X\overline{Y}}}$	[GBPP96]
15	F-measure	$\frac{2(n_X - n_{X\overline{Y}})}{n_X+n_Y}$	[vR79]
16	Gini-index	$\frac{(n_X - n_{X\overline{Y}})^2+n^2_{X\overline{Y}}}{n n_X} + \frac{n^2_{\overline{X}Y}+(n_{\overline{Y}}-n_{X\overline{Y}})^2}{n n_{\overline{X}}} - \frac{n^2_Y}{n^2} - \frac{n^2_{\overline{Y}}}{n^2}$	[TKS04]
17	II	$1 - \sum_{k=max(0,n_X-n_Y)}^{n_{X\overline{Y}}} \frac{C^{n_X-k}_{n_Y} C^k_{n_{\overline{Y}}}}{C^{n_X}_n}$	[GBPP96]
18	Implication index	$\frac{n_{X\overline{Y}} - \frac{n_X n_{\overline{Y}}}{n}}{\sqrt{\frac{n_X n_{\overline{Y}}}{n}}}$	[GBPP96]
19	IPEE	$1 - \frac{1}{2^{n_X}} \sum_{k=0}^{n_{X\overline{Y}}} C^k_{n_X}$	[BGGB05a]
20	Jaccard	$\frac{n_X - n_{X\overline{Y}}}{n_Y+n_{X\overline{Y}}}$	[TKS04]
21	J-measure	$\frac{n_X - n_{X\overline{Y}}}{n} log_2 \frac{n(n_X - n_{X\overline{Y}})}{n_X n_Y} + \frac{n_{X\overline{Y}}}{n} log_2 \frac{n n_{X\overline{Y}}}{n_X n_{\overline{Y}}}$	[TKS04]
22	Kappa	$\frac{2(n_X n_{\overline{Y}} - n n_{X\overline{Y}})}{n_X n_{\overline{Y}}+n_{\overline{X}} n_Y}$	[TKS04]
23	Klosgen	$\sqrt{\frac{n_X - n_{X\overline{Y}}}{n}}(\frac{n_{\overline{Y}}}{n} - \frac{n_{X\overline{Y}}}{n_X})$	[TKS04]
24	Laplace	$\frac{n_X+1-n_{X\overline{Y}}}{n_X+2}$	[TKS04]
25	Least Contradiction	$\frac{n_X - 2n_{X\overline{Y}}}{n_Y}$	[AK02]
26	Lerman	$\frac{n_X - n_{X\overline{Y}} - \frac{n_X n_Y}{n}}{\sqrt{\frac{n_X n_Y}{n}}}$	[GBPP96]
27	Lift/Interest factor	$\frac{n(n_X - n_{X\overline{Y}})}{n_X n_Y}$	[PSS00]
28	Loevinger/Certainty factor	$1 - \frac{n n_{X\overline{Y}}}{n_X n_{\overline{Y}}}$	[Loe47]
29	Mutual Information	$\frac{\frac{n_X - n_{X\overline{Y}}}{n} log(\frac{n(n_X - n_{X\overline{Y}})}{n_X n_Y}) + \frac{n_{X\overline{Y}}}{n} log(\frac{n n_{X\overline{Y}}}{n_X n_{\overline{Y}}}) + \frac{n_{\overline{X}Y}}{n} log(\frac{n n_{\overline{X}Y}}{n_{\overline{X}} n_Y}) + \frac{n_{\overline{XY}}}{n} log(\frac{n n_{\overline{XY}}}{n_{\overline{X}} n_{\overline{Y}}})}{min(-(\frac{n_X}{n} log(\frac{n_X}{n}) + \frac{n_{\overline{X}}}{n} log(\frac{n_{\overline{X}}}{n})), -(\frac{n_Y}{n} log(\frac{n_Y}{n}) + \frac{n_{\overline{Y}}}{n} log(\frac{n_{\overline{Y}}}{n})))}$	[TKS04]
30	Odd Multiplier	$\frac{(n_X - n_{X\overline{Y}})n_{\overline{Y}}}{n_Y n_{X\overline{Y}}}$	[LT04]
31	Odds Ratio	$\frac{(n_X - n_{X\overline{Y}})(n_{\overline{Y}} - n_{X\overline{Y}})}{n_{X\overline{Y}} n_{\overline{X}Y}}$	[TKS04]
32	Pavillon/Added value	$\frac{n_{\overline{Y}}}{n} - \frac{n_{X\overline{Y}}}{n_X}$	[TKS04]
33	Phi-Coefficient	$\frac{n_X n_{\overline{Y}} - n n_{X\overline{Y}}}{\sqrt{n_X n_Y n_{\overline{X}} n_{\overline{Y}}}}$	[TKS04]
34	Putative Causal Dependency	$\frac{3}{2} + \frac{4n_X - 3n_Y}{2n} - (\frac{3}{2n_X} + \frac{2}{n_{\overline{Y}}})n_{X\overline{Y}}$	[Kod01]
35	Rule Interest	$\frac{n_X n_{\overline{Y}}}{n} - n_{X\overline{Y}}$	[PS91]
36	Sebag & Schoenauer	$\frac{n_X}{n_{X\overline{Y}}} - 1$	[SS88]
37	Support	$\frac{n_X - n_{X\overline{Y}}}{n}$	[AS94]
38	TIC	$\sqrt{TI(X \rightarrow Y) \times TI(\overline{Y} \rightarrow X)}$	[BGGB05b]
39	Yule's Q	$\frac{n_X n_{\overline{Y}} - n n_{X\overline{Y}}}{n_X n_{\overline{Y}} + (n_Y - n_{\overline{Y}} - 2n_X)n_{X\overline{Y}} + 2n^2_{X\overline{Y}}}$	[TKS04]
40	Yule's Y	$\frac{\sqrt{(n_X - n_{X\overline{Y}})(n_{\overline{Y}} - n_{X\overline{Y}})} - \sqrt{n_{X\overline{Y}} n_{\overline{X}Y}}}{\sqrt{(n_X - n_{X\overline{Y}})(n_{\overline{Y}} - n_{X\overline{Y}})} + \sqrt{n_{X\overline{Y}} n_{\overline{X}Y}}}$	[TKS04]

Table 3.1: A list of objective measures.

N°	Interestingness Measure	Ind.	Equ.	Sym.	Var.	Des.	Sta.
1	Causal Confidence	○	○	○	○	●	○
2	Causal Confirm	○	○	○	●	●	○
3	Causal Confirmed-Confidence	○	○	○	○	●	○
4	Causal Support	○	○	●	●	●	○
5	Collective Strength	●	○	●	●	●	○
6	Confidence	○	●	○	○	●	○
7	Conviction	●	○	○	○	●	○
8	Cosine	○	○	●	●	●	○
9	Dependency	●	○	○	●	●	○
10	Descriptive Confirm	○	●	○	○	●	○
11	Descriptive Confirmed-Confidence/ Ganascia	○	●	○	○	●	○
12	EII ($\alpha = 1$)	●	○	○	●	○	●
13	EII ($\alpha = 2$)	●	○	○	●	○	●
14	Example & Contra-Example	○	●	○	○	●	○
15	F-measure	○	○	●	●	●	○
16	Gini-index	●	○	○	○	●	○
17	II	●	○	○	●	○	●
18	Implication index	●	○	○	○	○	●
19	IPEE	○	●	○	○	○	●
20	Jaccard	○	○	●	●	●	○
21	J-measure	●	○	○	○	●	○
22	Kappa	●	○	●	●	●	○
23	Klosgen	●	○	○	●	●	○
24	Laplace	○	●	○	○	●	○
25	Least Contradiction	○	●	○	●	●	○
26	Lerman	●	○	●	●	○	●
27	Lift/Interest factor	●	○	●	●	●	○
28	Loevinger/Certainty factor	●	○	○	●	●	○
29	Mutual Information	●	○	○	●	●	○
30	Odd Multiplier	●	○	○	●	●	○
31	Odds Ratio	●	○	●	●	●	○
32	Pavillon/Added Value	●	○	○	●	●	○
33	Phi-Coefficient	●	○	●	●	●	○
34	Putative Causal Dependency	○	○	○	○	●	○
35	Rule Interest	●	○	●	●	○	●
36	Sebag & Schoenauer	○	●	○	○	●	○
37	Support	○	○	●	○	●	○
38	TIC	●	○	○	●	●	○
39	Yule's Q	●	○	●	●	●	○
40	Yule's Y	●	○	●	●	●	○

Table 3.2: Some matched properties of objective measures (Ind.: Independency, Equ.: Equilibrium, Sym.: Symmetry, Var.: Variation, Des.: Descriptive, Sta.: Statistical. < ● >: matched, < ○ >: unmatched).

Nature Subject	**Descriptive**	**Statistical**
Equilibrium	– Confidence (6) – Descriptive Confirm (10) – Descriptive Confirmed-Confidence (11) – Example & Contra-Example (14) – Laplace (24) – Least Contradiction (25) – Sebag & Schoenauer (36)	– IPEE (19)
Independence	– Collective Strength (5) – Conviction (7) – Dependency (9) – Gini-index (16) – J-measure (21) – Kappa (22) – Klosgen (23) – Lift (27) – Loevinger (28) – Mutual Information (29) – Odd Multiplier (30) – Odds Ratio (31) – Pavillon (32) – Phi-Coefficient (33) – TIC (38) – Yule's Q (39) – Yule's Y (40)	– EII $\alpha = 1$ (12) – EII $\alpha = 2$ (13) – II (17) – Implication Index (18) – Lerman (26) – Rule Interest (35)
Others	– Causal Confidence (1) – Causal Confirm (2) – Causal Confirmed-Confidence (3) – Causal Support (4) – Cosine (8) – F-measure (15) – Jaccard (20) – Putative Causal Dependency (34) – Support (37)	

Table 3.3: A classification of objective measures.

N°	Formulae
1	$Laplace = \frac{Confidence \times (n \times Support + 1)}{n \times Support + 2 \times Confidence}$
2	$RuleInterest = \frac{n \times Support}{Confidence} \times Pavillon$
3	$Lift = \frac{Confidence}{Confidence - Pavillon}$
4	$Wang = \frac{Support}{Confidence} \times (Confidence - \alpha)$
5	$Gray\&Orlowska = (Lift^k - 1) \times (\frac{Support}{Lift})^m$
6	$Jmeasure = Support \times \log_2(Lift) + (Support - DescriptiveConfirm) \times \log 2(\frac{1}{Conviction})$
7	$JmeasureVariant = Support \times \log_2(Lift)$
8	$Jaccard = \frac{Support}{\frac{Support}{Confidence} + Confidence - Pavillon - Support}$
9	$Loevinger = 1 - \frac{1}{Conviction}$
10	$CausalConfirm = CausalSupport - 2 \times Support + 2 \times DescriptiveConfirm$
11	$CausalConfirmedConfidence = DescriptiveConfirmedConfidence - Confidence + CausalConfidence$
12	$Consine^2 = Lift \times Support$
13	$RuleInterestVariant = \lvert RuleInterest \rvert$
14	$TauxDeLiaison = Lift - 1$
15	$Yule'sQ = \frac{OddsRatio - 1}{OddsRatio + 1}$
16	$Yule'sY = \frac{\sqrt{OddsRatio} - 1}{\sqrt{OddsRatio} + 1}$
17	$LeastContradiction = \frac{DescriptiveConfirm}{Confidence - Pavillon}$
18	$Klosgen = \sqrt{Support} \times Pavillon$
19	$Sebag\&Schoenauer = \frac{Support}{Support - DescriptiveConfirm}$

Table 3.4: Some relations between objective measures

Chapter 4

Postprocessing approaches

Several syntheses on the postprocessing techniques of association rules are introduced in [BVV00] [NS05]. The postprocessing step is considered to be "an important step to help the user discover the useful knowledge nuggets in the huge set of generated association rules". In this chapter we reorganize the principal approaches in five important groups, *including our proposed method with representative measures.*

The first group is to use some constraints on association rules to filter the most interesting rules. Most of them are integrated into the mining process to efficiently reduce the number of rules and time execution (Fig. 1.2 (a) (c)). The second group eliminates the unnecessary or useless ones. The third group summarizes the elements corresponds to the underlying relationships. The fourth group is to group or cluster around the rules with similar property. It will be useful for the domain expert to evaluate all these rules together rather than individually. The last group visualizes the association rules in a graphical way to help the user choose the most interesting ones or to interact with them visually.

4.1 Constraints

4.1.1 Boolean expression constraints

A taxonomy of items (see Fig. 2.8 [Chap. 2] for an example) is used to filter out when a subset of association rules is interested by the user [SVA97]. For instance, those containing at least one item from a user-defined subset of items. Such constraints can be integrated into the mining algorithm (i.e. generalized association rules) as a postprocessing step to reduce dramatically the execution time. These constraints are considered such as boolean expressions[1] over the presence or absence of items in the rules. With a given taxonomy, the boolean expression is of the form *ancestors*(item) or *descendants*(item) rather than just a single item.

As an example inspired from the example illustrated in [SVA97] with a taxonomy (Fig. 2.8 [Chap. 2]), the constraint expression (Emmental $\wedge$ White Bread $\vee$ (**descendants**(Cheese) $\wedge$ $\neg$ **ancestors**(Brown Bread))) will give any rules that are (i) contain both Emmental and White Bread or (ii) contain Cheese or any descendants of Cheese and do not contain Brown Bread or any ascendants of Brown Bread (e.g., Bread).

[1] The boolean expression is in disjunctive normal form (DNF).

4.1.2 Constrained association query

To support a human-centered exploratory mining of association with constraints on attributes, a constrained association query (CAQ) structure is proposed [NLHP98]. The results are optimized via two properties *anti-monotonicity* and *succinctness.*

A CAQ is defined to be a query of the form: $\{(S_1, S_2)|C\}$, where C is a set of constraints on S_1, S_2 and S_1, S_2 are the sets of variables. Note that a CAQ does not make the notion of antecedent and consequent of an association explicitly. Single variable constraints (1-var) are useful in conditioning the antecedent and/or consequent separately, and two variable constraints (2-var) are useful in constraining them jointly. The constraints cover *domain*, *class*, and *aggregation* constraints. A *set variable* is either an identifier of the form S or is an expression of the form $S.A$, where A is an attribute in the minable view. A frequency constraints of the form $freq(S_i)$ saying that that the support of S_i must exceed some given threshold.

For instance, assume that the minable view to be *trans*(TID, Itemset), *itemInfor*(Item, Type, Price).

1-var constraints: $S \subset Item$ says that S is a set variable on the $Item$ domain, $S.price \leq 100$ says all items in S are of price less than or equal to \$100, $\{snacks, sodas\} \subseteq S.Type$ says S should include some items whose type is snacks and some items whose type is sodas, $S.Type \cap \{snacks, sodas\} = \emptyset$ says S should exclude such items.

2-var constraints: $S_1.Type \cap S_2.Type = \emptyset$, $max(S_1.Price) \leq avg(S_2.Price)$, $\{(S_1, S_2)|S_1 \subset Item\&S_2 \subset Item\&count(S_1) = 1\&count(S_2) = 1\&freq(S_2)\}$ asks for all pairs of single items satisfying frequency constraints.

4.1.3 Minimum improvement constraints

Motivated by the principle of Occam's Razor (i.e. plurality should not be posited without necessity) a minimum improvement constraints approach is proposed [BAG00]. With a fix consequent, the user can eliminate unnecessary rules by specifying a *minimum improvement* constraint. The necessary rules are defined as those have confidence greater than the confidence of any of its simplifications. A rule $X_s \rightarrow Y$ is said a simplification of a rule $X \rightarrow Y$ if $X_s \subset X$. Then the *improvement* of a rule $X \rightarrow Y$ is defined as:

$$imp(X \rightarrow Y) = MIN(\{conf(X \rightarrow Y) - conf(X_s \rightarrow Y)|X_sYX\}), \forall X_S \subset X$$

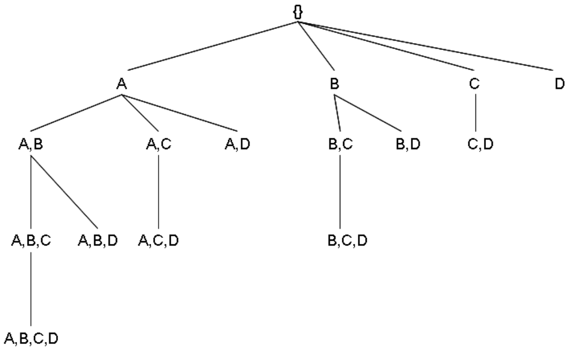


Figure 4.1: Rymon's set-enumeration tree for $U = \{A, B, C, D\}$.

A rule is less predictive power or undesirable when it has negative improvement. When the improvement of a rule is positive, it then considered as a desirable constraint. Given U is the set of all items that are not in the consequent. With the power set of U, rules which satisfy the minimum support, confidence and improvement constraints will be hold. Rymon's set-enumeration tree (see Fig. 4.1) [Rym92] is used to present the subset search problem of U efficiently.

4.1.4 Action hierarchy

By using the actionable definition (Sec. 3.1.1 [Chap. 3]), a taxonomy of action or an action hierarchy is proposed as constraints on association rules [AT97]. A hierarchy of actions is illustrated from more general actions at the top of the hierarchy to more specific actions at the bottom (see Fig. 4.2). The actions can be described in several stages: incremental or step-by-step.

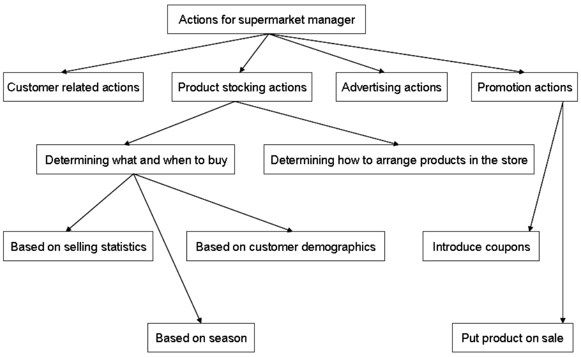

Actions for supermarket manager
Customer related actions
Product stocking actions
Advertising actions
Promotion actions
Determining what and when to buy
Determining how to arrange products in the store
Based on selling statistics
Based on customer demographics
Introduce coupons
Based on season
Put product on sale

Figure 4.2: Fragment of an action tree for the supermarket management.

When the action tree is completed, the actionable pattern corresponding for an action on the tree will be assigned (i.e. by the data mining queries or a pattern template [KMR$^+$94]). For example, the node "*Based on customer demographics*" of the tree in Fig. 4.2 can be assigned a query [KMR$^+$94]

$$ChildrenAge* \rightarrow Category(0.5, 0.01)$$

to find all rules representing the product categories that the customer with children are buying.

4.1.5 AND-OR taxonomy

A knowledge-based approach with an AND-OR taxonomy (AO-taxonomy as short) as constraints, is developed to mine generalized association rules (Sec. 2.3.5 - [Chap. 2]) [Sub98]. An AO-taxonomy has a single root and contains AND-nodes and OR-nodes (Fig. 4.3). An AND-taxonomy (OR-taxonomy) is a node has enough support if all (at least τ) of its child nodes have support $\geq$ minsup. An AND-node is represented by the symbol $\bullet$, $(+, \tau = c)$ for OR-node. The complement of a node $\neg X$ is hold if supp$(\neg X) \geq$ minsup.

4.1.6 Minimal set of unexpected patterns

A set of constraints is developed to find a minimal set of unexpected patterns [PT00] [PT06]. From the definition of unexpectedness [PT98], a rule $X \rightarrow Y$ is said to be *unexpected* with respect to the belief $X_b \rightarrow Y_b$ if it satisfies all the following conditions: (i) $Y \cap Y_b = \emptyset$, (ii) supp$(X \cap X_b) \geq$ minsup, (iii) the rule $XX_b \rightarrow Y$ hold (the rule $XX_b \rightarrow \neg Y_b$ also hold logically).

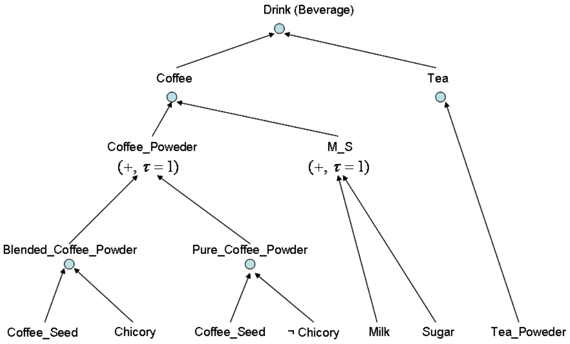


Figure 4.3: An AND-OR taxonomy.

Y is the *minimal set* of X if and only if the following conditions hold: (i) $Y \subseteq X$, (ii) $\forall x_i \in X, \exists y_i \in Y$ such that $y_i \models_M x_i$, (iii) $\forall y_1, y_2 \in Y, y_1 \nvDash_M y_2$. The operator $\models_M$ for a couple of rules $X_1 \rightarrow Y_1 \models_M X_2 \rightarrow Y_2$ if $X_2 \models X_1$ and $Y_1 = Y_2$ [2].

For instance, with the belief *diaper* → *beer* let the set of all unexpected patterns be $\{diaper \wedge weekday \rightarrow not\_beer$, $diaper \wedge unemployed \rightarrow not\_beer$, $diaper \wedge weekday \wedge unemployed \rightarrow not\_beer$, $weekday \rightarrow not\_beer$, $unemployed \rightarrow not\_beer\}$. The minimal set of unexpected patterns in this case is $\{weekday \rightarrow not\_beer$, $unemployed \rightarrow not\_beer\}$.

4.1.7 Interestingness preprocessing

The Interestingness Preprocessing Step (IPS) [Sah01] eliminates uninteresting rules independently of the domain, user and task. It can be integrated into the interestingness process as in Fig. 4.4. It simply filters out the unnecessary rules but does not consider the outputted rules as potentially interesting or having interestingness. Two IPS methods are proposed by considering the influence of confidence value:

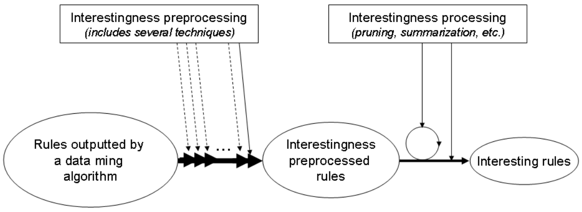

Interestingness preprocessing
(includes several techniques)
Interestingness processing
(pruning, summarization, etc.)
Rules outputted by a data ming algorithm
...
Interestingness preprocessed rules
Interesting rules

Figure 4.4: IPS in the interestingness process.

- IPS1 (Overfitting). The method tries to delete any rule $r = X \rightarrow Y$ if there exists a rule $r' = X' \rightarrow Y'$ such that: (i) $X' \subset X$, (ii) $Y' = Y$, and (iii) confidence$(r') \geq$ confidence(r), support$(r') >$ support(r).
- IPS2 (Transition). The method tries to to delete any rule r, $r = X \rightarrow Y$, for which $\exists r_2, r_3$ such that (i) $r_2 = X \rightarrow Y \wedge Z$, and (ii) $r_3 = Y \rightarrow Z$ where confidence(r_3)=1.

[2] $\models$: logical implication, $\models_M$ logical implication in the sense of minimality.

4.1.8 Negative association rules

Negative association rules is introduced to find what items a customer is not likely to buy [SON98] [BBJ00]. A statement "60% of the customers who buy potato chips do not buy bottled water" is an example of this type of rule.

The domain knowledge is used in form of a taxonomy for grouping similar items. One of the fundamental assumptions in this approach is that uniformity assumption, i.e., the items that belong to the same parent in a taxonomy are expected to have similar types of associations with other items. By considering only those cases where an expected support can be computed based on the taxonomy, the class of negative rules generated is not all possible negative rules but a subset of those rules. If additional domain knowledge is incorporated, it may be possible to mine additional types of negative rules.

A negative rule $X \not\rightarrow Y$ is measured by an interestingness measure $RI = \frac{\varepsilon[sup(X \cup Y)] - sup(X \cup Y)}{sup(X)}$, where $\varepsilon[sup(X \cup Y)]$ is the *expected support* of an itemset. The expected support value (i.e. based on a taxonomy) is calculated as:

$$\varepsilon[sup(p,q,r,...,t)] = \frac{sup(p \cup q \cup r' \cup \cdots \cup t') \times sup(r) \times \cdots sup(t)}{sup(r') \times \cdots sup(t')}$$

where $p', q', ..., r'$ are siblings of $p, q, ..., t$ respectively.

Given a database of customer transactions $\mathcal{D}$ and a taxonomy $\mathcal{T}$ on the set of items, find all rules $X \not\rightarrow Y$ such that: (i) sup(X) $\geq$ minsup and sup(Y) $\geq$ minsup, and (ii) $RI(X \not\rightarrow Y) \geq minRI$. *minsup* and *minRI* are both specified by the user.

4.1.9 Substitution

A model based on the idea of understanding the choice made by consumers which corresponding to the purchase of some items instead of others, is called *substitution* rules [THC02] [THC05].

To find *concrete* itemset, some complement items will be added to an itemset. For example with six items {A,B,C,D,E,F}, the itemset {BF} will become $\{\overline{A}B\overline{C}\overline{D}\overline{E}F\}$. A positive frequent itemset [THC05] $X = \{x_1, x_2, ..., x_k\}$ is called a concrete itemset if and only if (i) $k = 1$, or (ii) $k \geq 2$, $S_X > \prod_{x_i \in X} S_{x_i}$ and $Chi(X) \geq \chi^2_{df(X),\alpha}$, where $\prod_{x_i \in X} S_{x_i}$ corresponds to the expected support for itemset X and $\chi^2_{df(X),\alpha}$ is the value of chi-square distribution with degree of freedom $df(X)$ at probability α.

Given two itemsets X and Y, X is a substitute for Y, denoted by $X \rhd Y$, if and only if (i) both X and Y are concrete, (ii) X and Y are negatively correlated, and (iii) the negative association rule $X \rightarrow \overline{Y}$ is valid.

4.1.10 Cyclic/Calendric

The transaction model [AMS$^+$96] is added a time attribute that describes the time when the transaction was executed [ORS98]. This method is to find association rules that display regular hourly, daily, weekly, etc., variation that has the appearance of cycles. For example, such a rule $(Day = monday) \cup X \rightarrow Y$.

A cycle c [ORS98] is a tuple (l, o) consisting of a length l (in multiples of the time unit) and an offset o (the first time unit in which the cycle occurs), $0 \leq o \leq l$. An association rule that has a cycle is referred as *cyclic.*

An extended version of cyclic, called *calendric association rules* [RMS98], using a *calendar algebra* to describe complicated temporal phenomena of interest to the user.

4.1.11 Non-actionable

A significant rule or a non-redundant rule is not potentially useful for action so finding actionable rules is still a major problem. Instead of finding such actionable rules, an approach is proposed to find rules that are not actionable [LHM01]. A rule is said non-actionable rule if is not a potentially actionable rule. A rule is said potentially actionable rule if (i) it does not have any descendant rules, or (ii) it still significant with respect to "$\rightarrow Y$" after removing the data tuples that can be covered by its descendant potentially actionable rules. A rule $R' : X' \rightarrow Y'$ is a descendant rule of a rule $R : X \rightarrow Y$ (R is also called an ancestor rule of R') if we have $Y = Y'$ and $X \subseteq X'$.

4.1.12 Softness

To handle interestingness on mined patterns, an approach on soft constraints is proposed in [BB05]. The constraints are represented flexibly or as a "soft function". λ-interesting and top-k are two different problems are considered in this technique.

Based on the fuzzy semiring, a soft constraint C [BB05] on itemsets is defined by a quintuple $< Agg, Att, \theta, t, \alpha >$, where (i) $Agg \in \{supp, min, max, count, sum, range, avg, var, median, std, md\}$, (ii) agg will calculate on the attribute (Att), (iii) $\theta \in \{\leq, \geq\}$, (iv) $t \in \Re$ the center of the interval (see Fig. 4.5), associated with the semiring value 0.5, (v) $\alpha \in \Re^+$ is the softness parameter or the width of the interval.

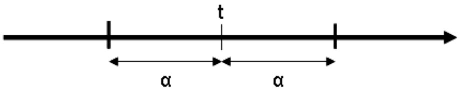


Figure 4.5: Softness interval.

4.2 Pruning

4.2.1 Rule cover

A set of rules can be pruned by a rule cover with a user- and domain- independent way [TKR+95]. A rule cover[3] is a subset of the original set of rules such that the cover matches all the rows that the original set matches. Consider a collection Γ of rules with the same consequent Y:

$$\Gamma = X_i \rightarrow Y | i = 1, ..., n$$

A subset $\Delta \subseteq \Gamma$ is a rule cover, if

$$\bigcup_{X \rightarrow Y \in \Gamma} m(XY) = \bigcup_{X \rightarrow Y \in \Delta} m(XY)$$

Δ is called a *structural rule cover* for Γ if there is no rule $X^{'} \rightarrow Y$ such that $X^{'} \in X$ for all rules $X \rightarrow Y \in \Delta$. A structural cover is a cover and contains the most general rules of the original ruleset. Given two rules $R_1 : AB \rightarrow C$ and $R_2 : ABD \rightarrow C$. R_2 has no additional predictive information than R_1 so one can prune R_2 from the structural rule cover.

[3] The usage of the term *cover* is borrowed from database theory.

4.2.2 Maximum entropy

An approach based on maximum entropy for removing redundant association rules is proposed in [JS02]. A constraint C on a set of attributes $\mathcal{I}$ is a pair $C = (I, p)$ where $I \subseteq \mathcal{I}$, $p \in [0, 1]$. The set of constraints generated by an association rule $X \to Y$ is defined as:

$$\mathcal{C}(X \to Y) = \{(X, sup(X)), (X \cup Y, sup(X \cup Y))\}$$

The maximum entropy distribution is computed by using the Generalized Iterative Scaling (GIS) algorithm [Rat97]. Let $\mathcal{C} = \{C_1, C_2, ..., C_n\}$ be a set of constraints, where $C_k = (I_k, p_k), 1 \leq k \leq n$. GIS proceeds by assigning some initial values to each probability in $P^{\mathcal{C}}$, and iteratively updating them until all the constraints are satisfied. Let $P^{\mathcal{C}(i)}$ denote the distribution after i iterations. Updating in each iteration is performed according to the formula (assuming that $\frac{0}{0} = 0$):

$$P^{\mathcal{C}(i+1)} = P^{\mathcal{C}(i)} \prod \left[\frac{p_k}{P^{\mathcal{C}(i)}(I_k)} \right]^{\frac{1}{C}}$$

Let $\mathcal{C} = \{(X, sup(X)), (Y, sup(Y)), (X \cup Y, sup(X \cup Y))\}, X, Y \subset \mathcal{I}, X \cap Y = \emptyset$ be a set of constraints. The maximum entropy distribution induced by $\mathcal{C}$ is given in Tab. 4.1.

$P^{\mathcal{C}}(X \to Y)$	Y	$\overline{Y}$
X	$\frac{sup(X \cup Y)}{n_{X \cup Y}}$	$\frac{sup(X) - sup(X \cup Y)}{n_X - n_{X \cup Y}}$
$\overline{X}$	$\frac{sup(Y) - sup(X \cup Y)}{n_Y - n_{X \cup Y}}$	$\frac{1 - sup(X) - sup(Y) + sup(X \cup Y)}{\lvert\mathcal{I}\rvert - n_X - n_Y + n_{X \cup Y}}$

Table 4.1: Maximum entropy induced.

4.2.3 Relative interestingness

The interestingness of a given rule can be captured through the amount of change in information relative to the common sense rules [HLL00][HLSL00]. Tab. 4.2 shows a rule structure with exception [Suz97]. Exceptions are usually minority, either unknown/new or omitted. They represent a way that contradicts the common belief. A common sense rule represents a common phenomenon that comes with high support and confidence in a particular domain while an exception rule is weak in terms of support but having high confidence similar to those common sense rule. A weak rule of low support may not be a reliable.

The interestingness of a rule can be considered as a function with three parameters: support, confidence and the knowledge about common sense rules. The knowledge of a rule $XX' \to Y$ is composed of the knowledge of the two rules $X \to Y$ and $X' \to Y$.

$X \to Y$	common sense rule
$X' \to \neg Y$	reference rule
$XX' \to \neg Y$	exception rule

Table 4.2: Rule structure for exceptions.

4.2.4 Interestingness-based filtering

Interestingness criteria are used to select only the most interesting rules [AT01]. Subjective and objective measures can be used as a part of this approach and the validation system can support different interestingness criteria. The domain expert can specify interestingness-based filters using a syntax similar to the syntax of the template-based filters (Sec. 4.2.8). For instance, the following statement specifies that all the high gain and unexpected rules should be accepted.

$$\mathbf{ACCEPT} : \mathbf{INTERESTINGNESS}\{\mathbf{gain} > \mathbf{0.5}, \mathbf{unexpected}\}$$

4.2.5 Ranking subset of contingency table

By looking at a small set of contingency tables, a whole set of association rules can be ranked with a desirable measure [TKS04]. A smaller set of contingency tables is provided to the expert for ranking and this information is used to determine the most appropriate measure. A small subset of contingency tables have to optimize the two following criteria: (i) small enough to rank manually, (ii) large enough to choose to best measure, (iii) diverse enough. The subset is then computed by the RANDOM or DISJOINT algorithms.

Fig. 4.6 shows the whole process of this technique. Let T be the set of all contingency tables and S be the tables selected by a subset selection algorithm. A subset that minimizes the difference between the similarity matrices computed from the subset S_S (i.e. by using the Pearson's cofficient) and the entire set of contingency tables S_T will be selected. The subset is then considered as a good representative for T. The difference is determined as:

$$d(S_S, S_T) = max_{i,j}\ |S_T(i,j) - S_S(i,j)|$$

4.2.6 Rule template

Rule templates [KMR+94], also called meta-rules, that describe the structure of interesting rules. A set of rules is described by the template. Which items in the antecedent and the consequent is determined clearly. A template is an expression

$$A_1, ..., A_k \rightarrow A_{k+1}$$

where A_i is an item, a class or an expression. A rule is considered interesting if it matches an *inclusive* template, uninteresting if it matches a *restrictive* template.

Another specification language [LHWC99] allows to specify three levels of domain knowledge: general impressions (i.e. a more generalized form of templates [LHC97]), reasonably precise concepts and precise knowledge. This meta-knowledge allows them to classify the discovered association rules into two categories: conforming and unexpected rules.

4.2.7 Uninteresting rules

A subjective approach is introduced in [Sah99] to find rules that are *not interesting* by eliminating a substantial portion of uninteresting association rules discovered. The approach can be incorporated into the mining process effectively [Sah02b]. The user can interact with a few and simple classification questions. Four possible classifications are given in Tab. 4.3 For example, the rule $Husband \rightarrow Married$ is classified as TNI because the word "husband" has the meaning as a male partner in marriage.

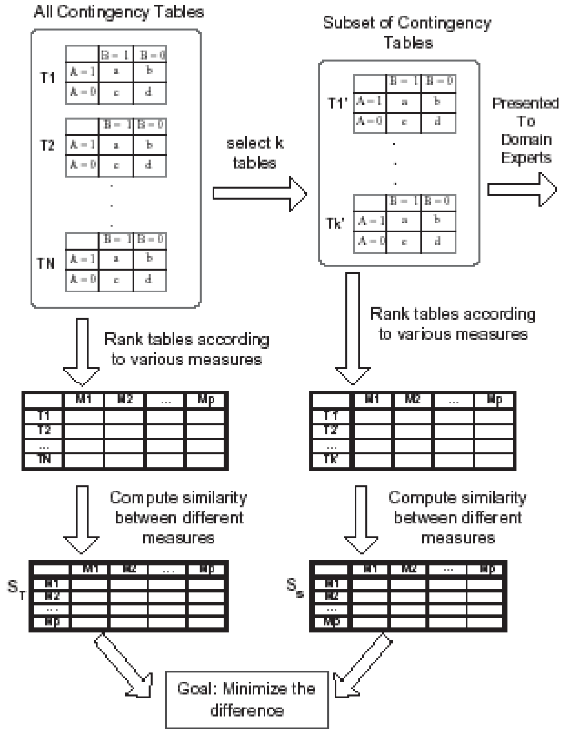

All Contingency Tables
T1
T2
TN
B = 1
B = 0
A = 1
A = 0
a
b
c
d
select k tables
Subset of Contingency Tables
T1'
Tk'
Presented To Domain Experts
Rank tables according to various measures
Rank tables according to various measures
M1
M2
...
Mp
T1
T2
...
TN
Tk'
Compute similarity between different measures
Compute similarity between different measures
S_T
S_s
Goal: Minimize the difference

Figure 4.6: Ranking subset of contingency tables.

4.2.8 Template-based rule filtering

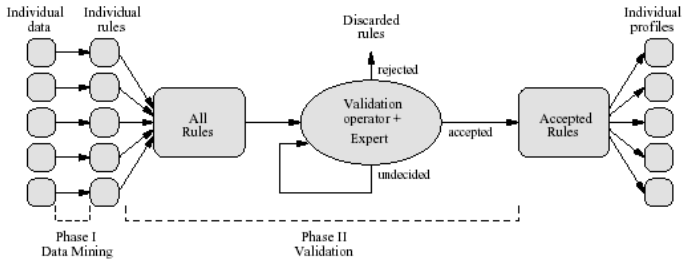


Figure 4.7: The profile building process.

An approach called template-based rule filtering operator is developed in [AT01]. Fig. 4.7 shows a model to illustrate. It allows the expert to specify in general terms the types of rules that he or she either wants to accept (*accepting* template) or reject (*rejecting* template). After a template is specified, unvalidated rules are "matched" against it. Rules that match an accepting template are accepted and put into user profiles, and rules that match a rejecting template are rejected. Rules that do not match a template remain unvalidated.

The operator is formally defined as a language with various constraints that the expert can impose on (Fig. 4.8): (i) the syntactic structure of the body (antecedent) and the head (consequent)

Rule type	Rule meaning
TNI	True-Not-Interesting
NTI	Not-True-Interesting
NTNI	Not-True-Not-Interesting
TI/I	True-Interesting/Interesting

Table 4.3: Rule classifications with a subjective measure.

of the rule, (ii) basic statistical parameters of the rule, (iii) the factual information about a user for whom the rule was discovered.

template	→	action : tmpl_expression
action	→	**ACCEPT** $\mid$ **REJECT**
tmpl_expression	→	atom_tmpl $\mid$ atom_tmpl logic_oper tmpl_expression
atom_tmpl	→	inverse pos_atom_tmpl
logic_oper	→	**AND** $\mid$ **OR**
inverse	→	ϵ $\mid$ **NOT**
pos_atom_tmpl	→	rule $\mid$ stats $\mid$ facts
rule	→	rule_part set_oper { trans_term_list }
rule_part	→	**BODY** $\mid$ **HEAD** $\mid$ **RULE**
stats	→	**STATS** { stat_term_list }
facts	→	**FACTS** { fact_term_list }
set_oper	→	$= \mid \neq \mid \subset \mid \subseteq \mid \supset \mid \supseteq$
trans_term_list	→	trans_term $\mid$ trans_term , trans_term_list
trans_term	→	attr_term $\mid$ aggr_attr_term
attr_term	→	attr_name $\mid$ attr_name compar_oper value $\mid$ attr_name = value_set
stat_term_list	→	stat_term $\mid$ stat_term , stat_term_list
stat_term	→	stat_name $\mid$ stat_name compar_oper stat_value
stat_name	→	**supp** $\mid$ **conf**
...	...	...

Figure 4.8: A fragment of the template specification language.

4.2.9 Merging

Adjacent intervals of numeric values are merged, in a bottom-up manner with a modification of a B-tree (M-tree, Fig. 4.9), to maximize the interestingness of a set of association rules [WTL98]. The user specifies a template of the form $C_1 \rightarrow C_2$ where C_1 and C_2 are conjunctions of uninstantiated and instantiated attributes: (i) each attribute appears at most once in the template, (ii) C_2 contains only categorical attributes, and (iii) all numeric attributes are uninstantiated.

4.2.10 Correlation between objective measures and real human interest

An approach is proposed to compute the correlation between the interestingness values issued from objective measures and the real human interest [CFE05]. Given a ruleset, each rule is assigned an interestingness value. For each measure, all the discovered rules are ranked according to the value

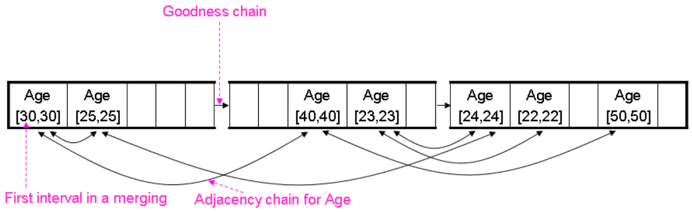


Figure 4.9: The leaf level of the M-Tree.

of that measure (from best value to worst value). Then an average rank is computed for each rule by calculating the average rank value of each rule. See Tab. 4.4 for an example with five measures and nine rules.

Measures Rules	Ranking					
	m_1	m_2	m_3	m_4	m_5	Average
r_1	1	2	1	1	7	2.4
r_2	2	3	2	2	3	2.4
r_3	3	1	4	3	5	3.2
r_4	4	4	3	4	4	3.8
r_5	5	6	5	5	2	4.6
r_6	6	7	6	7	1	5.6
r_7	7	5	7	6	6	6.2
r_8	6	7	5	6	7	6.2
r_9	5	7	7	6	6	6.2

Table 4.4: An example of average ranking.

The user is shown a set of 9 rules consists of three groups, each group has 3 rules: (i) the lowest rank (highest interestingness values), (ii) the median ranks (average interestingness values), and (iii) the highest ranks (lowest interestingness values). These 9 rules will then be assigned a subjective degree of interestingness from the user. The correlation between the ranks and the subjective degree of interestingness of these selected rules is then calculated by the Pearson's coefficient.

4.3 Summarization

4.3.1 Direction setting

After pruning insignificant rules, a special subset of association rules that represents the underlying relationship in the data is summarized [LHM99]. This subset is called *direction setting* (DS) rules (see Fig. 4.10). DS rules are the positively correlated association rules that set the direction for non-direction setting ($\overline{DS}$) rules to follow. The essential aspects or relevant details can be viewed efficiently.

4.3.2 GSE pattern

An approach with GSE patterns (i.e. General rule, Summary & Exception) is introduced in [LHH00]. A GSE pattern consists of three components: a single general rule (an if-then rule), a summary and

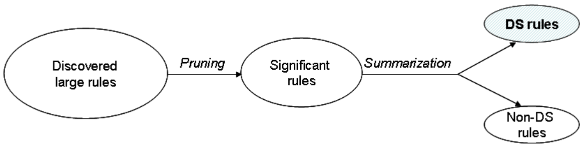


Figure 4.10: The process of finding DS rules.

a set of exceptions. A general relation is illustrated by the general rule. Some important information is highlighted by the summary. Unexpected rules with respect to the general rule is given in the set of exceptions. It has the following form:

$X \rightarrow c_i$ (sup, conf)
Summary
Except $E_1, ..., E_n$

4.4 Grouping

4.4.1 Differentiating with support values

After have put the rules with the same consequent Y into a cover (Sec. 4.2.1), the set of rules in the cover may still be quite large. Each corresponding set of rules (i.e. in a cover) can be made more understandable by restructuring or ordering based on their interestingness values issued from the support measure [AIS93a]. Such a method [TKR$^+$95] calculates a different interval in support values to grouping the rules .

General speaking, the distance between between two rules $X \rightarrow Y$ and $X' \rightarrow Y'$ is defined as the amount of rows in the set of transactions to differ the two rules:

$$d(X \rightarrow Y, X' \rightarrow Y') = sup(XY) + sup(X'Y') - 2 * sup(XYX'Y')$$

4.4.2 Neighborhood

An approach in terms of neighborhood-based unexpectedness by using a syntax-based distance is proposed in [DL98]. The distance function is defined to give different scales of importance for different parts of rules to differentiate. It differentiates three parts: (i) the symmetric difference of all items in the two rules, (ii) the symmetric difference of the antecedent of the two rules, (iii) the symmetric difference of the consequent. The itemset distance between two rules $R = X \rightarrow Y$ and $R' = X' \rightarrow Y'$ – given three positive real numbers δ_1, δ_2, δ_3 – is computed as:

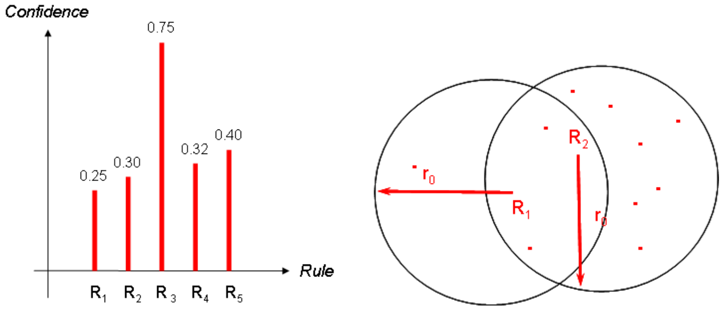
$$d(R, R') = \delta_1 * |(XY) \ominus (X'Y')| + \delta_2 * |X \ominus X'| + \delta_3 * |Y \ominus Y'|$$

where $\ominus$ denotes the symmetric difference between two elements (e.g. $X \ominus Y = X - Y \cup Y - X$). The user can give its preferences by giving the values of δ_1, δ_2, δ_3.

An r-neighborhood $\mathcal{N}(R, r)$ of a rule R ($r > 0$) is the following set of rules:

$$\mathcal{N}(R, r) = \{R_i | d(R_i, R) \leq r, R_i : \text{a potential rule}, i > 0\}$$

Three techniques are proposed to capture interesting rules: (i) unexpected confidence, (ii) sparse neighborhoods, and (iii) collection of rules. Suppose M is the set of rules and R is a rule in M and $r > 0$.



(a) Unexpected confidence (b) Sparse neighborhoods

Figure 4.11: Interesting rules with r-neighborhood.

- A rule R is considered interesting with the unexpected confidence in its r-neighborhood if

$$||conf(R) - avg(R, r)| - std(R, r)| \gg 0$$

where avg(R, r), std(R, r) are the average confidence and the standard deviation of the set $M \cap N(R, r) - \{R\}$ respectively. Fig. 4.11 (a) gives an example with five rules $R_1, ..., R_5$. We have avg$(R_3, r) = 0.3175$, sdt$(R_3, r) = 0.026$. So conf(R_3) - avg$(R_3, r) = 0.4325 \gg$ std(R_3, r)=0.026.

- A rule R is considered interesting (i.e. of the isolated type) if it has an unexpectedly sparse *r-neighborhood*: $\mathcal{N}(R_0, r)$ is large but $|M \cap N(R_0, r)|$ is relatively small. Fig. 4.11 (b) shows that the r_0-neighborhood of R_1 and R_2 can both be sparse neighborhoods. R_1 is more sparse than R_2 if $\mathcal{N}(R_1, r_0 = \mathcal{N}(R_2, r_0)$.

- Given two positive numbers $r_0 < r_1$. The r_0-neighborhood of R_0 has unexpected confidence in its r_1-neighborhood if (i) std$(M \cap \mathcal{N}(R_0, r_0))$ is small and (ii) $\mathcal{N}(R_0, r_0)$ is much larger or smaller than avg$(M \cap (R_0, r_1) - \mathcal{N}(R_0, r_0))$.

4.4.3 Triple factors

Based on the agglomerative hierarchical clustering method for exploring interestingness, a new similarity measure between two rules is proposed [Sah02a]. The measure integrates the features discussed in the two sections Sec. 4.4.1 and Sec. 4.4.2. It seems to be more "natural" according to the author because it can differentiate naturally two rules by a triple of factors such as their antecedents, consequents and attributes. The difference between two rules $R = X \to Y$ and $R' = X' \to Y'$ is calculated as follows:

$$\begin{aligned} d(R, R') = \quad & [1 + diff(X, X')] \tfrac{|X \oplus X'|}{|X \cup X'|} \gamma_1 + \\ & [1 + diff(Y, Y')] \tfrac{|Y \oplus Y'|}{|Y \cup Y'|} \gamma_2 + \\ & [1 + diff(X \cup Y, X' \cup Y')] \tfrac{|(X \cup Y) \oplus (X' \cup Y')|}{|X \cup Y \cup X' \cup Y'|} \gamma_3 \end{aligned}$$

where *diff* is a measure inspired from Sec. 4.4.1 to calculate the different proportion in supporting each portion in a rule:

$$diff(X, Y) = sup(X) + sup(Y) - 2 * sup(X \cup Y)$$

Also inspired from Sec. 4.4.2, the occurrences of the three parameters γ_1, γ_2 and γ_3 reflect user preferences.

4.4.4 Average distance

To select the rules that have the highest predictive power, a method is proposed to select the k rules that have the highest average distance between them [GB98]. The most heterogeneous set of rules that is possible to has high predictive capabilities in the assumption (i.e. antecedent of a rule) will be able chosen. Unlike the other works [KS96] [PS91] [MM95] [SA95] [ST96] [LHML99], the authors do not try to measure the interestingness of a rule. The distance between two rules $R = X \rightarrow Y$ and $R' = X' \rightarrow Y'$ is based on three attribute factors: (i) the number of attributes (i.e. items) in one rule and absent in the other, (ii) the number of attributes in both rules with overlapping[4] values, and (iii) the number of attributes in both rules with values slight or null overlapping.

$$d(R, R') = \begin{cases} \frac{\alpha_1\gamma_1 + \alpha_2\gamma_2 - \alpha_3\gamma_3}{\gamma_4} & \text{if } \beta = 0 \\ 2 & \text{otherwise} \end{cases}$$

where γ_1 is the number of attributes in rule R and not in rule R' plus the number of attributes in rule R' and not in rule R, γ_2 is the number of attributes both in rule R and in rule R', but with slightly overlapping values (i.e. an overlapping is considered if it is bellow 66%), γ_3 is the number of attributes in both rules, with overlapping values (i.e. an overlapping is considered if it is above 66%), γ_4 is the number of attributes in rule R plus the number of attributes in rule R', and β is the number of attributes both in rule R and in rule R' but with non overlapping values.

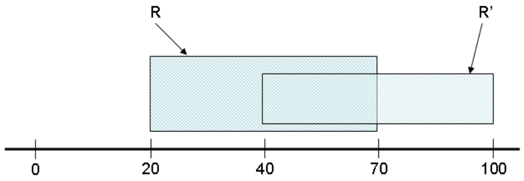


Figure 4.12: Overlapping of an attribute.

Fig. 4.12 shows a case when an attribute has a overlapping zone with both R and R'. The overlapping zone is an interval from 40 to 70. The overlapping percent for R is $\frac{70-40}{70-20} = \frac{30}{50} = 60\%$.

The parameters γ_1, γ_2 and γ_3 are weighted by constants α_1, α_2 and α_3. The values of these three constants are chosen as $\alpha_1 = 1$ and $\alpha_2 = \alpha_3 = 2$ to limit the measure value between -1 and 1. If the two rules have few attributes in common then the measure value is 1. The measure value is 1 indicating that the rules overlap strongly. It returns a value of 2 when the rules do not overlap absolutely.

4.4.5 Attribute hierarchy

A similarity measure using the notion of *attribute hierarchies* (i.e. a taxonomy of items) to find similar rules among all the discovered rules is proposed [AT01]. At the beginning, the human expert organizes an hierarchy of attributes/items. All the items are contained in the leaf nodes. The leaf nodes are then combined into some non-leaf nodes or *aggregated attributes*. Leaf-nodes and non-leaf nodes can also be combined into non-leaf nodes recursively (see Fig. 4.13 for an example).

The rules is then grouped by the following steps: specifying *rule aggregation level*, obtaining *aggregating rules*, and *grouping rules*.

[4] This technique is very useful with the quantitative association rules.

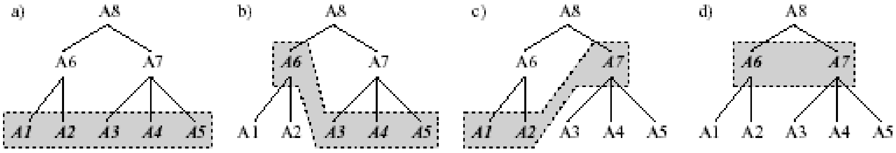


Figure 4.13: An attribute hierarchy.

4.4.6 Probability distance

A CMPB (Conditional Market-Basket Probability) distance between two rules $R = X \rightarrow Y$ and $R' = X' \rightarrow Y'$, based on a conditional probability, is proposed in a new agglomerative clustering technique using multi dimensional scaling (MDS) and self organizing map (SOM) [GSG99]:

$$\begin{aligned} d(R, R') &= p(\overline{(X \cup Y)} \vee \overline{(X' \cup Y')} | (X \cup Y) \vee (X' \cup Y')) \\ &= 1 - \frac{sup(X \cup Y \cup X' \cup Y')}{sup(X \cup Y) + sup(X' \cup Y') - sup(X \cup Y \cup X' \cup Y')} \end{aligned}$$

4.5 Visualizing

4.5.1 Rule visualizer

RuleVisualizer is a prototype tool for visualizing the association rules with the aid of templates [KMR+94]. The tool consists of three components: selection, browsing and graph. The first component (Fig. 4.14 (a)) specifies the criteria for rules to be presented: support, confidence and commonness. The other two components (Fig. 4.14 (b)(c)) present rule visually. A single rule is presented by a bar graph: the leftmost and rightmost bars represent the confidence and support of the rule while the middle bar represents the commonness. Several rules can be visualized simultaneously.

4.5.2 Human-centered rummaging

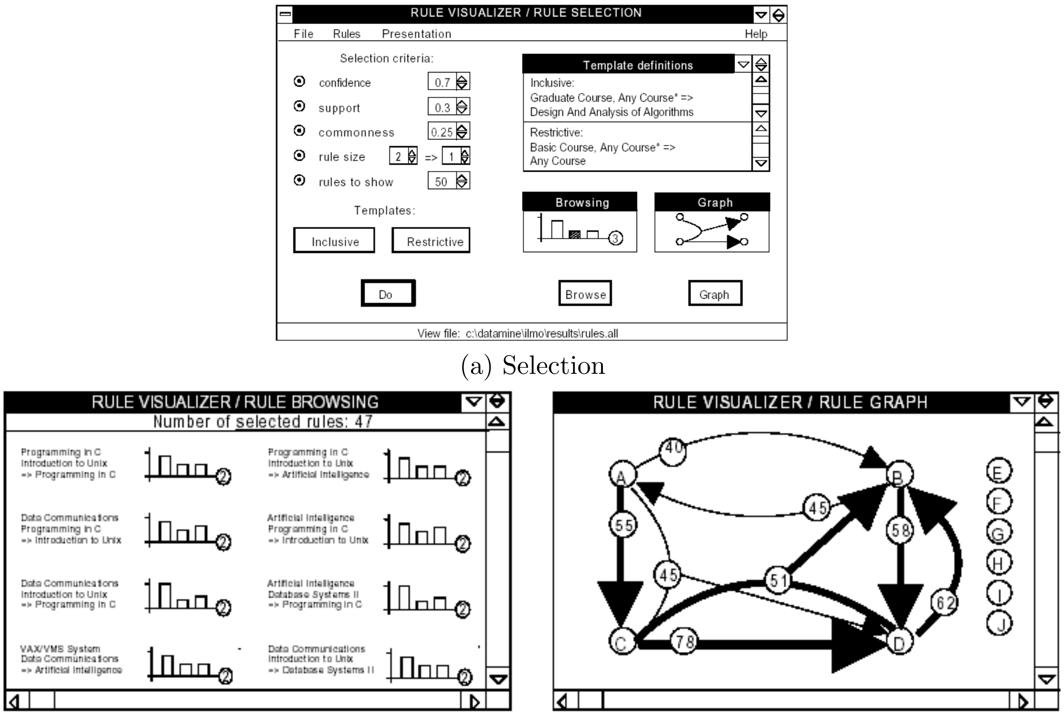
A rummaging model is proposed to let the user navigate/interacts as he/she wishes through the voluminous ruleset [BGB03b] [BGB03a]. It focuses on the successive limited subsets of association rules to explore (see Fig. 4.16). A series of local explorations by trial and error through the whole ruleset from which only the selected portion is gradually visited. The rules are grouped by subsets and combined by the neighborhood relation (see Fig. 4.15).

4.6 Representative measures

Instead of selecting only one measure as discussed in the approach of ranking a subset of contingency tables [Sec. 4.2.5], a set of objective measures called representative measures, representing different point of views on the datasets are captured [HGB06b]. Our approach is a pruning technique [Sec. 4.2] for the postprocessing of association rules.

Three data analysis techniques are used to illustrate: agglomerative hierarchical clustering (AHC), partitioning around medoids (PAM) [KR90], and correlation graph (CG). Each of these techniques is used as a means for achieving the results.

These three techniques are used with a $q \times q$ dissimilarity matrix, where $d(i, j) = d(j, i)$, measuring the difference or dissimilarity between two measures m_i and m_j. AHC finds the most similar clusters



Selection criteria:

- ● confidence 0.7
- ● support 0.3
- ● commonness 0.25
- ● rule size 2 => 1
- ● rules to show 50

Inclusive:
Graduate Course, Any Course* =>
Design And Analysis of Algorithms

Restrictive:
Basic Course, Any Course* =>
Any Course

View file: c:\datamine\ilmo\results\rules.all

(a) Selection

Number of selected rules: 47

Programming in C
Introduction to Unix
=> Programming in C

Programming in C
Introduction to Unix
=> Artificial Intelligence

Data Communications
Programming in C
=> Introduction to Unix

Artificial Intelligence
Programming in C
=> Introduction to Unix

Data Communications
Introduction to Unix
=> Programming in C

Artificial Intelligence
Database Systems II
=> Programming in C

VAX/VMS System
Data Communications
=> Artificial Intelligence

Data Communications
Introduction to Unix
=> Database Systems II

(b) Browsing

(c) Graph

Figure 4.14: RuleVisualizer prototype.

according to the average linkage method. PAM, is more robust than the k-means method, is to find a subset $m_1, m_2, ..., m_k \subset 1, ..., q$ which minimizes the objective function $\sum_{i=1}^{q} min_{t=1,...,k} d(i, m_t)$. CG, is a new approach implemented in the ARQAT tool (Chap. 5) while the results issued from AHC and PAM are illustrated with the R[5] tool.

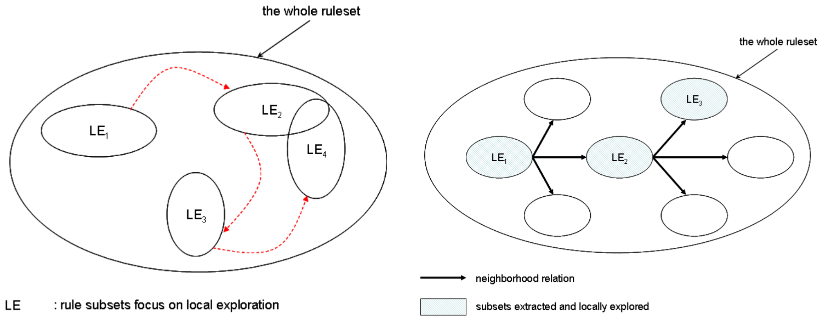
4.6.1 Dissimilarity between measures

Let $\mathcal{R}(\mathcal{D}) = \{r_1, r_2, ..., r_p\}$ denote input data as a set of p association rules derived from a dataset $\mathcal{D}$. Each rule $X \rightarrow Y$ is described by its itemsets (X, Y) and its cardinalities $(n, n_X, n_Y, n_{X\overline{Y}})$. Let $\mathcal{M}$ be the set of q available measures for our analysis $\mathcal{M} = \{m_1, m_2, ..., m_q\}$. Each measure is a numerical function on rule cardinalities: $m(X \rightarrow Y) = f(n, n_X, n_Y, n_{X\overline{Y}})$. For each measure $m_i \in \mathcal{M}$, we can construct a vector $m_i(\mathcal{R}) = \{m_{i1}, m_{i2}, ..., m_{ip}\}, i = 1..q$, where m_{ij} corresponds to the calculated value of the measure m_i for a given rule r_j.

We can then have a matrix $(p \times q)$ of interestingness values:

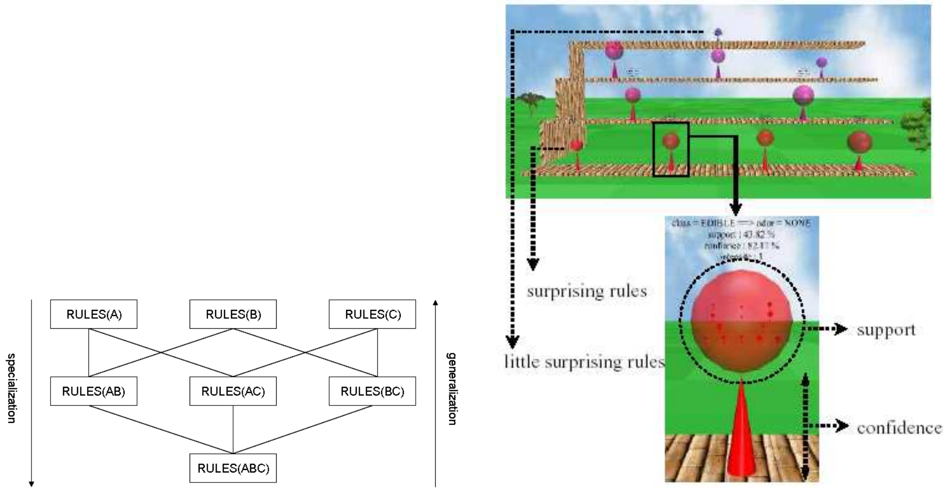
$$\mu = \begin{pmatrix} m_{11} & m_{12} & \ldots & m_{1q} \\ m_{21} & m_{22} & \ldots & m_{2q} \\ \ldots & \ldots & \ldots & \ldots \\ m_{p1} & m_{p2} & \ldots & m_{pq} \end{pmatrix}$$

[5] http://www.r-project.org/


the whole ruleset
LE1
LE2
LE3
LE4
LE : rule subsets focus on local exploration
the whole ruleset
LE1
LE2
LE3
neighborhood relation
subsets extracted and locally explored

(a) Local exploration (b) Neighborhood navigation

Figure 4.15: The neighborhood relation for rummaging.



(a) Specialization and generalization (b) Metaphor

Figure 4.16: Visualization with the ARVis platform.

Definition 4.6.1 (Similarity)**.** The similarity μ between two measures m_i and m_j is determined by the absolute correlation value calculated from a Pearson's, Spearman's or Kendall's coefficient ρ.

$$\mu(m_i, m_j) = |\rho(m_i, m_j)|$$

Definition 4.6.2 (Dissimilarity)**.** The dissimilarity ψ between two measures m_i and m_j is defined by:

$$\psi(m_i, m_j) = 1 - \mu(m_i, m_j) = 1 - |\rho(m_i, m_j)|$$

As correlation is symmetrical, the $\frac{q(q-1)}{2}$ dissimilarity/similarity values can be stored in one half of a matrix $q \times q$.

$$\psi = \begin{pmatrix} \psi_{11} & \psi_{12} & \dots & \psi_{1q} \\ \psi_{21} & \psi_{22} & \dots & \psi_{2q} \\ \dots & \dots & \dots & \dots \\ \psi_{q1} & \psi_{q2} & \dots & \psi_{qq} \end{pmatrix}$$

with: $\psi_{ii} = 0$, and $\forall (i,j), i \neq j, \psi_{ij} \geq 0, \psi_{ij} = \psi_{ji}$

Example. Given three measures m_1, m_2, m_3 and five association rules r_1,r_2,r_3,r_4,r_5. The interestingness of each rules with respect to each measure is illustrated the left partial table (Tab. 4.5). The dissimilarity values calculated between each couple of measures from the Pearson's coefficient are presented in the right partial table (Tab. 4.5). The dissimilarity values can be illustrated in a half of 3 × 3 table in which the dissimilarity of the measure with itself is zero.

$\mathcal{R}(\mathcal{D})$	m_1	m_2	m_3
r_1	0.84	0.89	0.91
r_2	0.86	0.90	0.93
r_3	0.88	0.94	0.97
r_4	0.94	0.95	0.99
r_5	0.83	0.87	0.84

ψ_{ij}	m_1	m_2	m_3
m_1		0.09	0.14
m_2			0.04
m_3			

Table 4.5: Dissimilarity values for three interestingness measures and five association rules.

4.6.2 AHC

Fig 4.17 illustrates a result obtained from a ruleset with a subset of 35 measures. The horizontal line goes throughout the cluster dendrogram has the small dissimilarity 0.15 determining the clusters of measures having strong relation. The user can intuitively choose a representative measure among the measures in a cluster. Intuitively, the user can choose the biggest cluster contains the measures Lift, Rule Interest, Phi-Coefficient, Kappa, Similarity Index, Putative Causal Dependency, Dependency, Klosgen, Pavillon for their first choice. In this cluster we can easily see two strong related clusters with four measures for each. This cluster gives the strongest effect on evaluation the similarity between two parts of an association rule. Another observation illustrates the existence of a confidence cluster with Causal Confidence, Causal Confirmed-Confidence, Laplace, Confidence, Descriptive Confirmed-Confidence. The user can then select this cluster to discover all the rules have the effect of high confidence.

The hierarchical structure also allows the user clearly seeing the clusters of measures that are connected closely with the hierarchical level computed.

4.6.3 PAM

The PAM clustering results are illustrated by three figures. The first one (Fig. 4.18) presents the results under a graphical form by projecting the measures according to the two principal axes obtained by a principal component analysis (PCA) [LMF82]. Each measure is presented by a symbol; each cluster, by a number. In spite of the deformation by projection, this visualization is very useful to have a synthetic view of the clustering. This graphical presentation is also useful to validate the

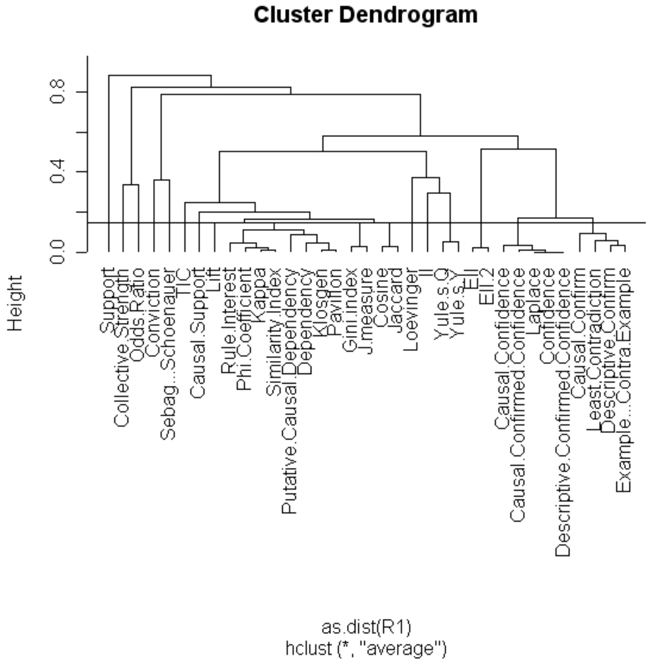


Figure 4.17: Clusters of measures with AHC on a ruleset.

choice of the number k of clusters, for example here $k = 16$ (the choice of these sixteen clusters is issued from the preliminary work presented in [HGB05c]).

Tab. 4.6 provides the obtained clusters and their medoids/representative measures. Tab. 4.7 details the numerical characteristics of the clusters. One can see that cluster 1 (Causal Confidence, Causal Confirmed-Confidence, Confidence, Descriptive Confirmed-Confidence, Laplace) and cluster 2 (Causal Confirm, Descriptive Confirm, Example & Contra-Example, Least Contradiction) are the largest and the least separated from the others since they have the lowest separation values (column 6, Tab. 4.7). Thus, the measures of these two clusters offer a very close point of views on the rules. This proximity is confirmed by observing the measures constitute these two clusters, as one can find the measures derived from the confidence measure. Furthermore, when observing the diameter (column 5, Tab. 4.7), we remark that cluster 1 is smaller than cluster 2. As a result, the measures of cluster 1 offer a more similar point of view on the rules than those of cluster 2. The clusters 5 (Conviction), 14 (Sebag-Schoenaueur), and 15 (Support) are constituted from only one measure and have the greatest values of separation. They illustrate a very different point of view on the studied rules.

4.6.4 Correlation graph

The third one is a graph-based view of the correlation matrix [HGB05b] [HGB06d] [HGB06c]. As graphs are a good means to offer relevant visual insights on data structure, the correlation matrix is used as the relation of an undirected and valued graph, called "correlation graph". In a correlation graph, a vertex represents a measure and an edge value is the correlation value between two vertices/measures. We also add the possibility to set a minimal threshold τ (maximal threshold

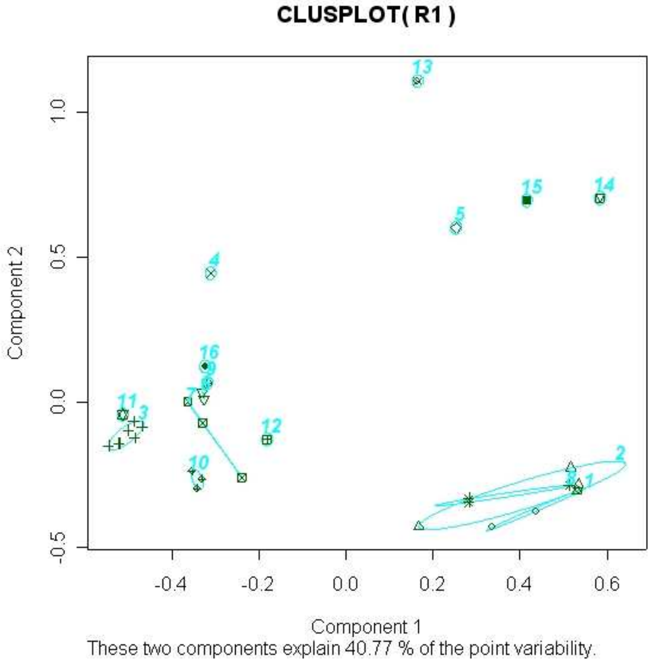


Figure 4.18: Measures and clusters projected on the two principal axes of a PCA. The measures are presented by the symbols and the clusters are represented by the numbers.

θ respectively) by absolute value to retain only the edges associated with a high correlation (respectively low correlation). When considering a large set of measures, the graph-based view of the correlation matrix may be quite complex. In order to highlight the more "natural" clusters, we propose to construct two types of subgraphs : the correlated ($CG+$) and the uncorrelated ($CG0$) partial subgraph. In this section we present the different filtering thresholds for their construction. We also extend the correlation graphs to graphs of stable clusters ($\overline{CG0}$ and $\overline{CG+}$) in order to compare several rulesets (see Chap. 7).

These two subgraphs can then be processed in order to extract clusters of measures: each cluster is defined as a connected subgraph. In CG+, each cluster gathers correlated or anti-correlated measures that may be interpreted similarly: they deliver a close point of view on data. Moreover, in CG0, each cluster contains uncorrelated measures, i.e. measures that deliver a different point of view. Hence, as each graph depends on a specific ruleset, the user can use the graphs as data insight, which graphically help him/her select the minimal set of the measures best adapted to his/her data. If a CG+ graph contains twelve clusters on 36 measures, the user can select the most representative measure in each cluster, and then retain it to validate the rules.

In order to make the interpretation of the large set of correlation values easier, we introduce the following definitions:

Definition 4.6.3 (τ-correlated/θ-uncorrelated)**.** Two measures m_i and m_j are τ-*correlated* with respect to the dataset $\mathcal{D}$ if their absolute correlation value is greater than or equal to a given threshold τ: $|\rho(m_i, m_j)| \geq \tau$. And, conversely, two measures m_i and m_j are θ-*uncorrelated* with respect to the dataset $\mathcal{D}$ if the absolute value of their correlation value is lower than or equal to a threshold value θ: $|\rho(m_i, m_j)| \leq \theta$.

N°	Name of measures	Representative measure
1	Causal Confidence, Causal Confirmed-Confidence, Confidence, Descriptive Confirmed-Confidence, Laplace	Causal Confirmed-Confidence
2	Causal Confirm, Descriptive Confirm, Example & Contra-Example, Least Contradiction	Example & Contra-Example
3	Causal Support,Kappa, Lerman, Phi-Coefficient, Rule Interest, Yule's Q, Yule's Y	Phi-Coefficient
4	Collective Strength	Collective Strength
5	Conviction	Conviction
6	Cosine, Jaccard	Jaccard
7	Dependency, Gini-index, J-measure	J-measure
8	EII, EII 2, IPEE	EII 2
9	II	II
10	Klosgen, Pavillon, Putative Causal Dependency	Klosgen
11	Lift	Lift
12	Loevinger	Loevinger
13	Odds Ratio	Odds Ratio
14	Sebag & Schoenauer	Sebag & Schoenauer
15	Support	Support
16	TIC	TIC

Table 4.6: Clusters of the measures obtained with PAM.

For θ-uncorrelated measures, we use a statistical test of significance by choosing a level of significance of the test $\alpha = 0.05$ for hypothesis testing (common values for α are: $\alpha = 0.1, 0.05, 0.005$). The threshold θ is then calculated by the following formula: $\theta = 1.960/\sqrt{p}$ in a population of size p [Ros87]. The assignment $\tau = 0.85$ of τ-correlated is used because this value is widely acceptable in the literature.

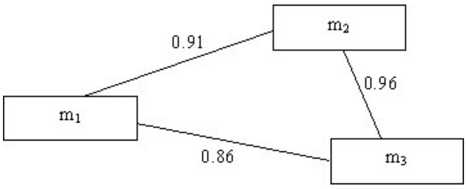


Figure 4.19: An illustration of the correlation graph.

As the correlation coefficient is symmetrical, the $q(q-1)/2$ correlation values can be stored in one half of the table $q \times q$. This table ($\mathcal{M} \times \mathcal{M}$) can also be viewed as the relation of an undirected and valued graph called correlation graph, in which a vertex value is a measure and an edge value is the correlation value between two vertices/measures. For instance, Fig. 4.19 can be the correlation graph obtained on five association rules $\mathcal{R}(\mathcal{D}) = \{r_1, r_2, r_3, r_4, r_5\}$ extracted from a dataset $\mathcal{D}$ and three measures $\mathcal{M} = \{m_1, m_2, m_3\}$ whose values and correlations are given in Tab. 4.8.

4.6.5 Correlated versus uncorrelated graphs

Unfortunately, when the correlation graph is complete, it is not directly human-readable. We need to define two transformations in order to extract more limited and readable subgraphs. By using definition 4.6.3, we can extract the *correlated partial subgraph (CG+)*: the subgraph composed of edges associated with a τ-correlated. On the same way, the *uncorrelated partial subgraph (CG0)* where we only retain edges associated with correlation values close to 0 (θ-uncorrelated).

These two partial subgraphs can then be used as a visualization support in order to observe the

N°	Size	Maximal distance	Average distance	Diameter	Separation
1	5	0.01705323	0.01260525	0.0641897	0.08648391
2	4	0.11602732	0.05102810	0.1267685	0.08648391
3	7	0.12459517	0.04592671	0.2599846	0.05986472
4	1	0.00000000	0.00000000	0.0000000	0.29375570
5	1	0.00000000	0.00000000	0.0000000	0.36002724
6	2	0.03107980	0.01553990	0.0310798	0.13697222
7	3	0.06636710	0.03336282	0.1212970	0.08094671
8	3	0.10800499	0.04583550	0.1251634	0.13059529
9	1	0.00000000	0.00000000	0.0000000	0.20598756
10	3	0.04345866	0.02029654	0.0548010	0.06509976
11	1	0.00000000	0.00000000	0.0000000	0.05986472
12	1	0.00000000	0.00000000	0.0000000	0.11382888
13	1	0.00000000	0.00000000	0.0000000	0.33963367
14	1	0.00000000	0.00000000	0.0000000	0.36002724
15	1	0.00000000	0.00000000	0.0000000	0.42706165
16	1	0.00000000	0.00000000	0.0000000	0.19300332

Table 4.7: The supplementary information obtained on the clusters of a ruleset.

$\mathcal{R} \times \mathcal{M}$	m_1	m_2	m_3
r_1	0.84	0.89	0.91
r_2	0.86	0.90	0.93
r_3	0.88	0.94	0.97
r_4	0.94	0.95	0.99
r_5	0.83	0.87	0.84

$\mathcal{M} \times \mathcal{M}$	m_1	m_2	m_3
m_1		0.91	0.86
m_2			0.96
m_3			

Table 4.8: Correlation values for three measures and five association rules.

correlative liaisons between measures. We can also observe the clusters of measures corresponding with the connected parts of the graphs.

4.7 Summary

Postprocessing of association rules is a difficult stage in the KDD process. A huge amount of rules in which the number of useful rules in point of view of the user is large. So the postprocessing task plays an important role to help the user focus on the most interesting ones. Fives such principal tasks are divided into groups: constraints, pruning, grouping, summarization and visualization.

Chapter 5

ARQAT tool

In this chapter, we present a new tool ARQAT to study the specific behavior of a set of objective measures in the context of several specific datasets and in an exploratory data analysis perspective. The tool implements 14 graphical and complementary views structured on 5 levels of analysis: ruleset analysis, correlation and clustering analysis, most interesting rules analysis, sensitivity analysis, and comparative analysis. The tool is described and illustrated on several datasets in order to show the interest of both the exploratory approach and the use of complementary views.

5.1 Principles

ARQAT (Fig. 5.1) is an exploratory analysis tool that embeds 40 objective measures (extended to 60 objective measures) studied in surveys (See Tab. 3.1 Chap. [3] for a complete list of selected measures). It provides graphical views structured in five task-oriented groups: ruleset analysis, correlation and clustering analysis, most interesting rules analysis, sensitivity analysis, and comparative analysis.

The ARQAT input is a set of association rules $\mathcal{R}$ and a set of objective measures $\mathcal{M}$. Each association rule $X \rightarrow Y$ must be associated with the four cardinalities n, n_X, n_Y, and $n_{X\overline{Y}}$. More precisely, n is the number of transactions, n_X (respectively n_Y) the number of transactions satisfying the itemset X (respectively Y), and $n_{X\overline{Y}}$ is the number of transactions satisfying $X \cap \overline{Y}$ (negative examples).

In **the first stage**, the input ruleset is preprocessed in order to compute the interestingness values of each rule, and the correlations between all measures pairs. The results are stored in two tables:

- an interestingness table ($\mathcal{R} \times \mathcal{M}$) where rows are rules and columns are interestingness measures,
- and a correlation matrix ($\mathcal{M} \times \mathcal{M}$) crossing measures.

At this stage, the ruleset may also be sampled (filtering box in Fig. 5.1) in order to focus the study on a more restricted subset of rules.

In **the second stage**, the data-analyst can drive the graphical exploration of results through a classical web-browser. ARQAT is structured in five groups of task-oriented views.

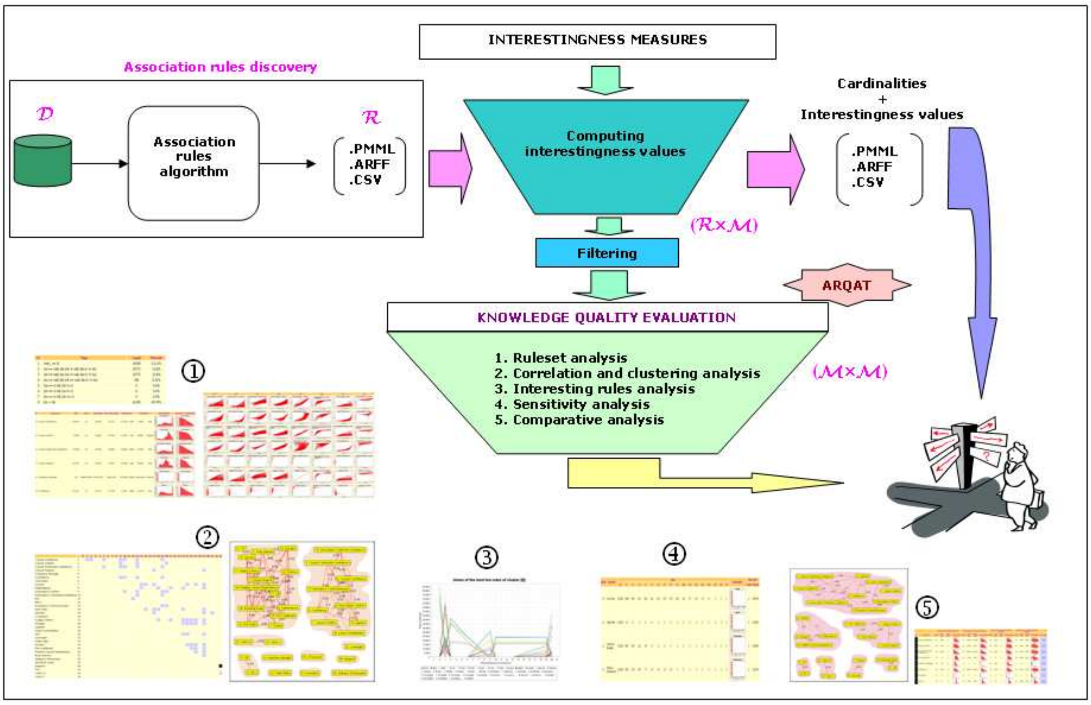

INTERESTINGNESS MEASURES
Association rules discovery
$\mathcal{D}$
Association rules algorithm
$\mathcal{R}$
.PMML
.ARFF
.CSV
Computing interestingness values
Cardinalities + Interestingness values
.PMML
.ARFF
.CSV
$(\mathcal{R}\times\mathcal{M})$
Filtering
ARQAT
KNOWLEDGE QUALITY EVALUATION
1. Ruleset analysis
2. Correlation and clustering analysis
3. Interesting rules analysis
4. Sensitivity analysis
5. Comparative analysis
$(\mathcal{M}\times\mathcal{M})$
①
②
③
④
⑤

Figure 5.1: ARQAT structure.

- The first group (1 in Fig. 5.1) is dedicated to ruleset and simple interestingness statistics to better understand the structure of the measure table ($\mathcal{R} \times \mathcal{M}$).
- The second group (2) is oriented to the study of measure correlation in table ($\mathcal{M} \times \mathcal{M}$) and measure clustering in order to select the most suitable measures.
- The third group (3) focuses on rule ordering to select the most interesting rules.
- The fourth group (4) proposes to study the sensitivity of measures.
- The last group (5) offers the possibility to compare the results obtained from different rulesets.

5.2 Modules

ARQAT is composed of three principal submodules: *preprocessing*, *evaluation*, and *display* (Fig. 5.2). Besides these three principal modules, there are the other two support modules: *utility* and *graph*. The UTILITY module supports the above modules to simplify the input/output facilities and some helpful manipulations. The GRAPH module helps the results issued from this tool to view the clusters obtained in a web browser. The ARQAT model is followed the framework (d) introduced in Fig. 1.2 - [Chap. 1].

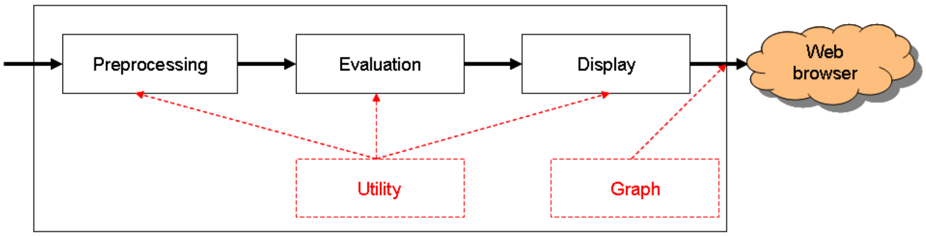


Figure 5.2: ARQAT modules.

5.3 Preprocessing

The preprocessing module (Fig. 5.3) firstly filters the input ruleset to have its list of cardinalities $f(n, n_X, n_Y, n_{X\overline{Y}})$. From these cardinalities, a list of interestingness values is computed with the help of objective measures. The interestingness values are then calculated and used to sampling the *original ruleset* to have a corresponding *sample ruleset*. The sample ruleset contains itself a set of most interesting rules where the number of rules is strongly less than the original ruleset. The sample rulesets are used to conduct a comparative study.

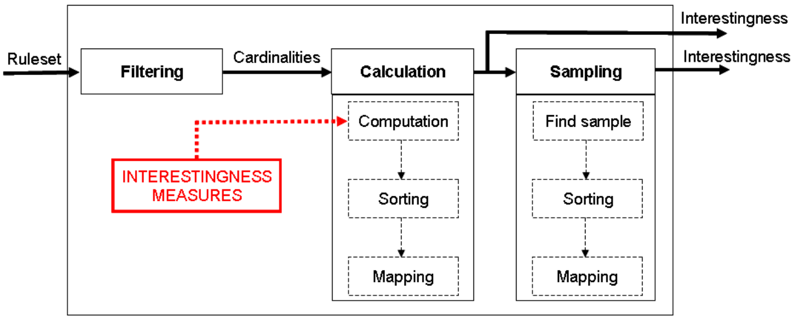


Figure 5.3: Preprocessing of rulesets.

5.3.1 Filtering

A ruleset, normally, has two files to identify its information achieved. The first one is an *item names* containing the *support count* [AIS93a] for each itemset with own name. The second one is usually a list of rules and contains some supplementary information for each rule (i.e.g. the support count for the rule). The other information, for example, are the interestingness values computed from some restricted objective measures (i.e. about five objective measures) such as the support count lift, support, confidence,.... We are not interested in these few interestingness values and recalculate all of them in the later stage for a more general analysis with 40 objective measures.

The input rulesets can be in the formats such as CSV, PMML, ARFF files. The cardinality set obtained is issued normally in the CSV format.

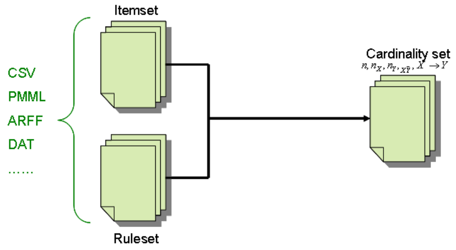


Figure 5.4: The filtering module.

5.3.2 Interestingness calculation

Each ruleset with its list of four cardinalities $(n, n_X, n_Y, n_{X\overline{Y}})$ is then calculated by an objective measure. The value obtained is called an interestingness value and stored in an *interestingness set*. The interestingness set is then sorted to have a *rank set*. The elements in the rank set is ranked due to its corresponding interestingness values. The higher the interestingness value the higher the rank obtained.

The other two necessary set are also created. The first set is an *order set*. Each element of the order set is an order mapping $f : 1 \longrightarrow 1$ for each element in the corresponding interestingness set. The *value set* contains the list of interestingness values correspond to the position of the elements int the rank set (i.e. mapping $f : 1 \longrightarrow 1$).

For example, with 40 objective measures, one can obtain 40 interestingness sets, 40 order sets, 40 rank sets and 40 value sets respectively (see Fig. 5.5). Each dataset type is saved in a corresponding folder. For instance, all the interestingness sets are stocked in an folder with the name INTERESTINGNESS. The other three folder names are ORDER, RANK and VALUE.

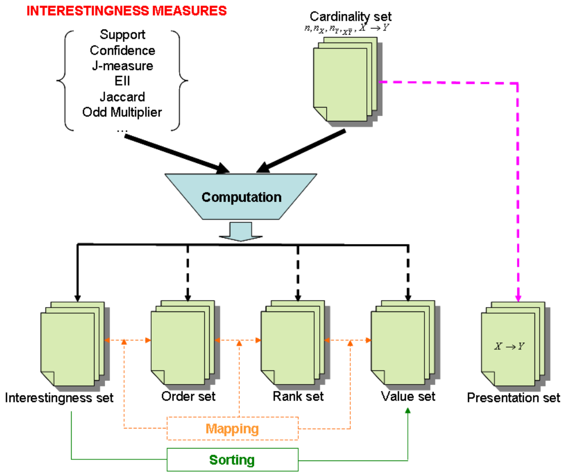


Figure 5.5: The calculation module.

5.3.3 Sampling

Definition 5.3.1 (Sample set)**.** The sample set is defined as a union of the k most interesting rules of each objective measure.

$$SampleSet = \bigcup_{i=1,j=1}^{i=q,j=p} r_{ij}\{rank(r_{ij}) \leq k\}$$

where q is the number of objective measures, p is the number of association rules. k is a parameter given by the user to quantify the interval of most interesting rules. The rule rank corresponds to the inversely statistical rank (i.e. the first most interesting rule has the rank of ”1”).

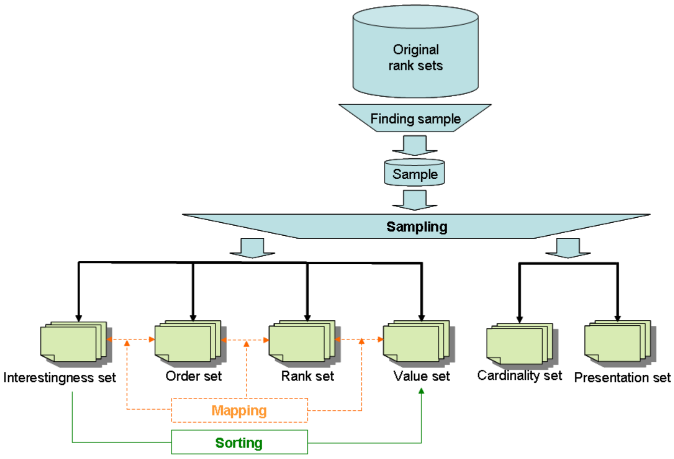


Figure 5.6: The sampling module.

5.4 Evaluation

5.4.1 Basic statistics

Ruleset characteristics

Each characteristic type is determined by a string representing its equation respectively. The purpose is to show the distributions underlying rule cardinalities, in order to detect ”borderline cases”. For instance, Tab. 5.1 gives 16 necessary characteristic types in our study in which the first line gives the number of ”logical” rules (i.e. rules without negative examples). The percentage of each characteristic type in the ruleset is also computed.

Initially, the counter of each characteristic type is set to zero. Each rule in the ruleset is then examined by its cardinalities to match the characteristic types. The complexity of the algorithm is linear $O(p)$.

N°	Type
1	$(n_{X\overline{Y}} = 0)$
2	$(n_X = n_{XY}) \wedge (n_Y \neq n_{XY}) \wedge (n \neq n_Y)$
3	$(n_Y = n_{XY}) \wedge (n_X \neq n_{XY}) \wedge (n \neq n_X)$
4	$(n_X = n_{XY}) \wedge (n_Y = n_{XY}) \wedge (n \neq n_X)$
5	$(n_X = n) \wedge (n_Y \neq n)$
6	$(n_Y = n) \wedge (n_X \neq n)$
7	$(n_X = n) \wedge (n_Y = n)$
8	$(n_X < n_Y)$
9	$(n_X < n_Y/2)$
10	$(n_X < n_Y/4)$
11	$(n_X < n_Y/6)$
12	$(n_X < n_Y/8)$
13	$(n_X < n_Y/10)$
14	$(n_X == n_Y)$
15	$(n_{X\overline{Y}} == n_X/2)$
16	$(n_{X\overline{Y}} == (n_X * n_Y)/2$

Table 5.1: Characteristic types (remind that $n_{XY} = n_X - n_{X\overline{Y}}$).

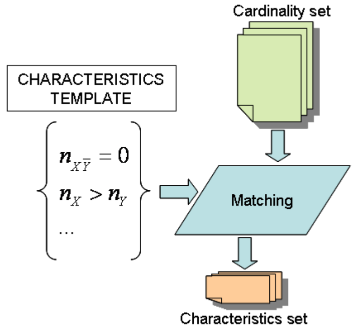


Figure 5.7: Evaluating some characteristics of a ruleset.

Histogram

To draw the histogram for each measures from its interestingness values, we used the JFreeChart package[1]. We added to this package a new histogram, namely inversely cumulative distribution, to explore the sensitivity concept (Sec. 5.4.1) proposed in our study. Tab. 5.2 gives an example with the four histogram types mentioned above.

Definition 5.4.1 (Cumulative distribution)**.** A cumulative distribution is a plot whose heights shows the proportion of data values that are smaller than or equal to any given number [Ros87]. This concept it is quite related with the well-known histogram representation.

For our purpose, we take inversely the cumulative distribution representation in order to show

[1] http://www.jfree.org/jfreechart/index.php

Bins / Histogram	1	2	3	4	5	6	7	8	9	10
Frequency	7	1	12	9	20	30	70	9	2	65
Relative frequency	0.031	0.004	0.053	0.040	0.88	0.133	0.311	0.040	0.008	0.288
Cumulative	7	8	20	29	49	79	149	158	160	225
Inverse cumulative	225	218	217	205	196	176	146	76	67	65

Table 5.2: Frequency and inverse-cumulative in bins.

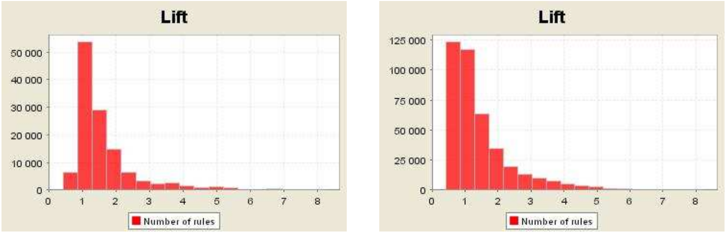


Figure 5.8: Frequency and inverse-cumulative histograms of the Lift measure.

the number of rules that have been ranked higher than an eventually specified minimum threshold. Intuitively, the user can see exactly the number of rules that he will have to deal with in the case in which he/she will choose a particular value for the minimum threshold.

The number of bins are directly dependent of the dataset size p. It is generated by the following Sturges formula:

$$BinWidth = \frac{Max - Min}{SturgesFormula} \text{ with } Sturges\ Formula = 1 + 3.3 * log(p)$$

Distribution

Statistical significance	Symbol	Formula
Min	min	$min(v_i)$
Max	max	$max(v_i)$
Mean	$mean$	$\frac{\sum_{i=1}^{p} v_i}{p}$
Variance	var	$\frac{\sum_{i=1}^{p}(v_i - mean)^2}{p-1}$
Standard deviation	std	$\sqrt{var}$
Skewness	$skewness$	$\frac{\sum_{i=1}^{p}(v_i - mean)^3}{(p-1)\times std}$
Kurtosis	$kurtosis$	$\frac{\sum_{i=1}^{p}(v_i - mean)^4}{(p-1)\times var^2} - 3$

Table 5.3: Some statistical indicators on a measure m.

The distribution of each measure can be useful to the users. From this information the user can have a quick evaluation on the ruleset. Some significant statistical characteristics about *minimum value*, *maximum value*, *average value*, *standard deviation value*, *skewness value*, *kurtosis value* are computed (Tab. 5.3). The shape information of the last two arguments are also determined. In addition, the histograms like frequency and inverse cumulative are also drawn (see Sec. 5.4.1).

Assume that $\mathcal{R}$ is the set of p association rules. Each association rule r_i $(i = 1..p)$ has an interestingness value v_i computed from a measure m.

Sensitivity

The sensitivity of an interestingness measure is referred at the number of most interesting rules that an interested user should have to analyze, and if these rules are still well distributed (have different assigned ranks), or all have ranks equal to the maximum assigned value for the specified data set. Tab. 5.4 shows a structure to be evaluated by the user.

rank	measure	inverse-cumulative bins				histogram	best rules
		1	2	3	...		

Table 5.4: Sensitivity structure.

Average

Due to the fact that the number of bins is not the same when we have many rulesets to evaluate the sensitivity, so the number of rules that returned in the last interval also has not the same significance. Assume that the total number of measures to rank is fixed, the average ranks is used. The latter one is calculated according to the rank of each measure obtained from each ruleset. A weight can be assigned to each ruleset to favorite the level of importance, given by the user.

We use the average ranks to rank the measure over a set of rulesets based on the sensitivity values computed. The complement rulesets are benefited from this evaluation.

rank	measure	ruleset 1					ruleset 2	...	avg. rank
		rank	first bin	last bin	image	best rule	...	...	

Table 5.5: Average structure to evaluate sensitivity on a set of rulesets.

An average structure (see Tab. 5.5) is constructed to have a quick evaluation on a set of rulesets. Each row represents a measure. The first two columns are represent the current rank of the measure. For each ruleset, the rank, first bin, last bin, image and best rule assigned for each measure are represented. A remark is that the first and last bins are taken from the inversely cumulative distribution. The last column is the average rank of each measure calculated from all the rulesets studied.

Scatterplot

The purpose of this technique is to check the pairwise relationships between variables.

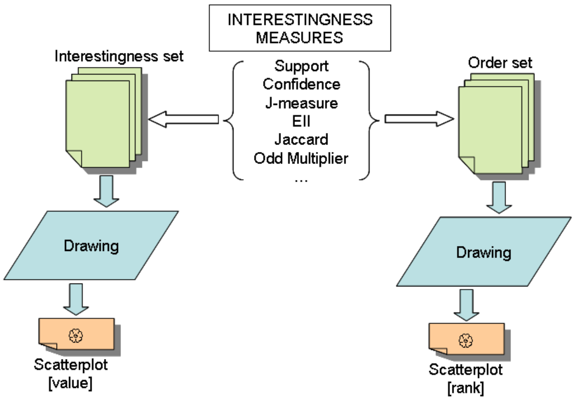
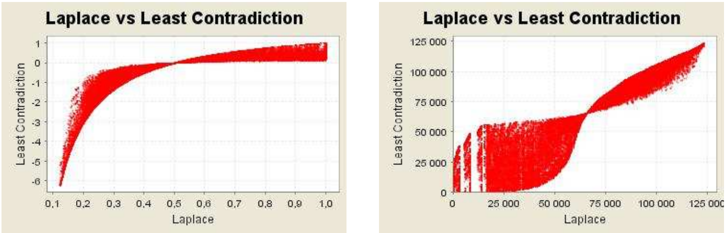


Figure 5.9: Drawing scatterplot view.

Definition 5.4.2 (Scatterplot matrix)**.** Given a set of variables, the scatterplot matrix contains all the pairwise scatter plots of the variables on a single page in a matrix format. That is, if there are variables, the scatterplot matrix will have rows and columns and the row and column of this matrix is a plot of versus.

In ARQAT, both the scatterplots on value and on rank are drawn to give a more insight into the intuitive view between each of the pairwise of any couple of measures.



(a) On value (b) On rank

Figure 5.10: An example of two scatterplots on the ruleset Mushroom.

5.4.2 Correlation analysis

Correlation

To compare the measures, we calculate all the correlations between the studied measures using the Person's, Spearman's or Kendall's coefficient (Fig. 5.11). The correlation values are stocked in a square matrix as given in Tab. 5.6.

- Pearson's coefficient

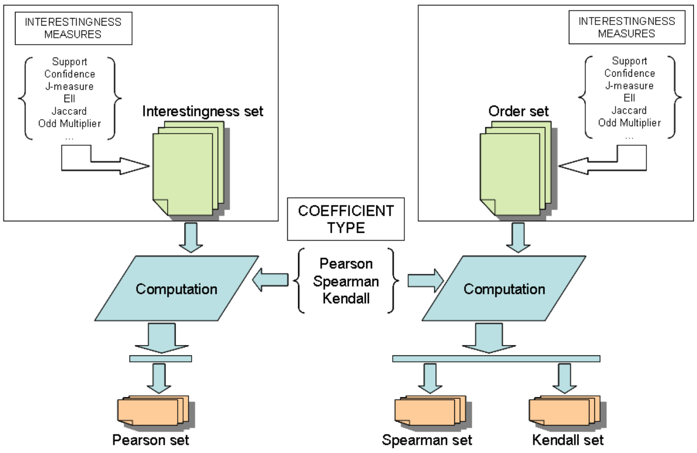


Figure 5.11: Computing correlation values.

	m_1	m_2	$\dots$	m_q
m_1	0	$v_{1,2}$	$\dots$	$v_{1,q}$
m_2		0	$\dots$	$v_{2,q}$
$\dots$			0	$\dots$
m_q				0

Table 5.6: Matrix of correlation values.

The correlation value between any two measures $m_i, m_j\{i,j = 1..q\}$ on the set of rules $\mathcal{R}$ is calculated by using a Pearson's correlation coefficient $\rho(m_i, m_j)$ [Ros87], where $\overline{m_i}, \overline{m_j}$ are the average values calculated of vector $m_i(\mathcal{R})$ and $m_j(\mathcal{R})$ respectively:

$$\rho_P(m_i, m_j) = \frac{\sum_{k=1}^{p}[(m_{ik} - \overline{m_i})(m_{jk} - \overline{m_j})]}{\sqrt{[\sum_{k=1}^{p}(m_{ik} - \overline{m_i})^2][\sum_{k=1}^{p}(m_{jk} - \overline{m_j})^2]}} \tag{5.4.1}$$

$\mathcal{R} \times \mathcal{M}$	m_1	m_2	m_3
r_1	0.95	0.97	0.10
r_2	0.80	0.91	0.30
r_3	0.87	0.91	0.94
r_4	0.90	0.80	0.94
r_5	-0.60	0.70	0.95

$\mathcal{M} \times \mathcal{M}$	m_1	m_2	m_3
m_1		0.83	-0.42
m_2			-0.73
m_3			

Table 5.7: Pearson's values for three measures and five association rules.

- Spearman's coefficient

  The correlation value between any two objective measures m_i, m_j, $\{i,j = 1..q\}$ on a ruleset $\mathcal{R}$ will be calculated by using the Spearman's rank correlation coefficient ρ_S [LD98]. The Spearman coefficient uses the ordering of the two interestingness values.

$$\rho_S(m_i, m_j) = \frac{\sum_{k=1}^{p}[(m_{ik} - \overline{m_i})(m_{jk} - \overline{m_j})]}{\sqrt{[\sum_{k=1}^{p}(m_{ik} - \overline{m_i})^2][\sum_{k=1}^{p}(m_{jk} - \overline{m_j})^2]}} \tag{5.4.2}$$

where $\overline{m_i}, \overline{m_j}$ are the average ranks calculated from vectors $m_i(\mathcal{R})$ and $m_j(\mathcal{R})$ respectively. m_{ik} and m_{jk} are the ranks of the rule r_k with respect to the two interestingness values calculated from the measures m_i, m_j respectively.

	Interestingness values		
$\mathcal{R}(\mathcal{D})$	m_1	m_2	m_3
r_1	0.95	0.97	0.10
r_2	0.80	0.91	0.30
r_3	0.87	0.91	0.94
r_4	0.90	0.80	0.94
r_5	-0.60	0.70	0.95

⇒

	Ranks		
$\mathcal{R}(\mathcal{D})$	m_1	m_2	m_3
r_1	5	5	1
r_2	2	4	2
r_3	3	4	4
r_4	4	2	4
r_5	1	1	5

⇒

	Correlation values		
$\mathcal{M} \times \mathcal{M}$	m_1	m_2	m_3
m_1		0.67	-0.67
m_2			-0.92
m_3			

Table 5.8: Spearman's values for three measures and five rules.

If there are no ties[2] in the rankings[3]:

$$\rho_S(m_i, m_j) = 1 - \frac{6}{p(p^2-1)}\sum_{k=1}^{p} d_k^2 \tag{5.4.3}$$

If there are ties:

$$\rho_S(m_i, m_j) = \frac{\frac{p(p^2-1)}{6} - \sum_{k=1}^{p} d_k^2 - \frac{1}{12}T_u - \frac{1}{12}T_v}{\sqrt{(\frac{p(p^2-1)}{6} - 2T_u)(\frac{p(p^2-1)}{6} - 2T_v)}} \tag{5.4.4}$$

where d_k $(k = 1..p)$ is the difference in statistical rank between the two objective measures m_i and m_j for the same association rule r_k. $T_u = \frac{1}{12}\sum t_u(t_u^2-1)$, $T_v = \frac{1}{12}\sum t_v(t_v^2-1)$, the rank of rule r_u is a tie and t_u is the number of tie beginning from the position u $\{u = 1..p\}$ of the ranking issued from m_i (v, t_v, m_j respectively). The range of $\rho_S(m_i, m_j)$ is $-1 \leq \rho_S(m_i, m_j) \leq 1$.

- Kendall's coefficient

  The correlation value between any two objective measures m_i, m_j, $\{i, j = 1..q\}$ on a ruleset $\mathcal{R}$ will be calculated by using the Kendall's rank correlation coefficient ρ_K [LD98]. The Kendall coefficient uses the ordering of the two interestingness values.

  Considering all pairs of objective measure values (m_{ik}, m_{jk}) and (m_{it}, m_{jt}) $\{k, t = 1..p\}$ computed from two rules r_k and r_t . A pair is called:

  - concordant if ($m_{ik} < m_{jk}$ and ($m_{it} < m_{jt}$)) or ($m_{ik} > m_{jk}$ and ($m_{it} > m_{jt}$)), and
  - discordant if ($m_{ik} < m_{jk}$ and ($m_{it} > m_{jt}$)) or ($m_{ik} > m_{jk}$ and ($m_{it} < m_{jt}$)).

[2] Tie: equal value.

[3] Five possible methods can be used: *average*, *first*, *min*, *max*, and *random* (here we use the *max* method). It is usually arranged in a decreasing order (or may be in an increasing order). In the experiment results, we use the increasing order to give a clear and "natural" view to the user.

where m_{ik}, m_{jk}, m_{it}, m_{jt} $(i, j = 1..m,\ k, t = 1..p)$ are the interestingness values of the objective measures m_i and m_j for the association rules r_k, r_t respectively.

The Kendall rank correlation coefficient value ρ_K is then calculated as:

$$\rho_K(m_i, m_j) = \frac{n_{concordant} - n_{discordant}}{\frac{p(p-1)}{2}} \tag{5.4.5}$$

where $n_{concordant}$, $n_{discordant}$, are the number of concordant and discordant pairs respectively. p is the number of association rules. The range of $\rho_K(m_i, m_j)$ is $-1 \leq \rho_K(m_i, m_j) \leq 1$.

If $m_{ik} = m_{jk}$ or $m_{it} = m_{jt}$ or both (i.e., a tie). Ties are not counted as concordant or discordant. The dominator $\begin{pmatrix} p \\ 2 \end{pmatrix} = \frac{p(p-1)}{2}$ will be replaced by $\sqrt{\left[\begin{pmatrix} p \\ 2 \end{pmatrix} - t_u\right] \times \left[\begin{pmatrix} p \\ 2 \end{pmatrix} - t_v\right]}$ when there is a large number of ties, where t_u (t_v respectively) is the number of ties involving u (v respectively).

	Interestingness values		
$\mathcal{R}(\mathcal{D})$	m_1	m_2	m_3
r_1	0.95	0.97	0.10
r_2	0.80	0.91	0.30
r_3	0.87	0.91	0.94
r_4	0.90	0.80	0.94
r_5	-0.60	0.70	0.95

⇒

	Ranks		
$\mathcal{R}(\mathcal{D})$	m_1	m_2	m_3
r_1	5	5	1
r_2	2	4	2
r_3	3	4	4
r_4	4	2	4
r_5	1	1	5

⇒

	Correlation values		
$\mathcal{M} \times \mathcal{M}$	m_1	m_2	m_3
m_1		0.53	-0.53
m_2			-0.89
m_3			

Table 5.9: Kendall's values for three measures and five rules.

	(0.95, 0.97)	(0.80, 0.91)	(0.87, 0.91)	(0.90, 0.80)	(-0.60, 0.70)
(0.95, 0.97)	–	●	●	●	●
(0.80, 0.91)		–	–	○	●
(0.87, 0.91)			–	○	●
(0.90, 0.80)				–	●
(-0.60, 0.70)					–

Table 5.10: An example of concordant and discordant between the pairs of interestingness values computed from two objective measures m_1 and m_2 ($< \bullet >$: concordant, $< \circ >$: discordant).

So that: $\rho_K(m_1, m_2) = \frac{7-2}{\sqrt{(10-0)(10-1)}} = \frac{5}{\sqrt{90}} = 0.53$

Line chart

From a pair of a rule and its corresponding interestingness value or rank, we draw a line chart to focus on the most interesting zone concentrated by a set of rules.

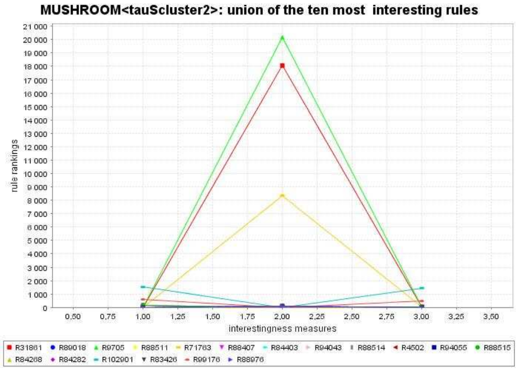


Figure 5.12: An example of a line chart on Mushroom ruleset.

Summary

Summary is a structure created to hold the correlation value between each pair of measures and more interestingly, its cluster type (e.g. τ-cluster or θ-cluster). The number of lines (couple of measures) in a summary is $\frac{n(n-1)}{2}$. The first two columns are the names of the measure pair. The next three columns are the correlation value calculated from the three well-known correlation coefficients Pearson, Spearman and Kendall (shortly P, S and K). The next three columns (PS, PK, and SK) are the stable clusters computed from the precedent three correlation values on each couple of coefficients. the last column is the stable cluster computed from the over all the three coefficients P, S and K.

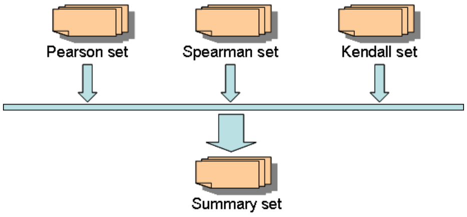


Figure 5.13: Summary scheme.

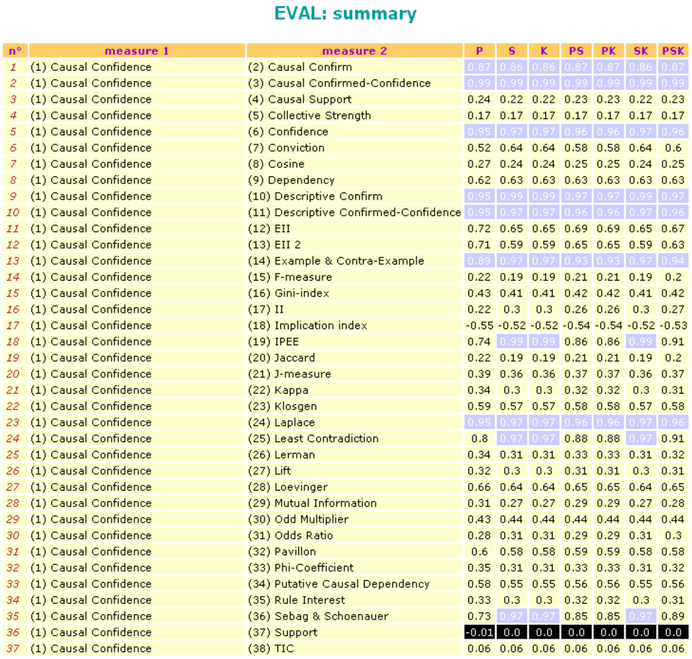
Measure 1	Measure 2	P	S	K	PS	PK	SK	PSK

Figure 5.14: A summary structure (P: Pearson, S: Spearman, K: Kendall).

Consider two measures namely *Causal Confidence* and *Causal Confirm* with their correlation values calculated from the P, S and K coefficients are 0.89, 0.84, 0.88 respectively. Easily, we can determine these three correlation values will determine three corresponding τ-clusters with a τ theshold = 0.85. Then we have PS, SK, PSK are neither a τ-cluster nor a θ-cluster but PK is a τ-cluster. The average value calculated are $PS = \frac{0.89+0.84}{2} = 0.87$, $PK = \frac{0.89+0.88}{2} = 0.89$, $SK = \frac{0.84+0.88}{2} = 0.86$, $PSK = \frac{0.89+0.84+0.88}{3} = 0.87$. Although the four average values are all greater than the τ-threshold but there is only PK is a τ-cluster.

Measure1	Measure 2	P		S		K		PS		PK		SK		PSK	
		value	cluster	. . .	. . .										

Figure 5.15: Stable combinations with correlation values (P: Pearson, S: Spearman, K: Kendall).



EVAL: summary

n°	measure 1	measure 2	P	S	K	PS	PK	SK	PSK
1	(1) Causal Confidence	(2) Causal Confirm	0.87	0.86	0.86	0.87	0.87	0.86	0.87
2	(1) Causal Confidence	(3) Causal Confirmed-Confidence	0.99	0.99	0.99	0.99	0.99	0.99	0.99
3	(1) Causal Confidence	(4) Causal Support	0.24	0.22	0.22	0.23	0.23	0.22	0.23
4	(1) Causal Confidence	(5) Collective Strength	0.17	0.17	0.17	0.17	0.17	0.17	0.17
5	(1) Causal Confidence	(6) Confidence	0.95	0.97	0.97	0.96	0.96	0.97	0.96
6	(1) Causal Confidence	(7) Conviction	0.52	0.64	0.64	0.58	0.58	0.64	0.6
7	(1) Causal Confidence	(8) Cosine	0.27	0.24	0.24	0.25	0.25	0.24	0.25
8	(1) Causal Confidence	(9) Dependency	0.62	0.63	0.63	0.63	0.63	0.63	0.63
9	(1) Causal Confidence	(10) Descriptive Confirm	0.95	0.99	0.99	0.97	0.97	0.99	0.97
10	(1) Causal Confidence	(11) Descriptive Confirmed-Confidence	0.95	0.97	0.97	0.96	0.96	0.97	0.96
11	(1) Causal Confidence	(12) EII	0.72	0.65	0.65	0.69	0.69	0.65	0.67
12	(1) Causal Confidence	(13) EII 2	0.71	0.59	0.59	0.65	0.65	0.59	0.63
13	(1) Causal Confidence	(14) Example & Contra-Example	0.89	0.97	0.97	0.93	0.93	0.97	0.94
14	(1) Causal Confidence	(15) F-measure	0.22	0.19	0.19	0.21	0.21	0.19	0.2
15	(1) Causal Confidence	(16) Gini-index	0.43	0.41	0.41	0.42	0.42	0.41	0.42
16	(1) Causal Confidence	(17) II	0.22	0.3	0.3	0.26	0.26	0.3	0.27
17	(1) Causal Confidence	(18) Implication index	-0.55	-0.52	-0.52	-0.54	-0.54	-0.52	-0.53
18	(1) Causal Confidence	(19) IPEE	0.74	0.99	0.99	0.86	0.86	0.99	0.91
19	(1) Causal Confidence	(20) Jaccard	0.22	0.19	0.19	0.21	0.21	0.19	0.2
20	(1) Causal Confidence	(21) J-measure	0.39	0.36	0.36	0.37	0.37	0.36	0.37
21	(1) Causal Confidence	(22) Kappa	0.34	0.3	0.3	0.32	0.32	0.3	0.31
22	(1) Causal Confidence	(23) Klosgen	0.59	0.57	0.57	0.58	0.58	0.57	0.58
23	(1) Causal Confidence	(24) Laplace	0.95	0.97	0.97	0.96	0.96	0.97	0.96
24	(1) Causal Confidence	(25) Least Contradiction	0.8	0.97	0.97	0.88	0.88	0.97	0.91
25	(1) Causal Confidence	(26) Lerman	0.34	0.31	0.31	0.33	0.33	0.31	0.32
26	(1) Causal Confidence	(27) Lift	0.32	0.3	0.3	0.31	0.31	0.3	0.31
27	(1) Causal Confidence	(28) Loevinger	0.66	0.64	0.64	0.65	0.65	0.64	0.65
28	(1) Causal Confidence	(29) Mutual Information	0.31	0.27	0.27	0.29	0.29	0.27	0.28
29	(1) Causal Confidence	(30) Odd Multiplier	0.43	0.44	0.44	0.44	0.44	0.44	0.44
30	(1) Causal Confidence	(31) Odds Ratio	0.28	0.31	0.31	0.29	0.29	0.31	0.3
31	(1) Causal Confidence	(32) Pavillon	0.6	0.58	0.58	0.59	0.59	0.58	0.58
32	(1) Causal Confidence	(33) Phi-Coefficient	0.35	0.31	0.31	0.33	0.33	0.31	0.32
33	(1) Causal Confidence	(34) Putative Causal Dependency	0.58	0.55	0.55	0.56	0.56	0.55	0.56
34	(1) Causal Confidence	(35) Rule Interest	0.33	0.3	0.3	0.32	0.32	0.3	0.31
35	(1) Causal Confidence	(36) Sebag & Schoenauer	0.73	0.97	0.97	0.85	0.85	0.97	0.89
36	(1) Causal Confidence	(37) Support	-0.01	0.0	0.0	0.0	0.0	0.0	0.0
37	(1) Causal Confidence	(38) TIC	0.06	0.06	0.06	0.06	0.06	0.06	0.06

Figure 5.16: Summary scheme extracted from a ruleset (A lighter color box shows a τ-cluster, the otherwise a black box is a θ-cluster).

Clustering

From the summary table discovered from each single ruleset (see also the synthesis subsection in Sec. 5.4.3), we find the connected components to form clusters. There are two types of clusters: τ-cluster (correlated) and θ-cluster (uncorrelated). The information from each column of a summary (P, S, K, PS, PK, SK, PSK) is considered to group the measures into the corresponding cluster types. So that, with a summary we have $2 \times 7 = 14$ cluster tables.

Matrix

The user has supplied three ways to intuitively evaluate the relation between each pair of measures. The first one is a significant matrix, represented by different colors assigned for each level of significance, to give an overview of the entire relation. The second one is a value matrix in which is pair of measures is computed by a coefficient or an average value. The third one gives a matrix with each cell is a scatterplot image. With the latter one, the user can have a quick comparison between the correlation value calculated with the cloud of pixels formed by the two measures. The latter is also implemented to resolve the gap between a correlation value and its scatterplot form.

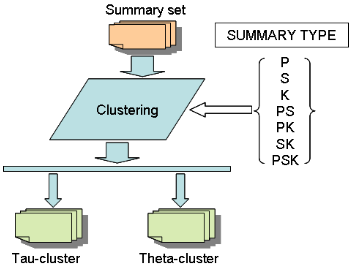


Figure 5.17: Clustering measures.

For the combination of coefficient (PS, PK, SK or PSK), a combination of value and cluster type is given in Fig. 5.18. These combinations are made on the value but not on the image (scatterplot).

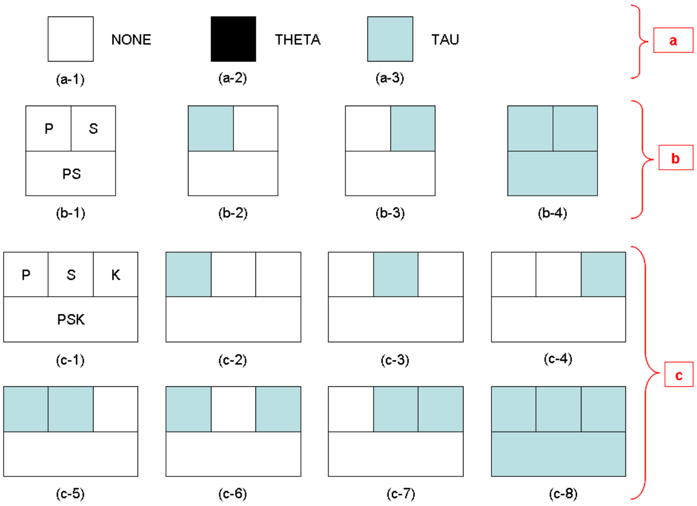


Figure 5.18: Combination of cluster type for a cell in a matrix ([**b**] and [**c**] are only visualized for τ-cluster but it is the same for θ-cluster). [**a**] is visualized for the matrix with only one coefficient such as P, S or K. [**b**] is visualized for the matrix with two coefficients PS, PK or SK. [**c**] is visualized for the matrix with three coefficients PSK.

Graph

As graphs are a good means to offer relevant visual insights on data structure, the correlation matrix is used as the relation of an undirected and valued graph, called "correlation graph". In a correlation graph, a vertex represents a measure and an edge value is the correlation value between two vertices/measures. We also add the possibility to set a minimal threshold τ (maximal threshold θ respectively) by absolute value to retain only the edges associated with a high correlation (respectively low correlation); the associated subgraphs are denoted by CG+ and CG0.

The value assigned for an edge is given from the correlation value calculated from Pearson (P), Spearman (S) or Kendall (K) coefficients. More general, an average correlation value is obtained when we interest in a combination between these coefficients: PS, PK, SK, or PSK.

The graph module is constructed to facilitate the view of the clusters discovered in the analyzed process. The user can drag and drop a node (i.e. an interestingness measure) to see the cluster.

5.4.3 Comparative study

Representative

In a cluster, all the measures can be represented by a single measure called representative measure. How we can choose a representative measure? The first way is that we let the user to select the measure that he/she feels the best. In the second way, we sort all the measures in the cluster by their number of relations with the other measures in the same cluster. From this point of view, the user can identify the "central of gravity" of the cluster.

Sample

Given a set of measures and a sample type (i.e. union or intersection), a set of rules is computed satisfying the two precedent conditions. Notice that the first ten most interesting rules of each measure with its corresponding ruleset are also stored before.

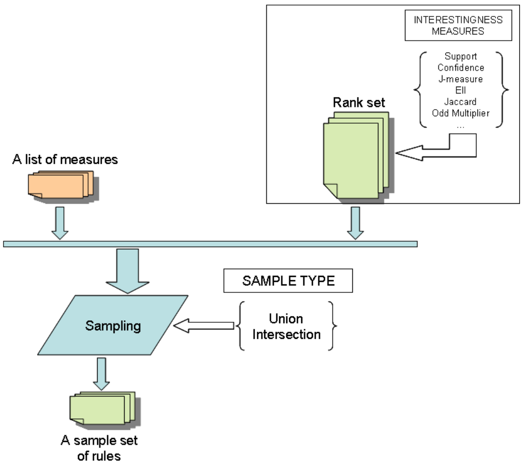


Figure 5.19: A process to compute a sample.

Synthesis

We proposed a stable hierarchy to discover the stable clusters between measures over all possibly combination sets of rulesets (see Fig. 5.20). This is the most interesting and important in ARQAT tool. In the above section (Sec. 5.13), a quick observation is computed via the four stable clusters in a ruleset (PS, PK, SK and PSK). In this component, we continue to examine a widely set of rulesets to discover the stable behaviors of measures. This widely set includes a couple set (between an original ruleset and its sample ruleset) and a multiple set (over all the original and sample rulesets, over all the original rulesets and over all the sample rulesets). If the number of original rulesets is 4, then the number of sample rulesets is 4, the number of couple rulesets is 4 and the number of

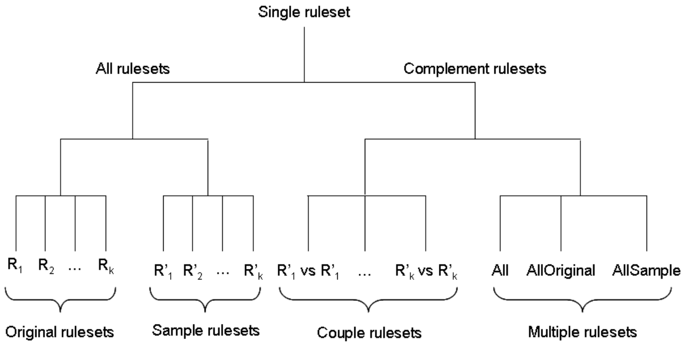
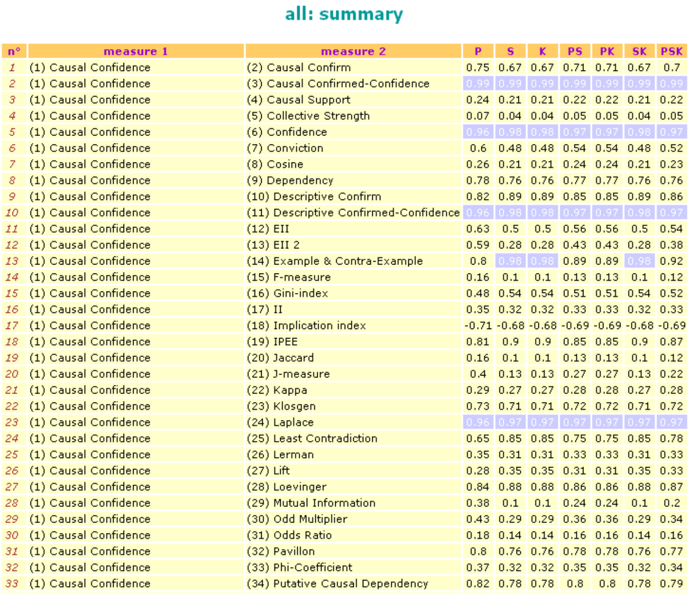


Figure 5.20: A stable hierarchy on a set of rulesets.

multiple rulesets is 3. Given k original rulesets, the total rulesets to evaluate is $k + k + k + 3 = 3k + 3 = 3(k + 1)$.

The set of all the original and sample rulesets is called "ALL" rulesets. The other rulesets (couple and multiple) are called "COMPLEMENT" rulesets. The entire rulesets are called "SINGLE" rulesets.

A summary structure is hold to determine the stable clusters and their average values calculated. A cluster is a τ-cluster (θ-cluster) if and only if all the clusters at the same level are all τ-cluster (θ-cluster).



all: summary

n°	measure 1	measure 2	P	S	K	PS	PK	SK	PSK
1	(1) Causal Confidence	(2) Causal Confirm	0.75	0.67	0.67	0.71	0.71	0.67	0.7
2	(1) Causal Confidence	(3) Causal Confirmed-Confidence	0.99	0.99	0.99	0.99	0.99	0.99	0.99
3	(1) Causal Confidence	(4) Causal Support	0.24	0.21	0.21	0.22	0.22	0.21	0.22
4	(1) Causal Confidence	(5) Collective Strength	0.07	0.04	0.04	0.05	0.05	0.04	0.05
5	(1) Causal Confidence	(6) Confidence	0.96	0.98	0.98	0.97	0.97	0.98	0.97
6	(1) Causal Confidence	(7) Conviction	0.6	0.48	0.48	0.54	0.54	0.48	0.52
7	(1) Causal Confidence	(8) Cosine	0.26	0.21	0.21	0.24	0.24	0.21	0.23
8	(1) Causal Confidence	(9) Dependency	0.78	0.76	0.76	0.77	0.77	0.76	0.76
9	(1) Causal Confidence	(10) Descriptive Confirm	0.82	0.89	0.89	0.85	0.85	0.89	0.86
10	(1) Causal Confidence	(11) Descriptive Confirmed-Confidence	0.96	0.98	0.98	0.97	0.97	0.98	0.97
11	(1) Causal Confidence	(12) EII	0.63	0.5	0.5	0.56	0.56	0.5	0.54
12	(1) Causal Confidence	(13) EII 2	0.59	0.28	0.28	0.43	0.43	0.28	0.38
13	(1) Causal Confidence	(14) Example & Contra-Example	0.8	0.98	0.98	0.89	0.89	0.98	0.92
14	(1) Causal Confidence	(15) F-measure	0.16	0.1	0.1	0.13	0.13	0.1	0.12
15	(1) Causal Confidence	(16) Gini-index	0.48	0.54	0.54	0.51	0.51	0.54	0.52
16	(1) Causal Confidence	(17) II	0.35	0.32	0.32	0.33	0.33	0.32	0.33
17	(1) Causal Confidence	(18) Implication index	-0.71	-0.68	-0.68	-0.69	-0.69	-0.68	-0.69
18	(1) Causal Confidence	(19) IPEE	0.81	0.9	0.9	0.85	0.85	0.9	0.87
19	(1) Causal Confidence	(20) Jaccard	0.16	0.1	0.1	0.13	0.13	0.1	0.12
20	(1) Causal Confidence	(21) J-measure	0.4	0.13	0.13	0.27	0.27	0.13	0.22
21	(1) Causal Confidence	(22) Kappa	0.29	0.27	0.27	0.28	0.28	0.27	0.28
22	(1) Causal Confidence	(23) Klosgen	0.73	0.71	0.71	0.72	0.72	0.71	0.72
23	(1) Causal Confidence	(24) Laplace	0.96	0.97	0.97	0.97	0.97	0.97	0.97
24	(1) Causal Confidence	(25) Least Contradiction	0.65	0.85	0.85	0.75	0.75	0.85	0.78
25	(1) Causal Confidence	(26) Lerman	0.35	0.31	0.31	0.33	0.33	0.31	0.33
26	(1) Causal Confidence	(27) Lift	0.28	0.35	0.35	0.31	0.31	0.35	0.33
27	(1) Causal Confidence	(28) Loevinger	0.84	0.88	0.88	0.86	0.86	0.88	0.87
28	(1) Causal Confidence	(29) Mutual Information	0.38	0.1	0.1	0.24	0.24	0.1	0.2
29	(1) Causal Confidence	(30) Odd Multiplier	0.43	0.29	0.29	0.36	0.36	0.29	0.34
30	(1) Causal Confidence	(31) Odds Ratio	0.18	0.14	0.14	0.16	0.16	0.14	0.16
31	(1) Causal Confidence	(32) Pavillon	0.8	0.76	0.76	0.78	0.78	0.76	0.77
32	(1) Causal Confidence	(33) Phi-Coefficient	0.37	0.32	0.32	0.35	0.35	0.32	0.34
33	(1) Causal Confidence	(34) Putative Causal Dependency	0.82	0.78	0.78	0.8	0.8	0.78	0.79

Figure 5.21: A summary computed from all the rulesets (original and sample).

Inside

This module gives special or complementary views on the most interesting rules from a cluster. The options given include: (a) the distributions of each measure in the cluster, (b) correlation values and scatterplot, (c) correlation graph, (d) k_m most interesting rules of each measure, (e-f) union and (g-h) intersection of the k_m most interesting rules in the cluster, (i-j) union and (k-l) intersection of the k_m most interesting rules of the other clusters in related with the cluster. We will describe some important features. The value of k_m in our study is $k_m = 10$. The cardinalities and the rule form are also showed to illustrate the measure properties of understanding the way that is a measure ranks a rule.

- The first k_m most interesting rules of each measure

  This is a list of the k_m highest ranked rules for each measure from the currently analyzed cluster, providing also access to the values that the interestingness values have assigned for the selected rules (see Fig. 5.25 (d)).

- Intersection of the k_m most interesting rules in a cluster

  We choose the k_m most interesting rules of each cluster and give the user an overview of their intersection. The rank of each rule is used to validate the result. The Y-axis holds the rank of the rule for the corresponding measure. Each rule is represented with parallel coordinates among interestingness measure values (see Fig. 5.25 (g)(h)). We can see the intersection in a horizontal line and if we obtain many rules having the same rank value so we will print these rules with only one line. If the user want to capture a small group of the most interesting rules for making their decision, he/she can use these rules for their first choice.

- Union of the k_m most interesting rules in a cluster

  With the same technique as above, we introduce the union of the k_m most interesting rules in each cluster, in order to give the user a more specific view in the cluster. The measures have the set of highest ranks (more interesting) rely concentratively on the low value of the Y-axis (see Fig. 5.25 (e)(f)).

- Union of all the clusters in relation to the current cluster

  Based on the k_m most interesting rules in the current cluster, we draw the parallel coordinates of each rule on other clusters. The user can see the zone that is the most interesting with the highest value (low zone) as in the Sec. 5.4.3. The effect of the ten most interesting rules on the other clusters gives the user a general sampling of the entire cluster. With the union approach, many best rules may be presented and compared (see Fig. 5.25 (i)(j)).

- Intersection of all the clusters in relation to the current cluster

  By decreasing the quantity of the k_m most interesting rules in one cluster, we will observe the rank distribution. The intersection is less interesting than the union because we generally do not have any interesting zone. Using the intersection in relation to the current cluster is important when the user just finds a small set of interesting and close rules (see Fig. 5.25 (k)(l)).

Union/Intersection

As in the above analysis of the union or intersection of the k_m most interesting rules in a cluster. We give another view with all the set of measures. By the disagree between measures when the number of measures is considerable, the intersection is often empty.

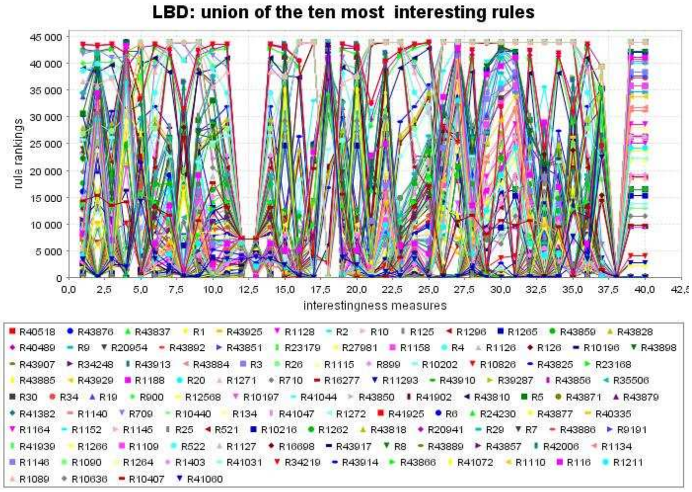


Figure 5.22: Union of the ten most interesting rules from all the measures on the LBD ruleset.

5.5 Display

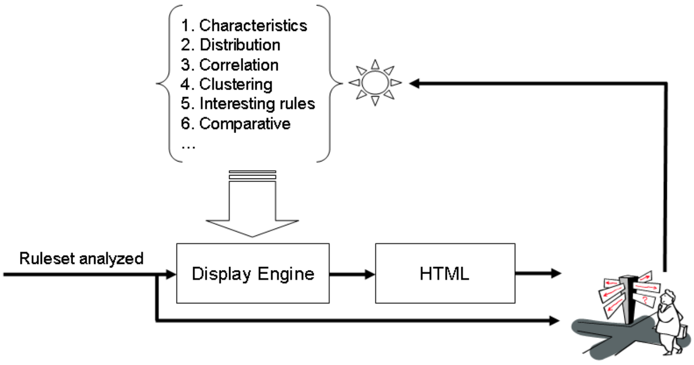
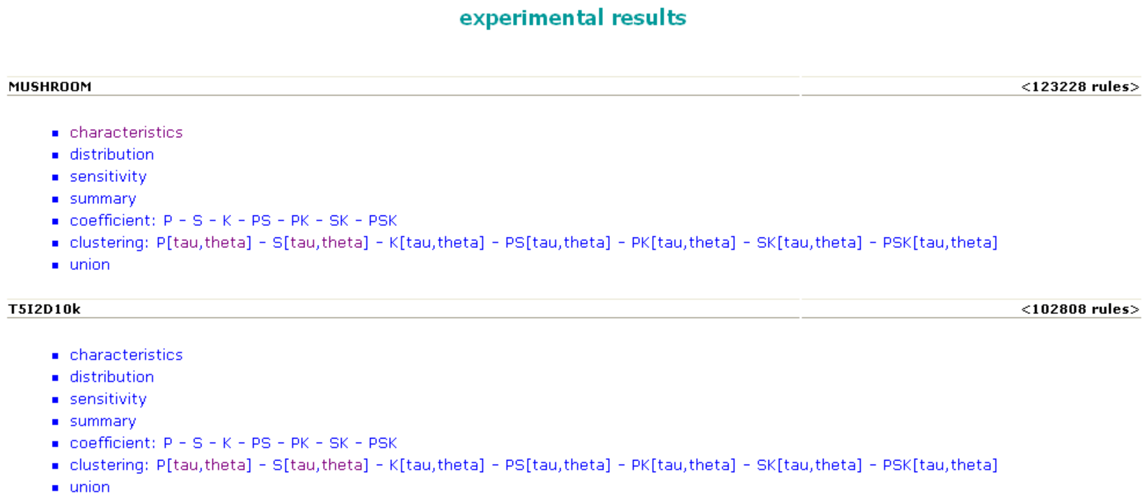


Figure 5.23: ARQAT Display.

The display module outputs all of the results in the HTML formats. The user can visualize and navigate through out a web browser to examine the aspects considered. Fig. 5.24 shows the principal links to access the evaluation results of a ruleset. All of the ruleset types such as original, sample, couple, and multiple are treated identically.



experimental results

MUSHROOM <123228 rules>

- characteristics
- distribution
- sensitivity
- summary
- coefficient: P - S - K - PS - PK - SK - PSK
- clustering: P[tau,theta] - S[tau,theta] - K[tau,theta] - PS[tau,theta] - PK[tau,theta] - SK[tau,theta] - PSK[tau,theta]
- union

T5I2D10k <102808 rules>

- characteristics
- distribution
- sensitivity
- summary
- coefficient: P - S - K - PS - PK - SK - PSK
- clustering: P[tau,theta] - S[tau,theta] - K[tau,theta] - PS[tau,theta] - PK[tau,theta] - SK[tau,theta] - PSK[tau,theta]
- union

Figure 5.24: Principal menu (extracted).

5.6 Utility

The utility module contains some common tools to facilitate the input/output process. It holds the descriptions of the files to analyze, the data file formats such as CSV, ARFF[4], PMML[5]. The common names, parameters used in the platform are also implemented int this module.

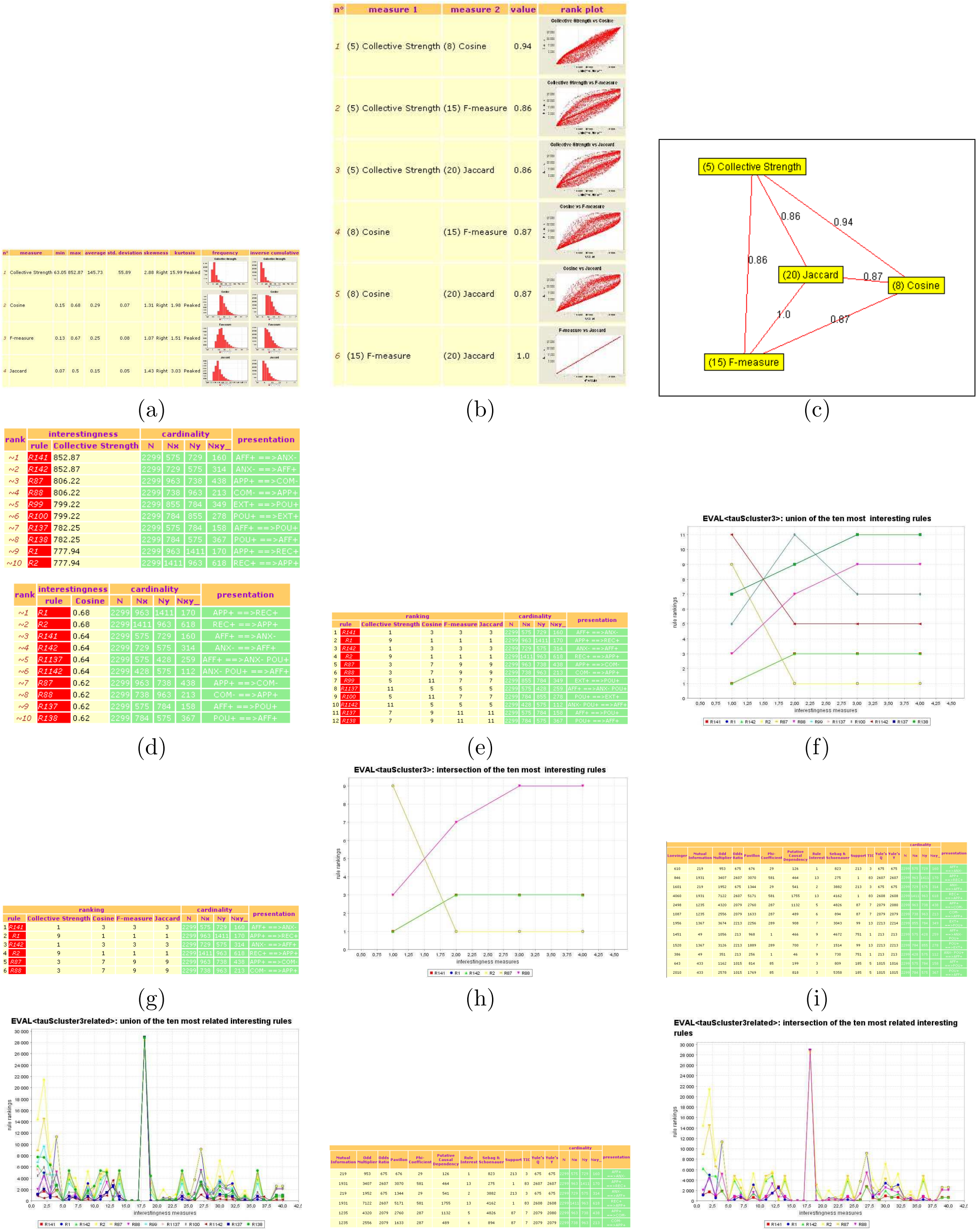
One of the most important component of this module is the statistics sub-package. This component implements the calculations of the τ, θ values and the other useful methods to compute the correlation values between measures such as Pearson's, Spearman's and Kendall's coefficients.

5.7 Summary

The ARQAT tool is organized in 14 complementary views with 5 analysis tasks. It processes the data via three stages preprocessing, evaluation and display. Three important correlation coefficients such as Pearson, Spearman and Kendall are implemented. A stable hierarchy is proposed to conduct comparative studies on several rulesets. A lot of views outputted in HTML files gives the user useful views to select the suitable measures or the most interesting rules. The intermediate data created during the process are stocked and be available to other postprocessing studies.

[4] ARFF and Sparse ARFF

[5] Supporting the current version PMML 3.0 with the help of parsing an XML document of the package XMLtp (http://mitglied.tripod.de/xmltp/).



n°	measure	min	max	average	std. deviation	skewness		kurtosis		frequency	inverse cumulative
1	Collective Strength	63.05	852.87	145.73	55.89	2.88	Right	15.99	Peaked	Collective Strength	Collective Strength
2	Cosine	0.15	0.68	0.29	0.07	1.31	Right	1.98	Peaked	Cosine	Cosine
3	F-measure	0.13	0.67	0.25	0.08	1.07	Right	1.51	Peaked	F-measure	F-measure
4	Jaccard	0.07	0.5	0.15	0.05	1.43	Right	3.03	Peaked	Jaccard	Jaccard

(a)

n°	measure 1	measure 2	value	rank plot
1	(5) Collective Strength	(8) Cosine	0.94	Collective Strength vs Cosine
2	(5) Collective Strength	(15) F-measure	0.86	Collective Strength vs F-measure
3	(5) Collective Strength	(20) Jaccard	0.86	Collective Strength vs Jaccard
4	(8) Cosine	(15) F-measure	0.87	Cosine vs F-measure
5	(8) Cosine	(20) Jaccard	0.87	Cosine vs Jaccard
6	(15) F-measure	(20) Jaccard	1.0	F-measure vs Jaccard

(b)

(c)

rank	interestingness		cardinality				presentation
	rule	Collective Strength	N	Nx	Ny	Nxy_	
~1	R141	852.87	2299	575	729	160	AFF+ ==>ANX-
~2	R142	852.87	2299	729	575	314	ANX- ==>AFF+
~3	R87	806.22	2299	963	738	438	APP+ ==>COM-
~4	R88	806.22	2299	738	963	213	COM- ==>APP+
~5	R99	799.22	2299	855	784	349	EXT+ ==>POU+
~6	R100	799.22	2299	784	855	278	POU+ ==>EXT+
~7	R137	782.25	2299	575	784	158	AFF+ ==>POU+
~8	R138	782.25	2299	784	575	367	POU+ ==>AFF+
~9	R1	777.94	2299	963	1411	170	APP+ ==>REC+
~10	R2	777.94	2299	1411	963	618	REC+ ==>APP+

rank	interestingness		cardinality				presentation
	rule	Cosine	N	Nx	Ny	Nxy_	
~1	R1	0.68	2299	963	1411	170	APP+ ==>REC+
~2	R2	0.68	2299	1411	963	618	REC+ ==>APP+
~3	R141	0.64	2299	575	729	160	AFF+ ==>ANX-
~4	R142	0.64	2299	729	575	314	ANX- ==>AFF+
~5	R1137	0.64	2299	575	428	259	AFF+ ==>ANX- POU+
~6	R1142	0.64	2299	428	575	112	ANX- POU+ ==>AFF+
~7	R87	0.62	2299	963	738	438	APP+ ==>COM-
~8	R88	0.62	2299	738	963	213	COM- ==>APP+
~9	R137	0.62	2299	575	784	158	AFF+ ==>POU+
~10	R138	0.62	2299	784	575	367	POU+ ==>AFF+

(d)

ranking						cardinality				presentation
rule		Collective Strength	Cosine	F-measure	Jaccard	N	Nx	Ny	Nxy_	
1	R141	1	3	3	3	2299	575	729	160	AFF+ ==>ANX-
2	R1	9	1	1	1	2299	963	1411	170	APP+ ==>REC+
3	R142	1	3	3	3	2299	729	575	314	ANX- ==>AFF+
4	R2	9	1	1	1	2299	1411	963	618	REC+ ==>APP+
5	R87	3	7	9	9	2299	963	738	438	APP+ ==>COM-
6	R88	3	7	9	9	2299	738	963	213	COM- ==>APP+
7	R99	5	11	7	7	2299	855	784	349	EXT+ ==>POU+
8	R1137	11	5	5	5	2299	575	428	259	AFF+ ==>ANX- POU+
9	R100	5	11	7	7	2299	784	855	278	POU+ ==>EXT+
10	R1142	11	5	5	5	2299	428	575	112	ANX- POU+ ==>AFF+
11	R137	7	9	11	11	2299	575	784	158	AFF+ ==>POU+
12	R138	7	9	11	11	2299	784	575	367	POU+ ==>AFF+

(e)

EVAL<tauScluster3>: union of the ten most interesting rules

(f)

ranking						cardinality				presentation
rule		Collective Strength	Cosine	F-measure	Jaccard	N	Nx	Ny	Nxy_	
1	R141	1	3	3	3	2299	575	729	160	AFF+ ==>ANX-
2	R1	9	1	1	1	2299	963	1411	170	APP+ ==>REC+
3	R142	1	3	3	3	2299	729	575	314	ANX- ==>AFF+
4	R2	9	1	1	1	2299	1411	963	618	REC+ ==>APP+
5	R87	3	7	9	9	2299	963	738	438	APP+ ==>COM-
6	R88	3	7	9	9	2299	738	963	213	COM- ==>APP+

(g)

EVAL<tauScluster3>: intersection of the ten most interesting rules

(h)

													cardinality				presentation
Loevinger	Mutual Information	Odd Multiplier	Odds Ratio	Pavillon	Phi-Coefficient	Putative Causal Dependency	Rule Interest	Sebag & Schoenauer	Support	TIC	Yule's Q	Yule's Y	N	Nx	Ny	Nxy_	
610	219	953	675	676	29	126	1	823	213	3	675	675	2299	575	729	160	AFF+ ==>ANX-
846	1931	3407	2607	3070	581	464	13	275	1	83	2607	2607	2299	963	1411	170	APP+ ==>REC+
1601	219	1952	675	1344	29	541	2	3882	213	3	675	675	2299	729	575	314	ANX- ==>AFF+
4060	1931	7122	2607	5171	581	1755	13	4162	1	83	2608	2608	2299	1411	963	618	REC+ ==>APP+
2498	1235	4320	2079	2760	287	1132	5	4826	87	7	2079	2080	2299	963	738	438	APP+ ==>COM-
1087	1235	2556	2079	1633	287	489	6	894	87	7	2079	2079	2299	738	963	213	COM- ==>APP+
1956	1367	3674	2213	2256	289	908	7	3043	99	13	2213	2214	2299	855	784	349	EXT+ ==>POU+
1451	49	1056	213	968	1	466	9	4672	751	1	213	213	2299	575	428	259	AFF+ ==>ANX- POU+
1520	1367	3126	2213	1889	289	700	7	1514	99	13	2213	2213	2299	784	855	278	POU+ ==>EXT+
386	49	351	213	256	1	46	9	730	751	1	213	213	2299	428	575	112	ANX- POU+ ==>AFF+
643	433	1162	1015	814	85	199	3	809	185	5	1015	1016	2299	575	784	158	AFF+ ==>POU+
2010	433	2578	1015	1769	85	818	3	5358	185	5	1015	1015	2299	784	575	367	POU+ ==>AFF+

(i)

EVAL<tauScluster3related>: union of the ten most related interesting rules

(j)

												cardinality				presentation
Mutual Information	Odd Multiplier	Odds Ratio	Pavillon	Phi-Coefficient	Putative Causal Dependency	Rule Interest	Sebag & Schoenauer	Support	TIC	Yule's Q	Yule's Y	N	Nx	Ny	Nxy_	
219	953	675	676	29	126	1	823	213	3	675	675	2299	575	729	160	AFF+ ==>ANX-
1931	3407	2607	3070	581	464	13	275	1	83	2607	2607	2299	963	1411	170	APP+ ==>REC+
219	1952	675	1344	29	541	2	3882	213	3	675	675	2299	729	575	314	ANX- ==>AFF+
1931	7122	2607	5171	581	1755	13	4162	1	83	2608	2608	2299	1411	963	618	REC+ ==>APP+
1235	4320	2079	2760	287	1132	5	4826	87	7	2079	2080	2299	963	738	438	APP+ ==>COM-
1235	2556	2079	1633	287	489	6	894	87	7	2079	2079	2299	738	963	213	COM- ==>APP+

(k)

EVAL<tauScluster3related>: intersection of the ten most related interesting rules

(l)

Figure 5.25: A set of complementary views on a cluster.

Chapter 6

Studies: on general evaluation

Making comparisons from the postprocessing of association rules have become a research challenge in KDD. By evaluating interestingness values calculated from interestingness measures on association rules, some approaches are proposed in this chapter to answer the questions: How we can capture the different aspects on the datasets via the behaviors of interestingness measures.

6.1 Datasets

A set of four datasets are collected, in which two datasets have opposite characteristics (i.e. correlated vs weakly corelated) and the others are two real-life datasets. Tab. 6.1 gives a quick description on theses four datasets studied.

Dataset	Number of items	Transactions	
		Total	Average length
$\mathcal{D}_1$	128	8416	23
$\mathcal{D}_2$	81	9650	5
$\mathcal{D}_3$	92	2883	8.5
$\mathcal{D}_4$	30	2299	10

Table 6.1: Dataset description.

- Mushroom

  The categorical Mushroom dataset ($\mathcal{D}_1$) from Irvine machine-learning database repository [NHBM98] has 23 nominal attributes corresponding to the species of gilled mushrooms (i.e., edible or poisonous).

- T5I2D10k

  The synthetic T5I2D10k dataset ($\mathcal{D}_2$) is obtained by simulating the transactions of customers in retailing businesses. The dataset was generated using the IBM synthetic data generator [AS94]. $\mathcal{D}_2$ has the typical characteristic of the Agrawal dataset T5I2D10k (T5: average size of the transactions is 5, I2: average size of the maximal potentially large itemsets is 2, D10k: number of items is 100).

- LBD

  The LBD dataset ($\mathcal{D}_3$) is a set of lift breakdowns from the breakdown service of a lift manufacturer.

- EVAL

  The EVAL dataset ($\mathcal{D}_4$) is a dataset of profiles of worker's performances which was used by the company *PerformanSe* to calibrate a decision support system in human resource management [Fle96].

The T5I2D10k synthetic dataset is generated according to the properties of market basket data that are typical weakly correlated data. In this dataset, the number of frequent patterns is small in compared with the total number of patterns. The MUSHROOM dataset constituted of correlated data, the proportion of patterns that are frequent is important.

The two datasets LBD and EVAL are called real-life datasets because both of them are issued from the real activities captured. While the other two datasets, T5I2D10k and MUSHROOM, are generated by simulating the real activities. For example, the size of a next transaction of the T5I2D10k is constructed by approximating a Poisson distribution with mean μ equal to T.

6.2 Rulesets

From the datasets discussed in the above section, the corresponding rulesets (i.e., the set of association rules) are generated with the rule mining techniques. Some abstract rulesets are also extracted to evaluate the behavior of interestingness measures[1].

Ruleset	Number of rules	θ	τ
$\mathcal{R}_1$	123228	0.005	0.85
$\mathcal{R}_2$	102808	0.006	0.85
$\mathcal{R}_3$	43930	0.009	0.85
$\mathcal{R}_4$	28938	0.011	0.85

(a) Original

Ruleset	Number of rules	θ	τ
$\mathcal{R}'_1$	11213	0.0185	0.85
$\mathcal{R}'_2$	8006	0.0219	0.85
$\mathcal{R}'_3$	8868	0.0208	0.85
$\mathcal{R}'_4$	6092	0.0251	0.85

(b) Sample

Table 6.2: Rulesets extracted.

- ORIGINAL

  The ruleset $\mathcal{R}_1$ is generated (respectively $\mathcal{R}_2$, $\mathcal{R}_3$, $\mathcal{R}_4$) from the dataset $\mathcal{D}_1$ (respectively $\mathcal{D}_2$, $\mathcal{D}_3$, $\mathcal{D}_4$) using the APRIORI algorithm [AS94][AMS$^+$96] (see Tab. 6.2 (a)).

- SAMPLE

  A sample ruleset $\mathcal{R}_1^{'}$ (respectively $\mathcal{R}_2^{'}$, $\mathcal{R}_3^{'}$, $\mathcal{R}_4^{'}$) of each the above ruleset is extracted from the $\mathcal{R}_1$ ruleset (respectively $\mathcal{R}_2$, $\mathcal{R}_3$, $\mathcal{R}_4$) as the union of the first 1000 rules ($\approx$ 1%, ordered by decreasing interestingness values) issued from each objective measures (see Tab. 6.2 (b)).

[1] In this chapter and the next chapter, we use the notion of interestingness measures and objective measures interchangeable.

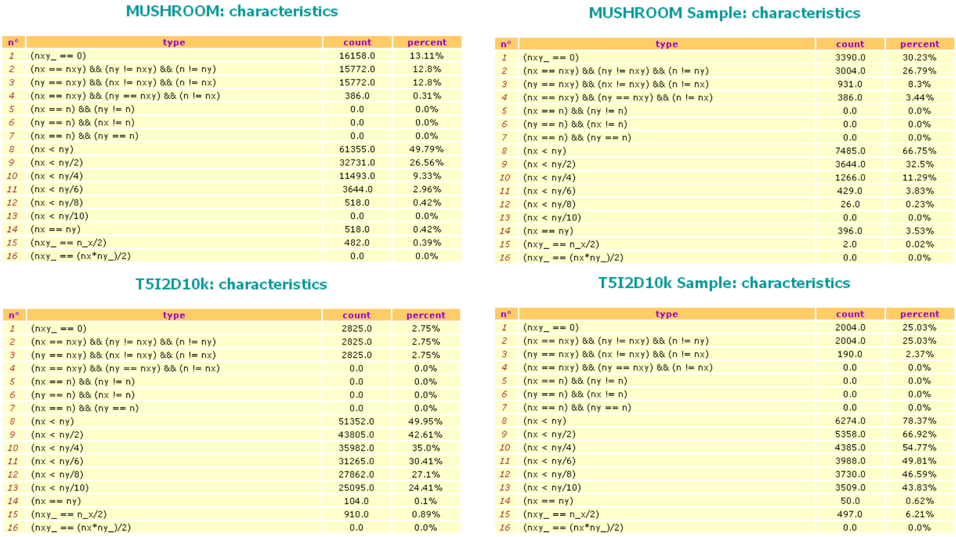
- Complement

  To evaluate the stable behavior of the measures on several rulesets, we have established some abstract rulesets namely *couple* rulesets ($\mathcal{R}_k$ vs $\mathcal{R}'_k$) and *multiple* rulesets (*All*, *AllOriginal*, and *AllSample*). These two type of rulesets are called *complement* rulesets (see Sec. 5.4.3 [Chap. 5]).

6.3 Used measures

In our experiment, we compared and analyzed the 40 objective measures defined in Tab. 3.1 [Chap. 3]. We must notice that EII($\alpha = 1$) and EII($\alpha = 2$) are two entropic versions of the II measure.

6.4 Efficiency of the sample model



MUSHROOM: characteristics

n°	type	count	percent
1	(nxy_ == 0)	16158.0	13.11%
2	(nx == nxy) && (ny != nxy) && (n != ny)	15772.0	12.8%
3	(ny == nxy) && (nx != nxy) && (n != nx)	15772.0	12.8%
4	(nx == nxy) && (ny == nxy) && (n != nx)	386.0	0.31%
5	(nx == n) && (ny != n)	0.0	0.0%
6	(ny == n) && (nx != n)	0.0	0.0%
7	(nx == n) && (ny == n)	0.0	0.0%
8	(nx < ny)	61355.0	49.79%
9	(nx < ny/2)	32731.0	26.56%
10	(nx < ny/4)	11493.0	9.33%
11	(nx < ny/6)	3644.0	2.96%
12	(nx < ny/8)	518.0	0.42%
13	(nx < ny/10)	0.0	0.0%
14	(nx == ny)	518.0	0.42%
15	(nxy_ == n_x/2)	482.0	0.39%
16	(nxy_ == (nx*ny_)/2)	0.0	0.0%

MUSHROOM Sample: characteristics

n°	type	count	percent
1	(nxy_ == 0)	3390.0	30.23%
2	(nx == nxy) && (ny != nxy) && (n != ny)	3004.0	26.79%
3	(ny == nxy) && (nx != nxy) && (n != nx)	931.0	8.3%
4	(nx == nxy) && (ny == nxy) && (n != nx)	386.0	3.44%
5	(nx == n) && (ny != n)	0.0	0.0%
6	(ny == n) && (nx != n)	0.0	0.0%
7	(nx == n) && (ny == n)	0.0	0.0%
8	(nx < ny)	7485.0	66.75%
9	(nx < ny/2)	3644.0	32.5%
10	(nx < ny/4)	1266.0	11.29%
11	(nx < ny/6)	429.0	3.83%
12	(nx < ny/8)	26.0	0.23%
13	(nx < ny/10)	0.0	0.0%
14	(nx == ny)	396.0	3.53%
15	(nxy_ == n_x/2)	2.0	0.02%
16	(nxy_ == (nx*ny_)/2)	0.0	0.0%

T5I2D10k: characteristics

n°	type	count	percent
1	(nxy_ == 0)	2825.0	2.75%
2	(nx == nxy) && (ny != nxy) && (n != ny)	2825.0	2.75%
3	(ny == nxy) && (nx != nxy) && (n != nx)	2825.0	2.75%
4	(nx == nxy) && (ny == nxy) && (n != nx)	0.0	0.0%
5	(nx == n) && (ny != n)	0.0	0.0%
6	(ny == n) && (nx != n)	0.0	0.0%
7	(nx == n) && (ny == n)	0.0	0.0%
8	(nx < ny)	51352.0	49.95%
9	(nx < ny/2)	43805.0	42.61%
10	(nx < ny/4)	35982.0	35.0%
11	(nx < ny/6)	31265.0	30.41%
12	(nx < ny/8)	27862.0	27.1%
13	(nx < ny/10)	25095.0	24.41%
14	(nx == ny)	104.0	0.1%
15	(nxy_ == n_x/2)	910.0	0.89%
16	(nxy_ == (nx*ny_)/2)	0.0	0.0%

T5I2D10k Sample: characteristics

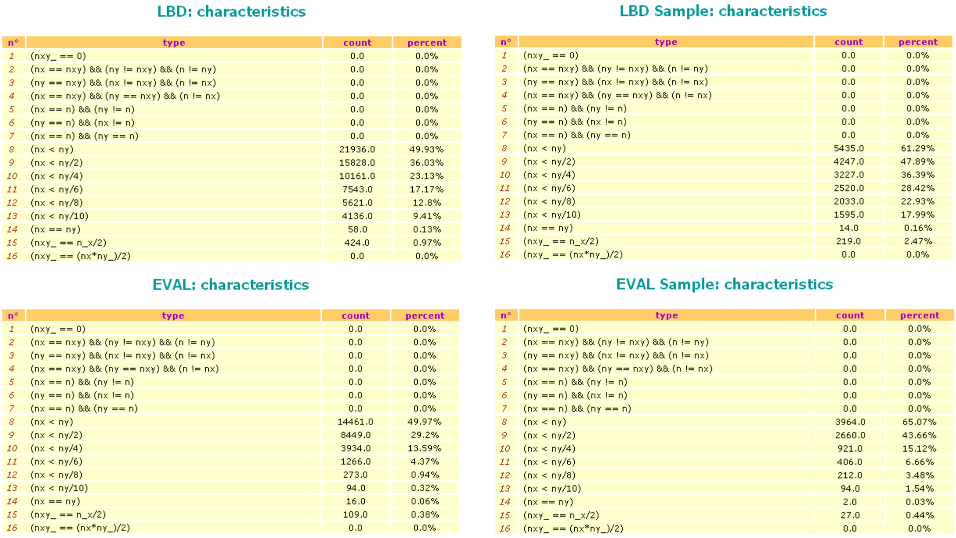
n°	type	count	percent
1	(nxy_ == 0)	2004.0	25.03%
2	(nx == nxy) && (ny != nxy) && (n != ny)	2004.0	25.03%
3	(ny == nxy) && (nx != nxy) && (n != nx)	190.0	2.37%
4	(nx == nxy) && (ny == nxy) && (n != nx)	0.0	0.0%
5	(nx == n) && (ny != n)	0.0	0.0%
6	(ny == n) && (nx != n)	0.0	0.0%
7	(nx == n) && (ny == n)	0.0	0.0%
8	(nx < ny)	6274.0	78.37%
9	(nx < ny/2)	5358.0	66.92%
10	(nx < ny/4)	4385.0	54.77%
11	(nx < ny/6)	3988.0	49.81%
12	(nx < ny/8)	3730.0	46.59%
13	(nx < ny/10)	3509.0	43.83%
14	(nx == ny)	50.0	0.62%
15	(nxy_ == n_x/2)	497.0	6.21%
16	(nxy_ == (nx*ny_)/2)	0.0	0.0%

Figure 6.1: Characteristics of the rulesets.

Based on the characteristics analyzed on the four original rulesets, we will evaluate the efficiency of the sample rulesets. One of the most easily method is to examine the sample rules with the same regard to the properties of objective measures.

Before evaluating the efficiency of the sample model, we can examine some ruleset characteristics showing the distributions underlying rule cardinalities, in order to detect "borderline cases". For instance, in Fig. 6.1 and Fig. 6.2, the first line on each ruleset gives the number of "logical" rules (i.e. rules without negative examples). We can notice that the number of logical rules is here very high (≈13%) on the $\mathcal{R}_1$ ruleset, but very low or zero on the others.

One of the most important properties discussed in Chap. 3 on objective measures is that a rule is seemed to be interesting if its cardinalities satisfy the condition $n_X < n_Y$. We can see from row 8 to row 13 in the characteristic table (Fig. 6.1 and Fig. 6.2), the percents on the right column (i.e. the sample rulesets) is higher on the left column (i.e. the original rulesets). Among these sample



LBD: characteristics

n°	type	count	percent
1	(nxy_ == 0)	0.0	0.0%
2	(nx == nxy) && (ny != nxy) && (n != ny)	0.0	0.0%
3	(ny == nxy) && (nx != nxy) && (n != nx)	0.0	0.0%
4	(nx == nxy) && (ny == nxy) && (n != nx)	0.0	0.0%
5	(nx == n) && (ny != n)	0.0	0.0%
6	(ny == n) && (nx != n)	0.0	0.0%
7	(nx == n) && (ny == n)	0.0	0.0%
8	(nx < ny)	21936.0	49.93%
9	(nx < ny/2)	15828.0	36.03%
10	(nx < ny/4)	10161.0	23.13%
11	(nx < ny/6)	7543.0	17.17%
12	(nx < ny/8)	5621.0	12.8%
13	(nx < ny/10)	4136.0	9.41%
14	(nx == ny)	58.0	0.13%
15	(nxy_ == n_x/2)	424.0	0.97%
16	(nxy_ == (nx*ny_)/2)	0.0	0.0%

LBD Sample: characteristics

n°	type	count	percent
1	(nxy_ == 0)	0.0	0.0%
2	(nx == nxy) && (ny != nxy) && (n != ny)	0.0	0.0%
3	(ny == nxy) && (nx != nxy) && (n != nx)	0.0	0.0%
4	(nx == nxy) && (ny == nxy) && (n != nx)	0.0	0.0%
5	(nx == n) && (ny != n)	0.0	0.0%
6	(ny == n) && (nx != n)	0.0	0.0%
7	(nx == n) && (ny == n)	0.0	0.0%
8	(nx < ny)	5435.0	61.29%
9	(nx < ny/2)	4247.0	47.89%
10	(nx < ny/4)	3227.0	36.39%
11	(nx < ny/6)	2520.0	28.42%
12	(nx < ny/8)	2033.0	22.93%
13	(nx < ny/10)	1595.0	17.99%
14	(nx == ny)	14.0	0.16%
15	(nxy_ == n_x/2)	219.0	2.47%
16	(nxy_ == (nx*ny_)/2)	0.0	0.0%

EVAL: characteristics

n°	type	count	percent
1	(nxy_ == 0)	0.0	0.0%
2	(nx == nxy) && (ny != nxy) && (n != ny)	0.0	0.0%
3	(ny == nxy) && (nx != nxy) && (n != nx)	0.0	0.0%
4	(nx == nxy) && (ny == nxy) && (n != nx)	0.0	0.0%
5	(nx == n) && (ny != n)	0.0	0.0%
6	(ny == n) && (nx != n)	0.0	0.0%
7	(nx == n) && (ny == n)	0.0	0.0%
8	(nx < ny)	14461.0	49.97%
9	(nx < ny/2)	8449.0	29.2%
10	(nx < ny/4)	3934.0	13.59%
11	(nx < ny/6)	1266.0	4.37%
12	(nx < ny/8)	273.0	0.94%
13	(nx < ny/10)	94.0	0.32%
14	(nx == ny)	16.0	0.06%
15	(nxy_ == n_x/2)	109.0	0.38%
16	(nxy_ == (nx*ny_)/2)	0.0	0.0%

EVAL Sample: characteristics

n°	type	count	percent
1	(nxy_ == 0)	0.0	0.0%
2	(nx == nxy) && (ny != nxy) && (n != ny)	0.0	0.0%
3	(ny == nxy) && (nx != nxy) && (n != nx)	0.0	0.0%
4	(nx == nxy) && (ny == nxy) && (n != nx)	0.0	0.0%
5	(nx == n) && (ny != n)	0.0	0.0%
6	(ny == n) && (nx != n)	0.0	0.0%
7	(nx == n) && (ny == n)	0.0	0.0%
8	(nx < ny)	3964.0	65.07%
9	(nx < ny/2)	2660.0	43.66%
10	(nx < ny/4)	921.0	15.12%
11	(nx < ny/6)	406.0	6.66%
12	(nx < ny/8)	212.0	3.48%
13	(nx < ny/10)	94.0	1.54%
14	(nx == ny)	2.0	0.03%
15	(nxy_ == n_x/2)	27.0	0.44%
16	(nxy_ == (nx*ny_)/2)	0.0	0.0%

Figure 6.2: Characteristics of the rulesets (cont.).

rulesets, the weekly correlated ruleset $\mathcal{R}_2'$ has the most efficiency with the value $(n_X < n_Y)$ goes from 50% to 78% and $(n_X < \frac{n_Y}{10})$ goes from 24% to 44%.

6.5 Distribution of interestingness values

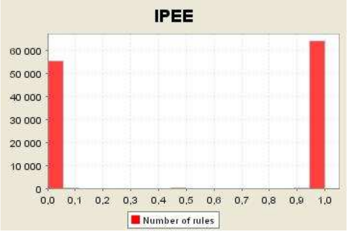
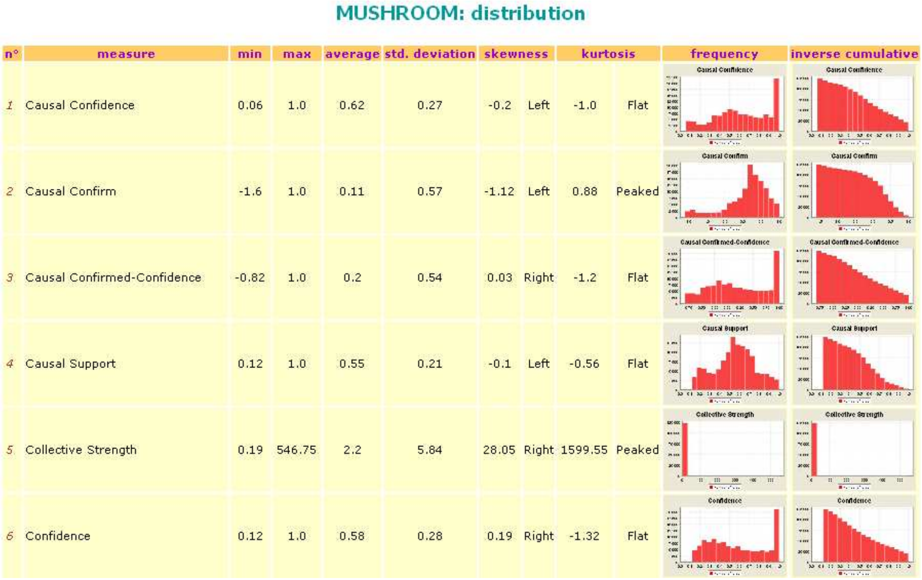


Figure 6.3: Variation of the measure IPEE on the $\mathcal{R}_1$ ruleset.

The interestingness distribution view (Sec. 5.4.1 [Chap. 5]), draws the histograms for each measure. The distributions are also completed with classically statistical indexes such as minimum, maximum, average, standard deviation, skewness and kurtosis values. In Fig. 6.4, one can see that Causal Confidence (line 1), Causal Confirmed-Confidence (line 3), Confidence (line 6) have an irregular distribution and a great number of rules with 100% confidence while it is very different from Causal Confirm (line 2).

An observation on the interestingness values from the all the measures on the $\mathcal{R}_1$ ruleset gives us three principal zones of distribution (Tab. 6.3). The first zone consists of the measures which interestingness values concentrate on the left – trending towards their minimum values –, on the right – trending towards their maximum values –, or on the center – trending towards their average



MUSHROOM: distribution

n°	measure	min	max	average	std. deviation	skewness		kurtosis		frequency	inverse cumulative
1	Causal Confidence	0.06	1.0	0.62	0.27	-0.2	Left	-1.0	Flat	Causal Confidence	Causal Confidence
2	Causal Confirm	-1.6	1.0	0.11	0.57	-1.12	Left	0.88	Peaked	Causal Confirm	Causal Confirm
3	Causal Confirmed-Confidence	-0.82	1.0	0.2	0.54	0.03	Right	-1.2	Flat	Causal Confirmed-Confidence	Causal Confirmed-Confidence
4	Causal Support	0.12	1.0	0.55	0.21	-0.1	Left	-0.56	Flat	Causal Support	Causal Support
5	Collective Strength	0.19	546.75	2.2	5.84	28.05	Right	1599.55	Peaked	Collective Strength	Collective Strength
6	Confidence	0.12	1.0	0.58	0.28	0.19	Right	-1.32	Flat	Confidence	Confidence

Figure 6.4: Distribution of some measures on the $\mathcal{R}_1$ ruleset (extracted).

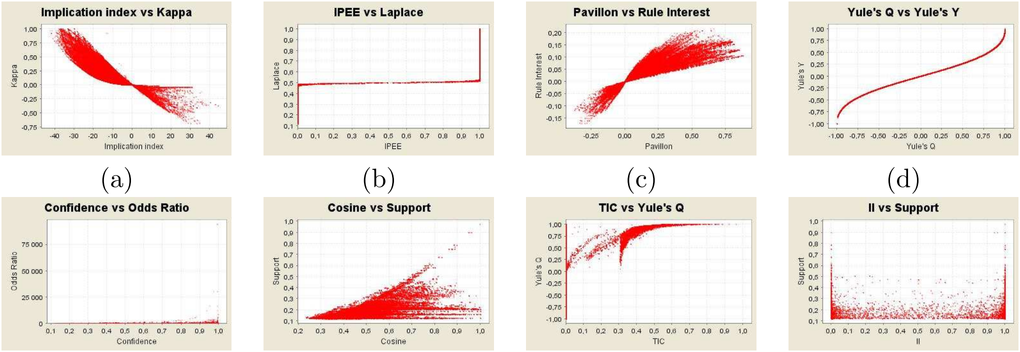
values –. Besides, IPEE has two unconnected poles of interestingness values: at minimum trend and at maximum trend (see Fig. 6.3).

Trend	Mesures
Left	Collective Strength, Conviction, Dependency, EII, EII 2, Gini-index, Jaccard, J-measure, Lift, Mutual Information, Odd Multilier, Odds Ratio, Sebag & Schoenauer, Support, TIC
Right	Causal Confidence, Causal Confirmed-Confidence, Confidence, Descriptive Confirmed-Confidence, Example & Contra-Example, II, IPEE, Laplace, Least Contradiction, Loevinger, Yule's Q
Center	Causal Confirm, Causal Support, Cosine, Descriptive Confirm, F-measure, Implication index, Kappa, Klosgen, Lerman, Pavillon, Phi-Coefficient, Putative Causal Dependency, Rule Interest, Yule's Y

Table 6.3: Trends on interestingness values

6.6 Joint-distribution matrix

The joint-distribution analysis shows the scatterplot matrix of all measure pairs. This graphical matrix is very useful to see the details of the relationships between measures. These relationships also depend on the seven summary types P, S, K, PS, PK, SK, PSK (Sec. 5.4.2 [Chap. 5]). For instance, Fig. 6.6 shows a small part in the scatterplot matrix on the interestingness values between measures.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 6.5: Some agreement and disagreement shapes.

Fig. 6.5 shows four agreement shapes: Implication index vs Kappa (a), IPEE vs Laplace (b), Pavillon vs Rule Interest (c), and Yule's Q vs Yule's Y (7) (highly correlated). On the other hand, four disagreement shapes on Confidence vs Odds Ratio (e), Cosine vs Support (f), TIC vs Yue's Q (g), and II vs Support (h) (highly uncorrelated) is also be noticed.

6.7 Correlation analysis

This task aims at delivering the clustering of measures and facilitating the choice of a subset of measures that is best-adapted to describe the ruleset. The correlation values between measure pairs are computed by using the Pearson's, Spearman's or Kendall's correlation coefficients and stored in the correlation matrix ($\mathcal{M} \times \mathcal{M}$). Two visual representations are proposed. The first one is a simple summary matrix in which each significant correlation value is visually associated with a different color (a level of gray). For instance, the dark cell from Fig. 6.7 shows a low correlation value between II and Support. The other 110 gray cells correspond to high correlation values.

The second one (Fig. 6.8) is a graph-based view of the correlation matrix, called correlation graph (Sec. 5.4.2 [Chap. 5]) with two types: CG0 and CG+.

These two subgraphs can then be processed in order to extract clusters of measures: each cluster is defined as a connected subgraph. In CG+, each cluster gathers correlated or anti-correlated measures that may be interpreted similarly: they deliver a close point of view on data. Moreover, in CG0, each cluster contains uncorrelated measures, i.e. measures that deliver a different point of view.

Hence, as each graph depends on a specific ruleset, the user can use the graphs as data insight, which graphically help him/her select the minimal set of the measures best adapted to his/her data. For instance in Fig. 6.8, CG+ graph contains 10 clusters on 40 measures. The user can select the most representative measure in each cluster (Sec. 4.6 [Chap. 4]), and then retain it to validate the rules.

A close observation on the CG0 graph (Fig. 6.8) shows an uncorrelated cluster formed by the measures II and Support (also the two dark cells in Fig. 6.7). This observation is confirmed on Fig. 6.6, particularly with the scatterplot (h) in Fig. 6.5. CG+ graph shows a trivial cluster where Yule's Q and Yule's Y are strongly correlated. This is also confirmed in Fig. 6.5 (d), showing a functional dependency between the two measures. These two examples show the interest of using

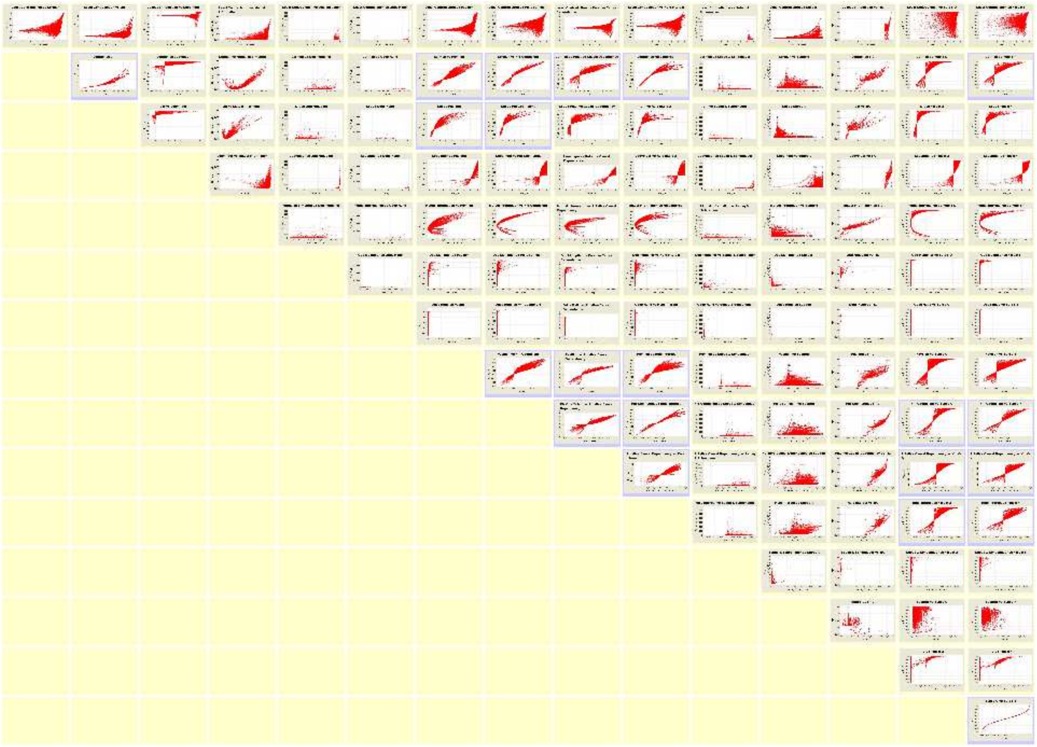


Figure 6.6: Scatterplot matrix of joint-distributions on the $\mathcal{R}_1$ ruleset (extracted).

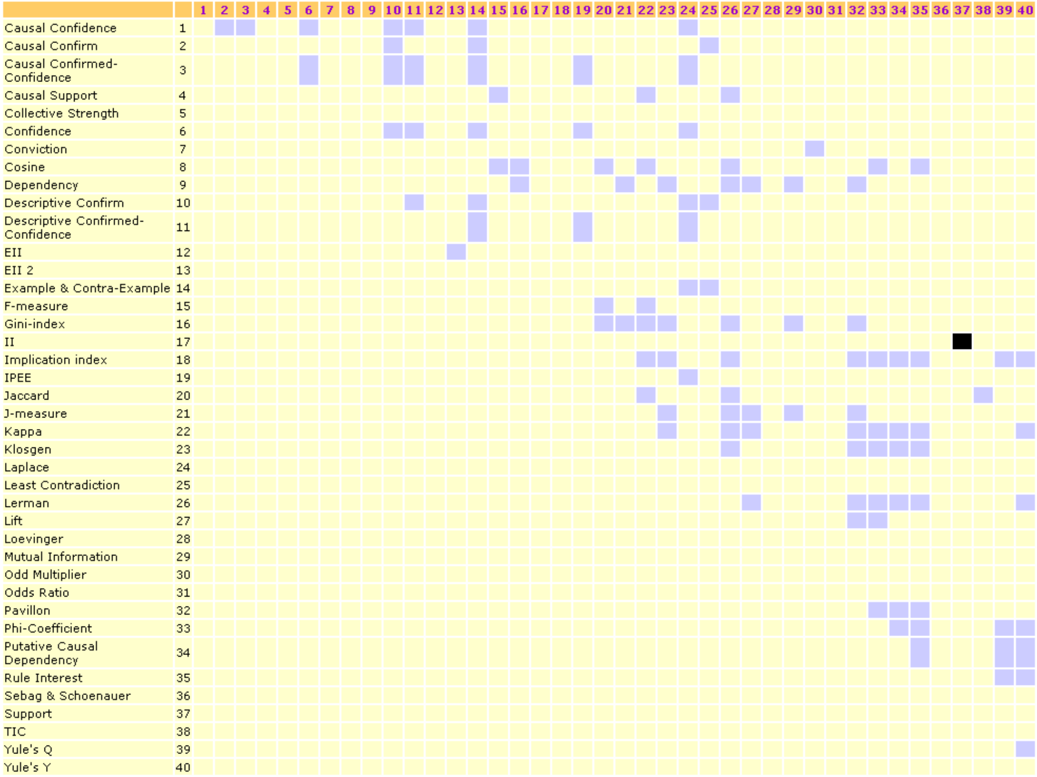
the scatterplot matrix complementarily (Fig. 6.6) with the correlation graphs CG0, CG+ (Fig. 6.8) in order to evaluate the nature of the correlation links, and overcome the limits of the correlation coefficient.

6.8 Interesting rule analysis

In order to help a user select the most interesting rules, two specific views are implemented. The first view (Fig. 6.9) collects a set of a given number of interesting rules for each measure in one cluster, in order to answer the question: how interesting are the rules according to this cluster. The selected rules can be visualized with parallel coordinate drawing (Fig. 6.10) alternatively. The main interest of such a drawing is to rapidly see the measure rankings of the rules.

These two views can be used with the measure values of a rule or alternatively with the rank of the value. For instance, Fig. 6.9 and Fig. 6.10 use the rank to evaluate the union, for all the measures in cluster C1, of the ten most interesting rules for each measure (see Fig. 6.8). The Y-axis in Fig. 6.10 holds the rule rank for the corresponding measure. By observing the concentration lines on low rank values, one can obtain four measures: Confidence(4 → 6[2]), Descriptive Confirmed-Confidence(6 → 11), Example & Contra-Example(7 → 14), and IPEE (10 → 19) (on points 1, 2, 3, 4 respectively) that are good for a majority of interesting rules. This can also be retrieved from the

[2] 4 → 6: 4 is the order of the measures in the corresponding cluster, 6 is the exact position of the measure in the table of all objective measures studied.



		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
Causal Confidence	1		▨	▨			▨				▨	▨			▨										▨																
Causal Confirm	2										▨				▨											▨															
Causal Confirmed-Confidence	3						▨				▨	▨			▨					▨					▨																
Causal Support	4															▨							▨				▨														
Collective Strength	5																																								
Confidence	6										▨	▨			▨					▨					▨																
Conviction	7																														▨										
Cosine	8															▨	▨				▨		▨				▨							▨		▨					
Dependency	9																▨					▨		▨			▨	▨		▨			▨								
Descriptive Confirm	10											▨			▨										▨	▨															
Descriptive Confirmed-Confidence	11														▨					▨					▨																
EII	12													▨																											
EII 2	13																																								
Example & Contra-Example	14																								▨	▨															
F-measure	15																				▨		▨																		
Gini-index	16																				▨	▨	▨	▨			▨			▨			▨								
II	17																																					■			
Implication index	18																						▨	▨			▨						▨	▨	▨	▨				▨	▨
IPEE	19																								▨																
Jaccard	20																						▨				▨												▨		
J-measure	21																							▨			▨	▨		▨			▨								
Kappa	22																							▨			▨	▨					▨	▨	▨	▨					▨
Klosgen	23																										▨						▨	▨	▨	▨					
Laplace	24																																								
Least Contradiction	25																																								
Lerman	26																											▨					▨	▨	▨	▨					▨
Lift	27																																▨	▨							
Loevinger	28																																								
Mutual Information	29																																								
Odd Multiplier	30																																								
Odds Ratio	31																																								
Pavillon	32																																	▨	▨	▨					
Phi-Coefficient	33																																		▨	▨				▨	▨
Putative Causal Dependency	34																																			▨				▨	▨
Rule Interest	35																																							▨	▨
Sebag & Schoenauer	36																																								
Support	37																																								
TIC	38																																								
Yule's Q	39																																								▨
Yule's Y	40																																								

Figure 6.7: Coefficient matrix on the $\mathcal{R}_1$ ruleset.

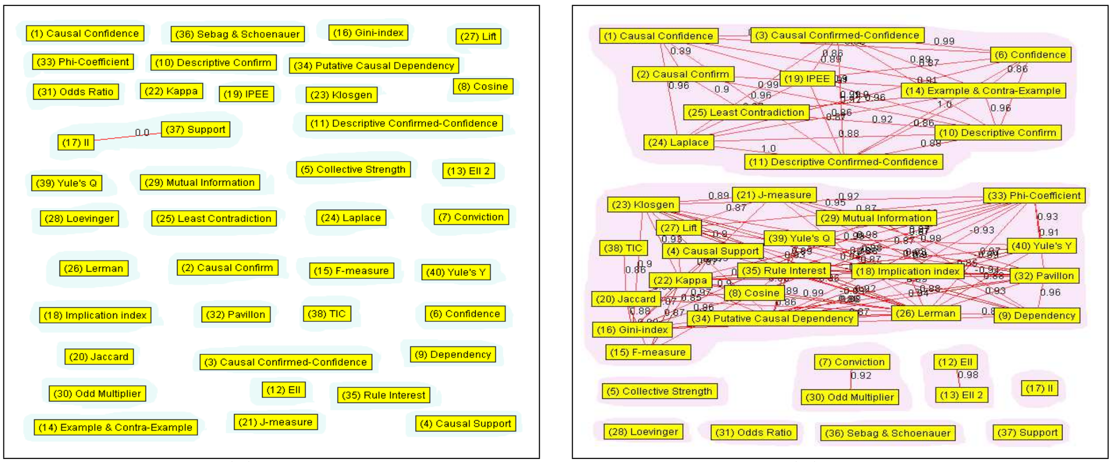
columns with the names of the corresponding measures in Fig. 6.9. Among these four measures, IPEE is the most suitable choice because of the lowest rule ranks obtained.

6.9 Ranking measures by sensitivity values

6.9.1 On a ruleset

The sensitivity evaluation is based on the number of rules that falls in each interval is compared to rank the measures . For a measure on a ruleset, the most significance interval will be the last bin (i.e., interval) of the inversely cumulative distribution. To have an approximation view on the sensitivity value, the number of rules has the maximum value is also retained. Fig. 6.11 shows the first seven measures that obtain the highest ranks. A remark is that the number of rules in the first interval is not always the same for all the measures because of the affectation of the number of NaN values.

An example of ranking two measures is given in Fig. 6.12 on the $\mathcal{R}_1$ ruleset. The measure *Implication index* is ranked at the 13^{th} place from a set of 40 measures while the measure *Rule Interest* is ranked at the 14^{th} place. The meaning for this ranking is that the measure Implication index is more sensitive than the measure Rule Interest on the $\mathcal{R}_1$ ruleset even if the number of the



(a) CG0 (b) CG+

Figure 6.8: CG0 and CG+ graphs on the $\mathcal{R}_1$ ruleset (clusters are highlighted with a gray background).

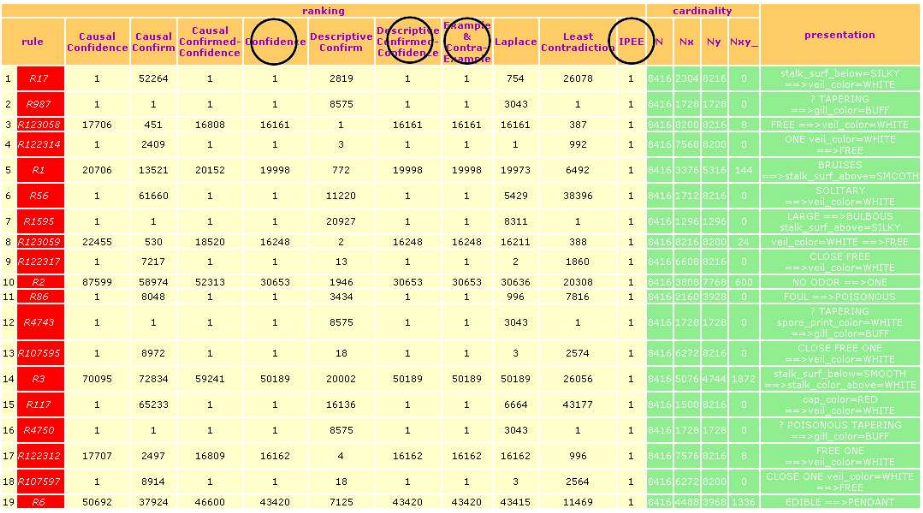
most interesting rules returned with the maximum value is greater for the measure Rule Interest ($3 > 2$). The differences counted from each couple intervals, beginning from the last interval are quite important because the user will feel easier when looking at 11 rules in the last interval of the measure Implication index instead of looking at 64 rules from the same interval of the measure Rule Interest.

6.9.2 On a set of rulesets

In Fig. 6.13, we can see the measure Implication Index goes strongly from place 13^{rd} in the $\mathcal{R}_1$ ruleset to place 9^{th} over all the original rulesets while the measure Rule Interest goes lightly from place 14^{th} to place 13^{rd}.

6.10 Summary

This chapter gives an overview on some aspects of the datasets studied (e.g. characteristics, ...) to the user. The efficiency of the sample model, variation of interestingness values, joint-distribution matrix, correlation analysis, interesting rules analysis, and ranking measures by sensitivity values both on a ruleset and a set of rulesets are represented.



	ranking											cardinality				
	rule	Causal Confidence	Causal Confirm	Causal Confirmed-Confidence	Confidence	Descriptive Confirm	Descriptive Confirmed-Confidence	Example & Contra-Example	Laplace	Least Contradiction	IPEE	N	Nx	Ny	Nxy_	presentation
1	R17	1	52264	1	1	2819	1	1	754	26078	1	8416	2304	8216	0	stalk_surf_below=SILKY ==>veil_color=WHITE
2	R987	1	1	1	1	8575	1	1	3043	1	1	8416	1728	1728	0	? TAPERING ==>gill_color=BUFF
3	R123058	17706	451	16808	16161	1	16161	16161	16161	387	1	8416	8200	8216	8	FREE ==>veil_color=WHITE
4	R122314	1	2409	1	1	3	1	1	1	992	1	8416	7568	8200	0	ONE veil_color=WHITE ==>FREE
5	R1	20706	13521	20152	19998	772	19998	19998	19973	6492	1	8416	3376	5316	144	BRUISES ==>stalk_surf_above=SMOOTH
6	R56	1	61660	1	1	11220	1	1	5429	38396	1	8416	1712	8216	0	SOLITARY ==>veil_color=WHITE
7	R1595	1	1	1	1	20927	1	1	8311	1	1	8416	1296	1296	0	LARGE ==>BULBOUS stalk_surf_above=SILKY
8	R123059	22455	530	18520	16248	2	16248	16248	16211	388	1	8416	8216	8200	24	veil_color=WHITE ==>FREE
9	R122317	1	7217	1	1	13	1	1	2	1860	1	8416	6608	8216	0	CLOSE FREE ==>veil_color=WHITE
10	R2	87599	58974	52313	30653	1946	30653	30653	30636	20308	1	8416	3808	7768	600	NO ODOR ==>ONE
11	R86	1	8048	1	1	3434	1	1	996	7816	1	8416	2160	3928	0	FOUL ==>POISONOUS
12	R4743	1	1	1	1	8575	1	1	3043	1	1	8416	1728	1728	0	? TAPERING spore_print_color=WHITE ==>gill_color=BUFF
13	R107595	1	8972	1	1	18	1	1	3	2574	1	8416	6272	8216	0	CLOSE FREE ONE ==>veil_color=WHITE
14	R3	70095	72834	59241	50189	20002	50189	50189	50189	26056	1	8416	5076	4744	1872	stalk_surf_below=SMOOTH ==>stalk_color_above=WHITE
15	R117	1	65233	1	1	16136	1	1	6664	43177	1	8416	1500	8216	0	cap_color=RED ==>veil_color=WHITE
16	R4750	1	1	1	1	8575	1	1	3043	1	1	8416	1728	1728	0	? POISONOUS TAPERING ==>gill_color=BUFF
17	R122312	17707	2497	16809	16162	4	16162	16162	16162	996	1	8416	7576	8216	8	FREE ONE ==>veil_color=WHITE
18	R107597	1	8914	1	1	18	1	1	3	2564	1	8416	6272	8200	0	CLOSE ONE veil_color=WHITE ==>FREE
19	R6	50692	37924	46600	43420	7125	43420	43420	43415	11469	1	8416	4488	3968	1336	EDIBLE ==>PENDANT

Figure 6.9: Union of the ten most interesting rules for each measure of cluster $C1$ on the $\mathcal{R}_1$ ruleset (extracted).

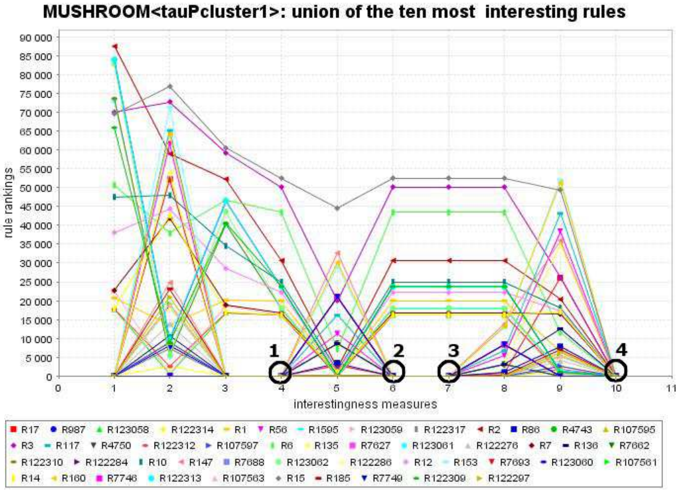
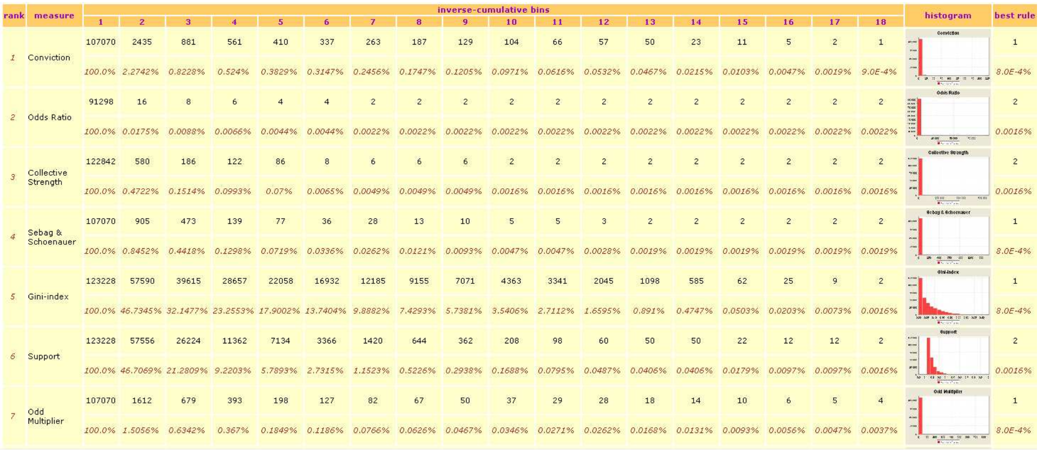
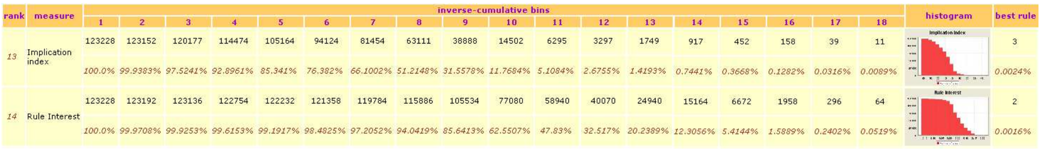


Figure 6.10: Plot of the union of the ten most interesting rules for each measure of cluster $C1$ on the $\mathcal{R}_1$ ruleset.



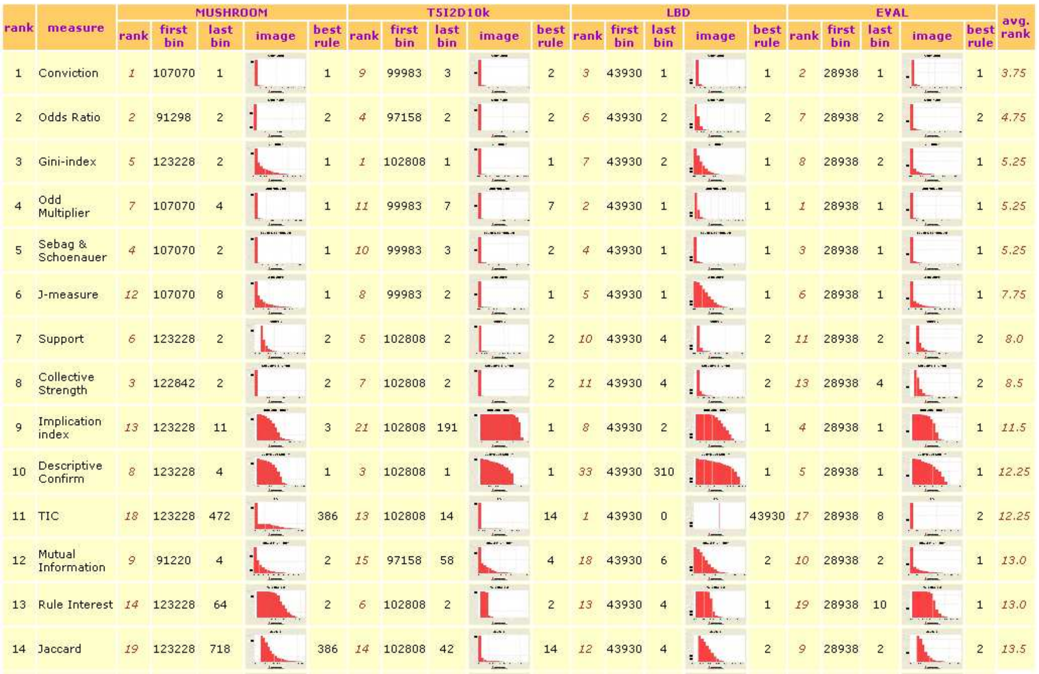
rank	measure	inverse-cumulative bins																		histogram	best rule
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18		
1	Conviction	107070	2435	881	561	410	337	263	187	129	104	66	57	50	23	11	5	2	1	Conviction	1
		100.0%	2.2742%	0.8228%	0.524%	0.3829%	0.3147%	0.2456%	0.1747%	0.1205%	0.0971%	0.0616%	0.0532%	0.0467%	0.0215%	0.0103%	0.0047%	0.0019%	9.0E-4%		8.0E-4%
2	Odds Ratio	91298	16	8	6	4	4	2	2	2	2	2	2	2	2	2	2	2	2	Odds Ratio	2
		100.0%	0.0175%	0.0088%	0.0066%	0.0044%	0.0044%	0.0022%	0.0022%	0.0022%	0.0022%	0.0022%	0.0022%	0.0022%	0.0022%	0.0022%	0.0022%	0.0022%	0.0022%		0.0016%
3	Collective Strength	122842	580	186	122	86	8	6	6	6	2	2	2	2	2	2	2	2	2	Collective Strength	2
		100.0%	0.4722%	0.1514%	0.0993%	0.07%	0.0065%	0.0049%	0.0049%	0.0049%	0.0016%	0.0016%	0.0016%	0.0016%	0.0016%	0.0016%	0.0016%	0.0016%	0.0016%		0.0016%
4	Sebag & Schoenauer	107070	905	473	139	77	36	28	13	10	5	5	3	2	2	2	2	2	2	Sebag & Schoenauer	1
		100.0%	0.8452%	0.4418%	0.1298%	0.0719%	0.0336%	0.0262%	0.0121%	0.0093%	0.0047%	0.0047%	0.0028%	0.0019%	0.0019%	0.0019%	0.0019%	0.0019%	0.0019%		8.0E-4%
5	Gini-index	123228	57590	39615	28657	22058	16932	12185	9155	7071	4363	3341	2045	1098	585	62	25	9	2	Gini-index	1
		100.0%	46.7345%	32.1477%	23.2553%	17.9002%	13.7404%	9.8882%	7.4293%	5.7381%	3.5406%	2.7112%	1.6595%	0.891%	0.4747%	0.0503%	0.0203%	0.0073%	0.0016%		8.0E-4%
6	Support	123228	57556	26224	11362	7134	3366	1420	644	362	208	98	60	50	50	22	12	12	2	Support	2
		100.0%	46.7069%	21.2809%	9.2203%	5.7893%	2.7315%	1.1523%	0.5226%	0.2938%	0.1688%	0.0795%	0.0487%	0.0406%	0.0406%	0.0179%	0.0097%	0.0097%	0.0016%		0.0016%
7	Odd Multiplier	107070	1612	679	393	198	127	82	67	50	37	29	28	18	14	10	6	5	4	Odd Multiplier	1
		100.0%	1.5056%	0.6342%	0.367%	0.1849%	0.1186%	0.0766%	0.0626%	0.0467%	0.0346%	0.0271%	0.0262%	0.0168%	0.0131%	0.0093%	0.0056%	0.0047%	0.0037%		8.0E-4%

Figure 6.11: Sensitivity rank on the $\mathcal{R}_1$ ruleset (extracted).



rank	measure	inverse-cumulative bins																		histogram	best rule
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18		
13	Implication index	123228	123152	120177	114474	105164	94124	81454	63111	38888	14502	6295	3297	1749	917	452	158	39	11	Implication index	3
		100.0%	99.9383%	97.5241%	92.8961%	85.341%	76.382%	66.1002%	51.2148%	31.5578%	11.7684%	5.1084%	2.6755%	1.4193%	0.7441%	0.3668%	0.1282%	0.0316%	0.0089%		0.0024%
14	Rule Interest	123228	123192	123136	122754	122232	121358	119784	115886	105534	77080	58940	40070	24940	15164	6672	1958	296	64	Rule Interest	2
		100.0%	99.9708%	99.9253%	99.6153%	99.1917%	98.4825%	97.2052%	94.0419%	85.6413%	62.5507%	47.83%	32.517%	20.2389%	12.3056%	5.4144%	1.5889%	0.2402%	0.0519%		0.0016%

Figure 6.12: Comparison of sensitivity values on a couple of measures of the $\mathcal{R}_1$ ruleset.



rank	measure	MUSHROOM					T5I2D10k					LBD					EVAL					avg. rank
		rank	first bin	last bin	image	best rule	rank	first bin	last bin	image	best rule	rank	first bin	last bin	image	best rule	rank	first bin	last bin	image	best rule	
1	Conviction	1	107070	1		1	9	99983	3		2	3	43930	1		1	2	28938	1		1	3.75
2	Odds Ratio	2	91298	2		2	4	97158	2		2	6	43930	2		2	7	28938	2		2	4.75
3	Gini-index	5	123228	2		1	1	102808	1		1	7	43930	2		1	8	28938	2		1	5.25
4	Odd Multiplier	7	107070	4		1	11	99983	7		7	2	43930	1		1	1	28938	1		1	5.25
5	Sebag & Schoenauer	4	107070	2		1	10	99983	3		2	4	43930	1		1	3	28938	1		1	5.25
6	J-measure	12	107070	8		1	8	99983	2		1	5	43930	1		1	6	28938	1		1	7.75
7	Support	6	123228	2		2	5	102808	2		2	10	43930	4		2	11	28938	2		2	8.0
8	Collective Strength	3	122842	2		2	7	102808	2		2	11	43930	4		2	13	28938	4		2	8.5
9	Implication index	13	123228	11		3	21	102808	191		1	8	43930	2		1	4	28938	1		1	11.5
10	Descriptive Confirm	8	123228	4		1	3	102808	1		1	33	43930	310		1	5	28938	1		1	12.25
11	TIC	18	123228	472		386	13	102808	14		14	1	43930	0		43930	17	28938	8		2	12.25
12	Mutual Information	9	91220	4		2	15	97158	58		4	18	43930	6		2	10	28938	2		1	13.0
13	Rule Interest	14	123228	64		2	6	102808	2		2	13	43930	4		1	19	28938	10		1	13.0
14	Jaccard	19	123228	718		386	14	102808	42		14	12	43930	4		2	9	28938	2		2	13.5

Figure 6.13: Sensitivity rank on all the original rulesets (extracted).

Chapter 7

Comparative studies

In the context of data mining, we use Pearson's, Spearman's and Kendall's correlation coefficients in order to compare the behavior of 40 objective measures of association rules. Via a the graph-based approach [Sec. 4.6.4, Chap. 4], we can visualize not only the strong but also the weak correlations between interestingness measures. We propose to discover the stable clusters of objective measures (i.e. subsets of objective measures delivering a close rule ranking) by making comparative study on two opposite datasets (a highly correlated one and a lowly correlated one) and two real-life datasets. The results – are performed with the techniques such as correlation graph, AHC and PAM [Sec. 4.6.4/4.6.2/4.6.3, Chap. 4] –, show that the correlation between objective measures depends on data nature and rule ranks, and also show some stable clusters of measures.

7.1 Graph of stable clusters

In order to facilitate the comparison between several correlation matrices, we have introduced some extensions to define the stable clusters between objective measures.

Definition 7.1.1 ($\overline{CG+}/\overline{CG0}$ graph)**.** The $\overline{CG+}$ graph (respectively $\overline{CG0}$ graph) of a set of k rulesets $\mathcal{R} = \{\mathcal{R}(\mathcal{D}_1), ..., \mathcal{R}(\mathcal{D}_k)\}$ is defined as the average graph of intersection of the k partially correlated (respectively uncorrelated) subgraphs $CG+_k$ (respectively $CG0_k$) calculated on $\mathcal{R}$. Hence, each edge of $\overline{CG+}$ (respectively $\overline{CG0}$) is associated with the average value of the corresponding edge in the k $CG+_k$ graphs. Therefore, the $\overline{CG+}$ (respectively $\overline{CG0}$) graph allows visualizing the strongly (respectively weakly) stable correlations, as being common to k studied rulesets.

Definition 7.1.2 (Stable cluster)**.** We call τ-stable (respectively θ-stable) cluster a connected part of the $\overline{CG+}$ (respectively $\overline{CG0}$) graph.

7.2 Comparative analysis: Pearson's coefficient

7.2.1 Discussion

The analysis aims at finding stable relations between the objective measures studied over the eight rulesets affecting by Pearson's coefficient [Sec. 5.4.2, Chap. 5]. We investigate in: (i) the $\overline{CG0}$ graphs in order to identify the objective measures that do not agree for ranking the rules, (ii) the $\overline{CG+}$ graph in order to find the objective measures that do agree for ranking the rules.

Ruleset	Number of correlations		Number of clusters	
	τ-correlated	θ-uncorrelated	CG+	CG0
$\mathcal{R}_1$	110	1	10	39
$\mathcal{R}_2$	75	0	14	40
$\mathcal{R}_3$	99	11	10	29
$\mathcal{R}_4$	108	4	9	36
$\mathcal{R}_1^{'}$	114	8	11	32
$\mathcal{R}_2^{'}$	82	12	14	29
$\mathcal{R}_3^{'}$	103	17	7	23
$\mathcal{R}_4^{'}$	105	32	8	20

Table 7.1: Comparison of correlations with Pearson's coefficient.

7.2.2 $CG+$ and $CG0$

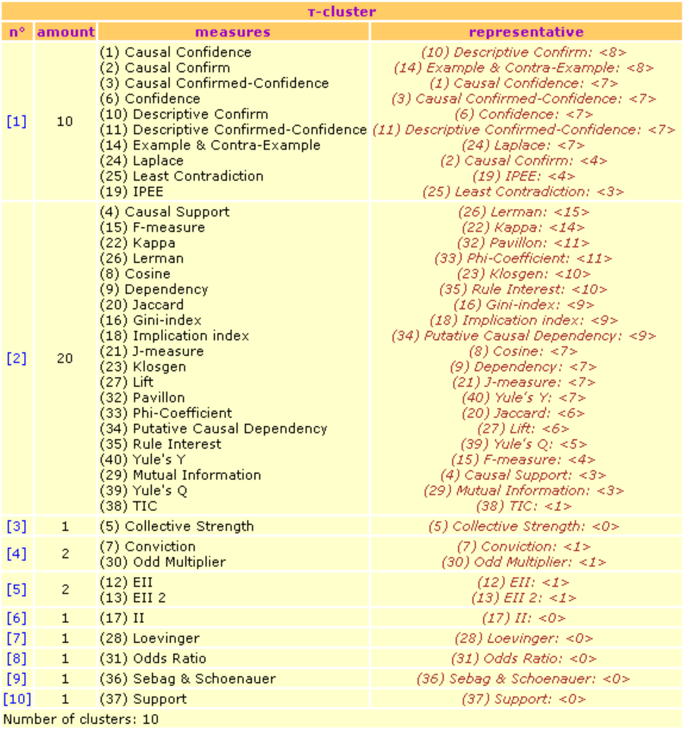
Fig. 7.2 and Fig. 7.3 show the $CG+$ graphs obtained from the eight corresponding rulesets. As seen before, the sample rulesets and the original rulesets have close results so we can use the sample rulesets ($\mathcal{R}_1^{'}$, $\mathcal{R}_2^{'}$, $\mathcal{R}_3^{'}$ and $\mathcal{R}_4^{'}$) for representing the original rulesets. This observation is useful when we evaluate the CG+ graphs but not for CG0 graphs (Fig. 7.4 and Fig. 7.5).

As in [Sec. 4.6, Chap. 4] for the representative measure within a cluster, the user can choose a cluster to examine. For example, with the CG+ graph of $\mathcal{R}_1$ (Fig. 7.2), one can choose the second cluster as the largest one containing twenty measures for his/her first choice. In this cluster one can obtain Lerman as the representative measure by choosing the measure has a strong relation with the others (column 4 in Fig. 7.1). One can also see the weak relation between TIC and the other measures of the cluster with the number of relations is only 1. Generally, one can use the 10 representative measures obtained from Fig. 7.1 to replace the 40 studied measures (25%).

Tab. 7.1 also shows a quite stable tendency in counting the number of τ-correlated: 110($\mathcal{R}_1$) → 114($\mathcal{R}_1^{'}$), 75($\mathcal{R}_2$) → 82($\mathcal{R}_2^{'}$), 99($\mathcal{R}_3$) → 103($\mathcal{R}_3^{'}$), 108($\mathcal{R}_4$) → 105($\mathcal{R}_4^{'}$) but a quite different with θ-uncorrelated: 1($\mathcal{R}_1$) → 8($\mathcal{R}_1^{'}$), 0($\mathcal{R}_2$) → 12($\mathcal{R}_2^{'}$), 11($\mathcal{R}_3$) → 17($\mathcal{R}_3^{'}$), 4($\mathcal{R}_4$) → 32($\mathcal{R}_4^{'}$).

7.2.3 $\overline{CG0}$ graphs: uncorrelated stability

CG0 graphs first show that there are no θ-stable clusters appearing on the eight rulesets studied in Fig. 7.4 and Fig. 7.5. Secondly, there is no $\overline{CG0}$ graph from these rulesets.



τ-cluster			
n°	amount	measures	representative
[1]	10	(1) Causal Confidence (2) Causal Confirm (3) Causal Confirmed-Confidence (6) Confidence (10) Descriptive Confirm (11) Descriptive Confirmed-Confidence (14) Example & Contra-Example (24) Laplace (25) Least Contradiction (19) IPEE	*(10) Descriptive Confirm: <8>* *(14) Example & Contra-Example: <8>* *(1) Causal Confidence: <7>* *(3) Causal Confirmed-Confidence: <7>* *(6) Confidence: <7>* *(11) Descriptive Confirmed-Confidence: <7>* *(24) Laplace: <7>* *(2) Causal Confirm: <4>* *(19) IPEE: <4>* *(25) Least Contradiction: <3>*
[2]	20	(4) Causal Support (15) F-measure (22) Kappa (26) Lerman (8) Cosine (9) Dependency (20) Jaccard (16) Gini-index (18) Implication index (21) J-measure (23) Klosgen (27) Lift (32) Pavillon (33) Phi-Coefficient (34) Putative Causal Dependency (35) Rule Interest (40) Yule's Y (29) Mutual Information (39) Yule's Q (38) TIC	*(26) Lerman: <15>* *(22) Kappa: <14>* *(32) Pavillon: <11>* *(33) Phi-Coefficient: <11>* *(23) Klosgen: <10>* *(35) Rule Interest: <10>* *(16) Gini-index: <9>* *(18) Implication index: <9>* *(34) Putative Causal Dependency: <9>* *(8) Cosine: <7>* *(9) Dependency: <7>* *(21) J-measure: <7>* *(40) Yule's Y: <7>* *(20) Jaccard: <6>* *(27) Lift: <6>* *(39) Yule's Q: <5>* *(15) F-measure: <4>* *(4) Causal Support: <3>* *(29) Mutual Information: <3>* *(38) TIC: <1>*
[3]	1	(5) Collective Strength	*(5) Collective Strength: <0>*
[4]	2	(7) Conviction (30) Odd Multiplier	*(7) Conviction: <1>* *(30) Odd Multiplier: <1>*
[5]	2	(12) EII (13) EII 2	*(12) EII: <1>* *(13) EII 2: <1>*
[6]	1	(17) II	*(17) II: <0>*
[7]	1	(28) Loevinger	*(28) Loevinger: <0>*
[8]	1	(31) Odds Ratio	*(31) Odds Ratio: <0>*
[9]	1	(36) Sebag & Schoenauer	*(36) Sebag & Schoenauer: <0>*
[10]	1	(37) Support	*(37) Support: <0>*
Number of clusters: 10			

Figure 7.1: τ-cluster on $\mathcal{R}_1$ with Pearson's coefficient.

7.2.4 $\overline{CG+}$ graph: correlated stability

From Tab. 7.1, we can see that, $\mathcal{R}_1$ ($\mathcal{R}'_1$) is approximately 1.5 times as correlated as $\mathcal{R}_2$ ($\mathcal{R}'2$). $\mathcal{R}_3$ ($\mathcal{R}'_3$) is approximately as correlated as $\mathcal{R}_4$ ($\mathcal{R}'_4$). As seen in Fig. 7.6, seven τ-stable clusters found come from the rulesets studied.

By briefly analyzing these τ-stable clusters, some interesting observations are drawn.

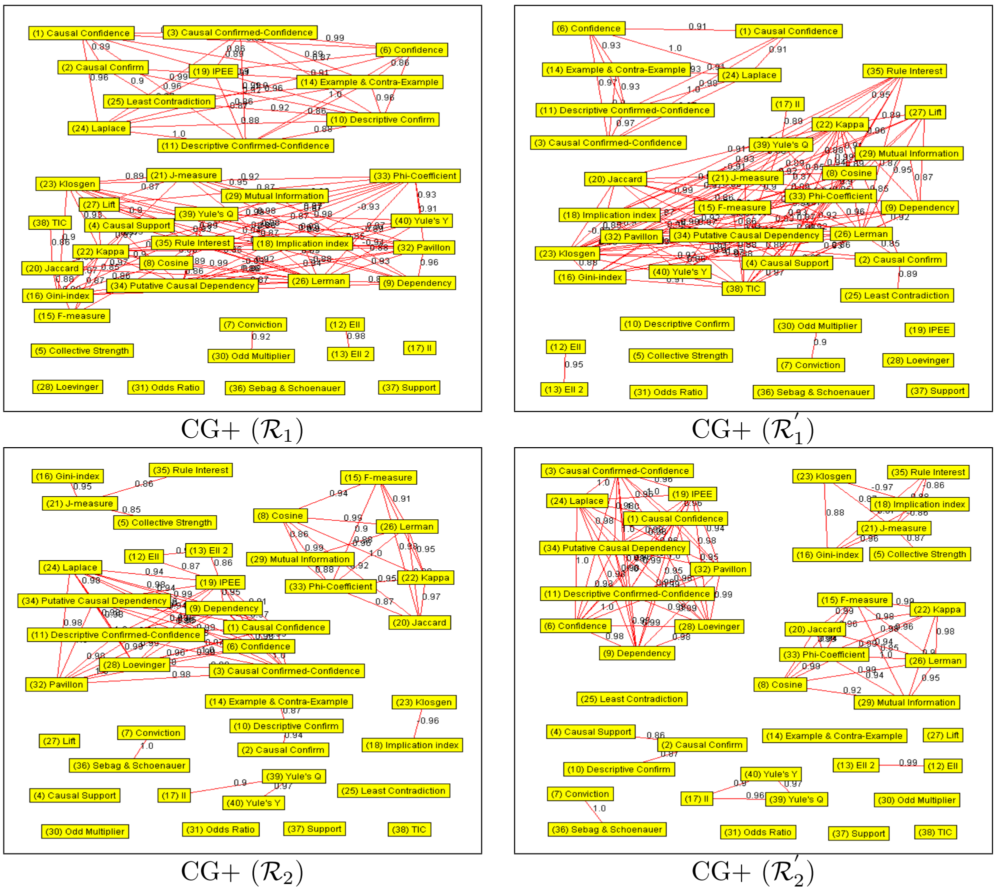
(C1), the largest cluster, (Confidence, Causal Confidence, Causal Confirmed-Confidence, Descriptive Confirmed-Confidence, Laplace) has most of its measures extended from Confidence measure. From this cluster, we can easily see a highly connected component – each vertex must have an edge with the other vertices – indicating the strong agreement of the five measures. According to the classification in Tab. 3.3 [Chap. 3], this cluster is associated with the descriptive measures that are sensible to equilibrium.

(C2)(C5)(C6), another three clusters, have themselves each a highly connected components which are formed by (Cosine, Jaccard, F-measure), (Phi-Coefficient, Lerman, Kappa) and (Implication Index, Klosgen). Most of these measures are similarity measures. These three clusters are useful to measure the deviation from independence (see the classification in Tab. 3.3 [Chap. 3]).

(C3), this cluster (Dependency, Pavillon, Putative Causal Dependency) is interesting because almost all the measures of this cluster are reasonably well correlated.

(C4), is a statistical cluster formed by EII and EII 2, which are two measures obtained with different parameters of the same original formula.

(C7), Yule's Q and Yule's Y, brings out a trivial observation because these measures are derived from Odds Ratio measure. Both measures are in the group of measuring the deviation from



CG+ ($\mathcal{R}_1$)

CG+ ($\mathcal{R}_1'$)

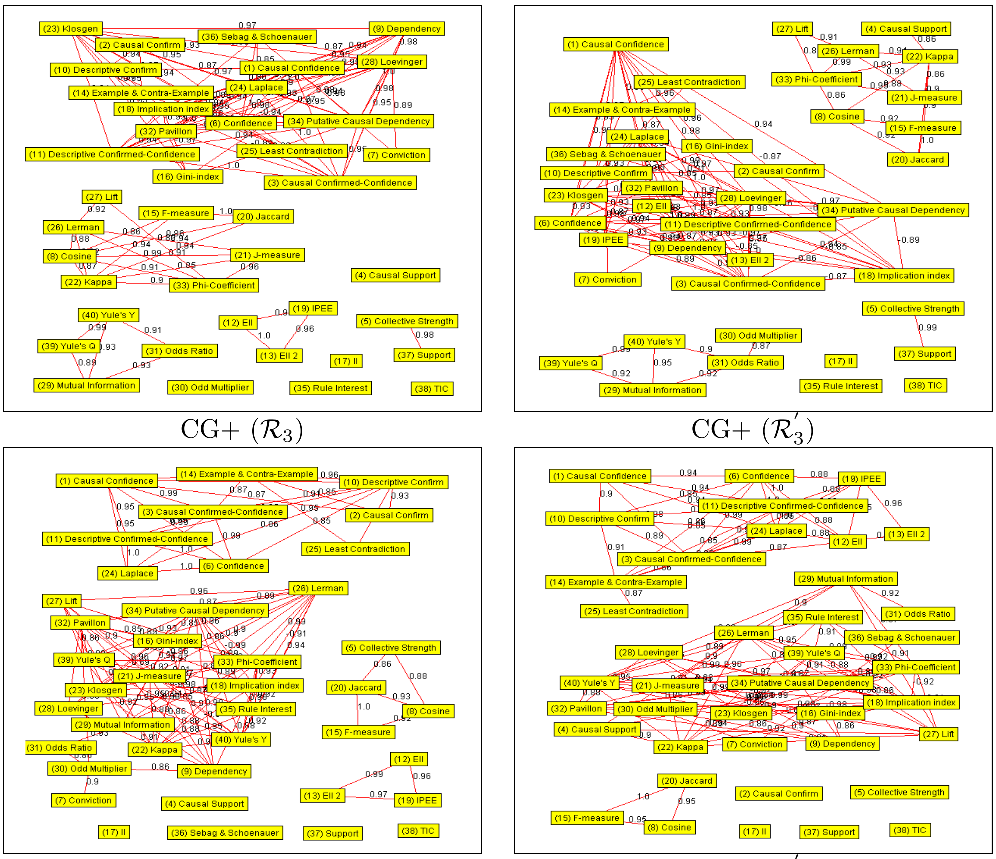
CG+ ($\mathcal{R}_2$)

CG+ ($\mathcal{R}_2'$)

Figure 7.2: CG+ graphs with Pearson's coefficient.

independence (see the classification in Tab. 3.3 [Chap. 3]).

In looking for τ-stable clusters, we have found the τ-correlated that exist between various measures. Seven τ-stable clusters have been identified. Each τ-stable cluster, that forms a subgraph in a $\overline{CG+}$ graph, also contains a highly connected component. Therefore, we can choose a representative measure for each one of these clusters. For example with our experiment, we have chosen seven representative measures from the 40 objective measures studied. How we can choose a representative measure is also an interesting study for the future. In the first approach, the user can choose the measure that is the best one from his/her point of view. The second approach is that we can select the measure that has the highest number of relations with the others (e.g., the measures Descriptive Confirmed-Confidence, Jaccard, Lerman, Klosgen, Pavillon, EII($\alpha = 1$), and Yule's Q from Fig. 7.6). The stronger the τ-stable cluster, the more interesting the representative measure. An important observation is that, the existence of highly connected graphs represents a strong agreement with a τ-stable cluster. We have reached significant information: *τ-stable clusters can be obtained from different measures and rulesets.* The different measures imply taking into account both their mathematical definitions and their respective properties.



CG+ ($\mathcal{R}_3$)

CG+ ($\mathcal{R}_3'$)

CG+ ($\mathcal{R}_4$)

CG+ ($\mathcal{R}_4'$)

Figure 7.3: CG+ graphs with Pearson's coefficient (cont.).

7.3 Comparative analysis: Spearman's coefficient

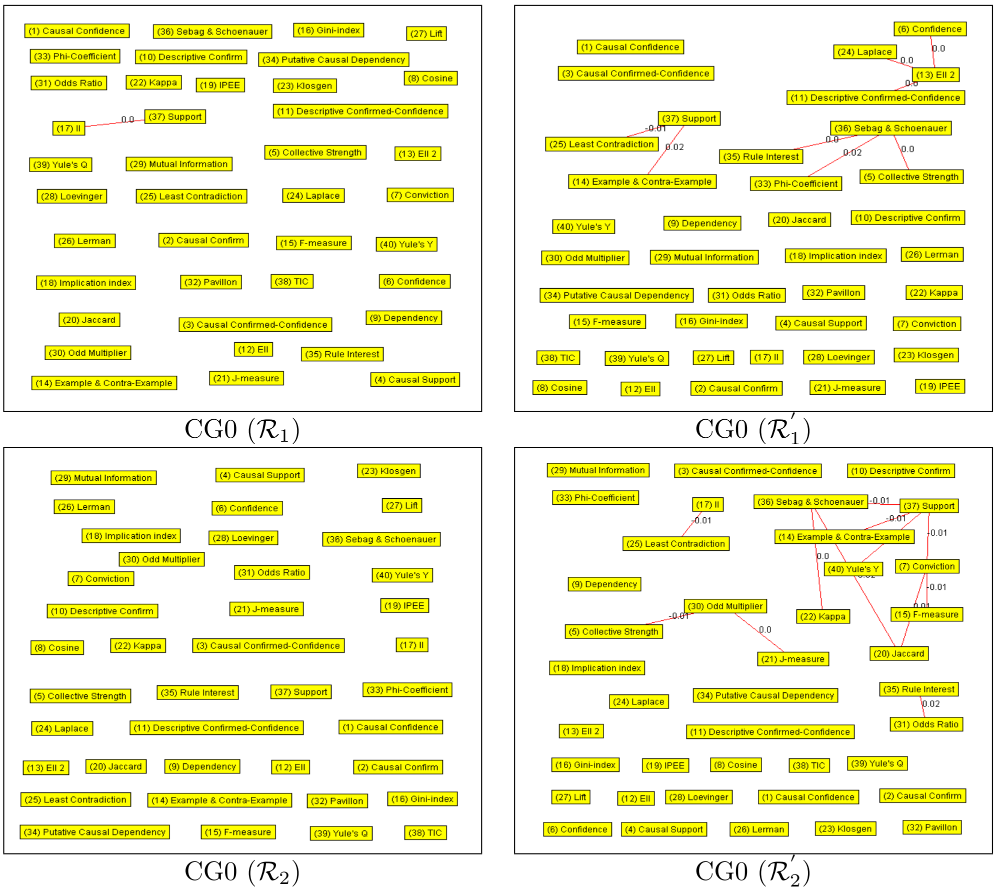
To resolve the outlier problem on Pearson's coefficient, an analysis with Spearman's coefficient is performed.

7.3.1 Discussion

As in the precedent discussion, this analysis also aims at finding stable relations between the objective measures studied over the eight rulesets with Spearman's coefficient (see Sec. 5.4.2 [Chap. 5]). The same aspects are investigated in: (i) the $\overline{CG0}$ graphs, and (ii) the $\overline{CG+}$ graph.

7.3.2 $CG+$ and $CG0$

Fig. 7.8 and Fig. 7.9 show the $CG+$ graphs obtained from the eight corresponding rulesets. As the same precedent observation, we can use the sample rulesets ($\mathcal{R}_1'$, $\mathcal{R}_2'$, $\mathcal{R}_3'$ and $\mathcal{R}_4'$) for representing the original rulesets. This observation is also helpful for the CG+ graphs but not for the CG0 graphs



CG0 ($\mathcal{R}_1$)

CG0 ($\mathcal{R}_1^{'}$)

CG0 ($\mathcal{R}_2$)

CG0 ($\mathcal{R}_2^{'}$)

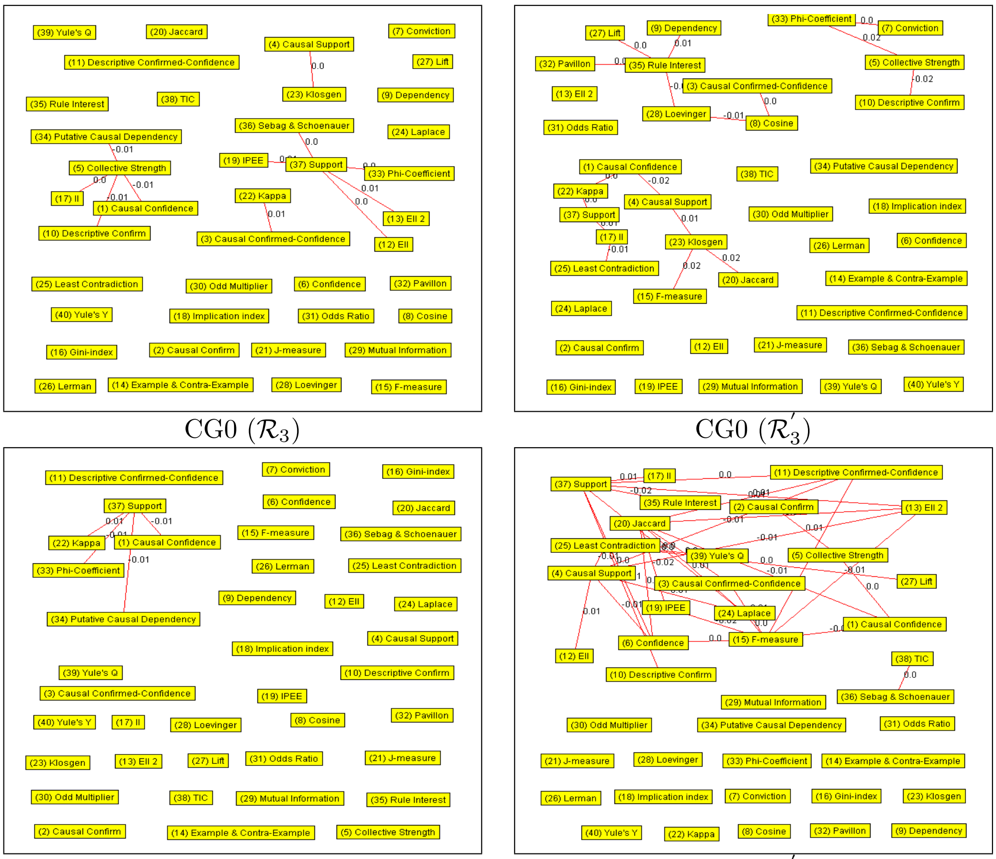
Figure 7.4: CG0 graphs with Pearson's coefficient.

(Fig. 7.10 and Fig. 7.11).

As in Sec. 4.6 [Chap. 4] for the representative measure within a cluster, the user can choose a cluster to examine. For example, with the CG+ graph of $\mathcal{R}_1$ (Fig. 7.2), one can choose the first cluster as the largest one containing twenty-four measures for his/her first choice. In this cluster one can obtain Putative Causal Dependency as the representative measure by choosing the measure has a strong relation with the others (column 4 in Fig. 7.7). One can also see the weak relation between TIC and the other measures of the second cluster in the precedent analysis with Pearson's coefficient is now broken and TIC is currently an independent cluster. The same number of the representative measures obtained from Fig. 7.7 to replace the 40 studied measures (25%) as in the discussion with Pearson's coefficient.

Tab. 7.2 also shows a quite stable tendency in counting the number of τ-correlated: $113(\mathcal{R}_1) \rightarrow 113(\mathcal{R}_1^{'})$, $95(\mathcal{R}_2) \rightarrow 90(\mathcal{R}_2^{'})$, $182(\mathcal{R}_3) \rightarrow 171(\mathcal{R}_3^{'})$, $208(\mathcal{R}_4) \rightarrow 189(\mathcal{R}_4^{'})$ but a quite different with θ-uncorrelated: $0(\mathcal{R}_1) \rightarrow 14(\mathcal{R}_1^{'})$, $3(\mathcal{R}_2) \rightarrow 13(\mathcal{R}_2^{'})$, $4(\mathcal{R}_3) \rightarrow 36(\mathcal{R}_3^{'})$, $8(\mathcal{R}_4) \rightarrow 32(\mathcal{R}_4^{'})$.

With Spearman's coefficient, the number of τ-correlated in the real-life rulesets $\mathcal{R}_3, \mathcal{R}_4$ $(\mathcal{R}'_3, \mathcal{R}'_4)$ is twice as correlated as in the first two rulesets $\mathcal{R}_1, \mathcal{R}_2$ $(\mathcal{R}'_1, \mathcal{R}'_2)$.



CG0 ($\mathcal{R}_3$)

CG0 ($\mathcal{R}_3^{'}$)

CG0 ($\mathcal{R}_4$)

CG0 ($\mathcal{R}_4^{'}$)

Figure 7.5: CG0 graphs with Pearson's coefficient (cont.).

7.3.3 $\overline{CG0}$ graphs: uncorrelated stability

As the same observation with Pearson's coefficient from the precedent section. There are no θ-stable clusters that appear on CG0 graphs obtained from the eight rulesets studied (Fig. 7.4 and Fig. 7.5). There is no $\overline{CG0}$ graph from these rulesets.

7.3.4 $\overline{CG+}$ graph: correlated stability

From Tab. 7.2, we can see that $\mathcal{R}_1$ ($\mathcal{R}'_1$) is approximately 1.2 times as correlated as $\mathcal{R}_2$ ($\mathcal{R}'2$). $\mathcal{R}_3$ ($\mathcal{R}'_3$) is approximately 0.9 times as correlated as $\mathcal{R}_4$ ($\mathcal{R}'_4$). As seen in Fig. 7.12, six τ-stable clusters found come from the rulesets studied.

By comparing the $\overline{CG+}$ graph obtained with which is established by Pearson's coefficient in the precedent section, some remarks can be drawn.

- In general, there are not much different between the two graphs.

- With clusters C1(with Pearson's coefficient)-C1(with Spearman's coefficient) (C6-C4, C7-C5 respectively), we can see the participation of Example & Contra-Example (Lift, Odds Ratio

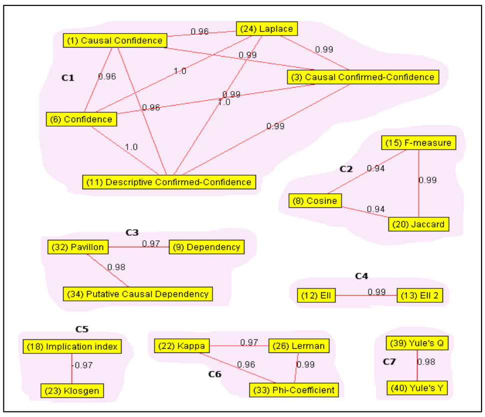


Figure 7.6: $\overline{CG+}$ graph with Pearson's coefficient.

Ruleset	Number of correlations		Number of clusters	
	τ-correlated	θ-uncorrelated	CG+	CG0
$\mathcal{R}_1$	113	0	10	40
$\mathcal{R}_2$	95	3	10	37
$\mathcal{R}_3$	182	4	7	36
$\mathcal{R}_4$	208	8	8	32
$\mathcal{R}_1^{'}$	113	14	12	26
$\mathcal{R}_2^{'}$	90	13	10	30
$\mathcal{R}_3^{'}$	171	36	9	17
$\mathcal{R}_4^{'}$	189	32	7	24

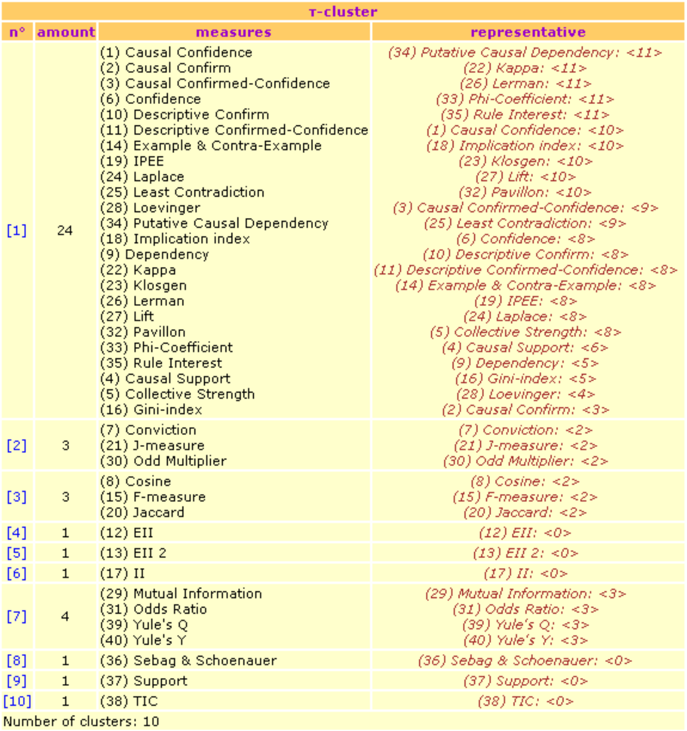
Table 7.2: Comparison of correlations with Spearman's coefficient.

respectively) in cluster C1 formed with Spearman's coefficient. Both of these two clusters have the strong agreement between measures.

- Inversely observation in clusters C2-C2 (C3-C6 respectively) with the disappearance of Cosine (Dependency respectively) measure.
- The couple clusters C3-C5 are identical.
- The disappearance of the cluster C4 with EII and EII 2.

7.4 Comparative analysis: Kendall's coefficient

Another important coefficient, namely Kendall's coefficient, is integrated in our analysis. Despite the close results between Pearson's and Spearman's coefficients, the results obtained from Kendall's coefficient are not the same.



τ-cluster			
n°	amount	measures	representative
[1]	24	(1) Causal Confidence (2) Causal Confirm (3) Causal Confirmed-Confidence (6) Confidence (10) Descriptive Confirm (11) Descriptive Confirmed-Confidence (14) Example & Contra-Example (19) IPEE (24) Laplace (25) Least Contradiction (28) Loevinger (34) Putative Causal Dependency (18) Implication index (9) Dependency (22) Kappa (23) Klosgen (26) Lerman (27) Lift (32) Pavillon (33) Phi-Coefficient (35) Rule Interest (4) Causal Support (5) Collective Strength (16) Gini-index	*(34) Putative Causal Dependency: <11>* *(22) Kappa: <11>* *(26) Lerman: <11>* *(33) Phi-Coefficient: <11>* *(35) Rule Interest: <11>* *(1) Causal Confidence: <10>* *(18) Implication index: <10>* *(23) Klosgen: <10>* *(27) Lift: <10>* *(32) Pavillon: <10>* *(3) Causal Confirmed-Confidence: <9>* *(25) Least Contradiction: <9>* *(6) Confidence: <8>* *(10) Descriptive Confirm: <8>* *(11) Descriptive Confirmed-Confidence: <8>* *(14) Example & Contra-Example: <8>* *(19) IPEE: <8>* *(24) Laplace: <8>* *(5) Collective Strength: <8>* *(4) Causal Support: <6>* *(9) Dependency: <5>* *(16) Gini-index: <5>* *(28) Loevinger: <4>* *(2) Causal Confirm: <3>*
[2]	3	(7) Conviction (21) J-measure (30) Odd Multiplier	*(7) Conviction: <2>* *(21) J-measure: <2>* *(30) Odd Multiplier: <2>*
[3]	3	(8) Cosine (15) F-measure (20) Jaccard	*(8) Cosine: <2>* *(15) F-measure: <2>* *(20) Jaccard: <2>*
[4]	1	(12) EII	*(12) EII: <0>*
[5]	1	(13) EII 2	*(13) EII 2: <0>*
[6]	1	(17) II	*(17) II: <0>*
[7]	4	(29) Mutual Information (31) Odds Ratio (39) Yule's Q (40) Yule's Y	*(29) Mutual Information: <3>* *(31) Odds Ratio: <3>* *(39) Yule's Q: <3>* *(40) Yule's Y: <3>*
[8]	1	(36) Sebag & Schoenauer	*(36) Sebag & Schoenauer: <0>*
[9]	1	(37) Support	*(37) Support: <0>*
[10]	1	(38) TIC	*(38) TIC: <0>*
Number of clusters: 10			

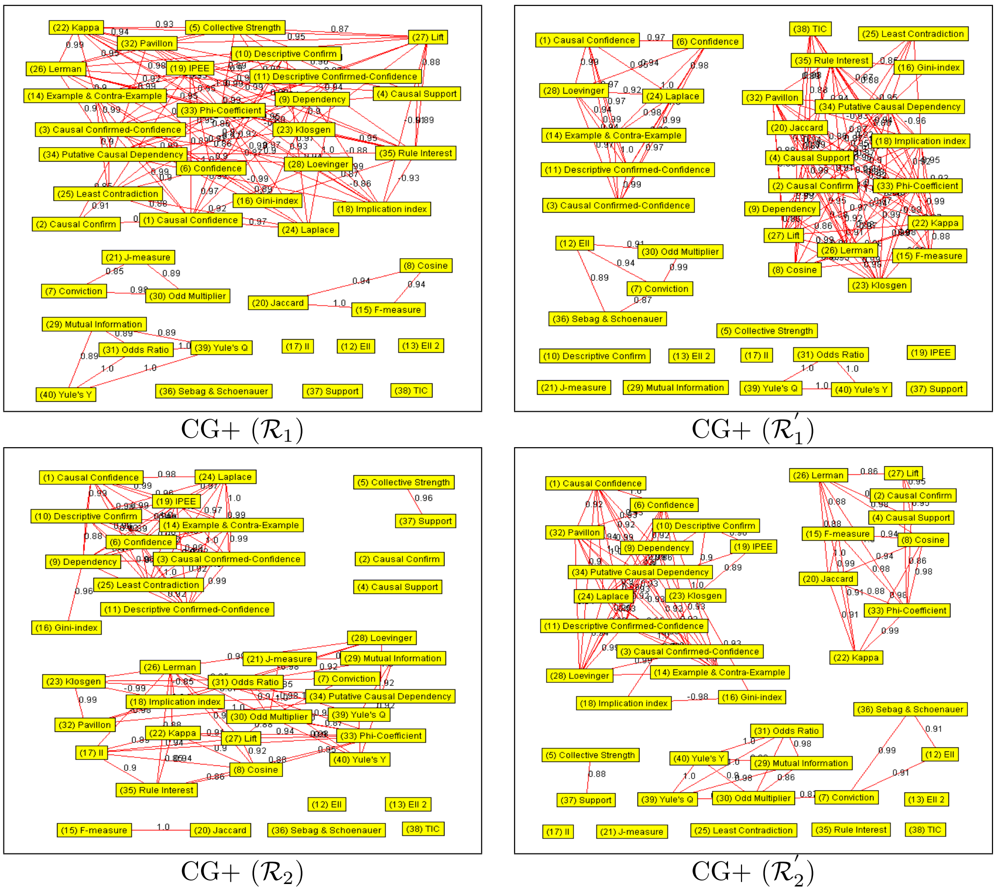
Figure 7.7: τ-cluster on $\mathcal{R}_1$ with Spearman's coefficient.

7.4.1 Discussion

In comparing Tab. 7.2 and Tab. 7.3, both of these two tables come from the two commonly rank coefficients Spearman and Kendall respecively, we can see a strong decrease of the number of τ-correlated computing with Kendall's coefficients over all the original rulesets. For instance, $\mathcal{R}_1 - S(113) \rightarrow \mathcal{R}_1 - K(37)$ (33%), $\mathcal{R}_2 - S(95) \rightarrow \mathcal{R}_2 - K(56)$ (59%), $\mathcal{R}_3 - S(182) \rightarrow \mathcal{R}_3 - K(88)$ (48%), $\mathcal{R}_4 - S(208) \rightarrow \mathcal{R}_4 - K(106)$ (51%) with an average percent of decrease of $\approx$ 48%!. The decrease of number of τ-correlated from sample rulesets also gives a similar change: $\mathcal{R}'_1 - S(113) \rightarrow \mathcal{R}'_1 - K(27)$ (24%), $\mathcal{R}'_2 - S(90) \rightarrow \mathcal{R}'_2 - K(35)$ (39%), $\mathcal{R}'_3 - S(171) \rightarrow \mathcal{R}'_3 - K(80)$ (47%), $\mathcal{R}'_4 - S(189) \rightarrow \mathcal{R}'_1 - K(61)$ (32%). The average percent of number of τ-correlated in the sample rulesets is smaller than in the original rulesets, varies from 48% to 35.5%.

This observation expresses the rules in the sample rulesets have a strong attachment in ranking relatively with each other after the sample process. Both of the two average decreases 48% and 35.5% also says that the relative increase or decrease of ranks (or interestingness values) between the objective measures over all the rulesets often lays down approximately 50%.

For θ-uncorelated, there are not much changes in the first two original rulesets but a strong change happens with the last two original rulesets: $\mathcal{R}_3 - S(4) \rightarrow \mathcal{R}_3 - K(46)$ (115%), $\mathcal{R}_4 - S(8) \rightarrow \mathcal{R}_3 - K(47)$ (59%). The same observations with the two sample rulesets. This is conformable with the above observation when the number of τ-correlated decrease then the number of θ-uncorrelated increase.



CG+ ($\mathcal{R}_1$)

CG+ ($\mathcal{R}_1'$)

CG+ ($\mathcal{R}_2$)

CG+ ($\mathcal{R}_2'$)

Figure 7.8: CG+ graphs with Spearman's coefficient.

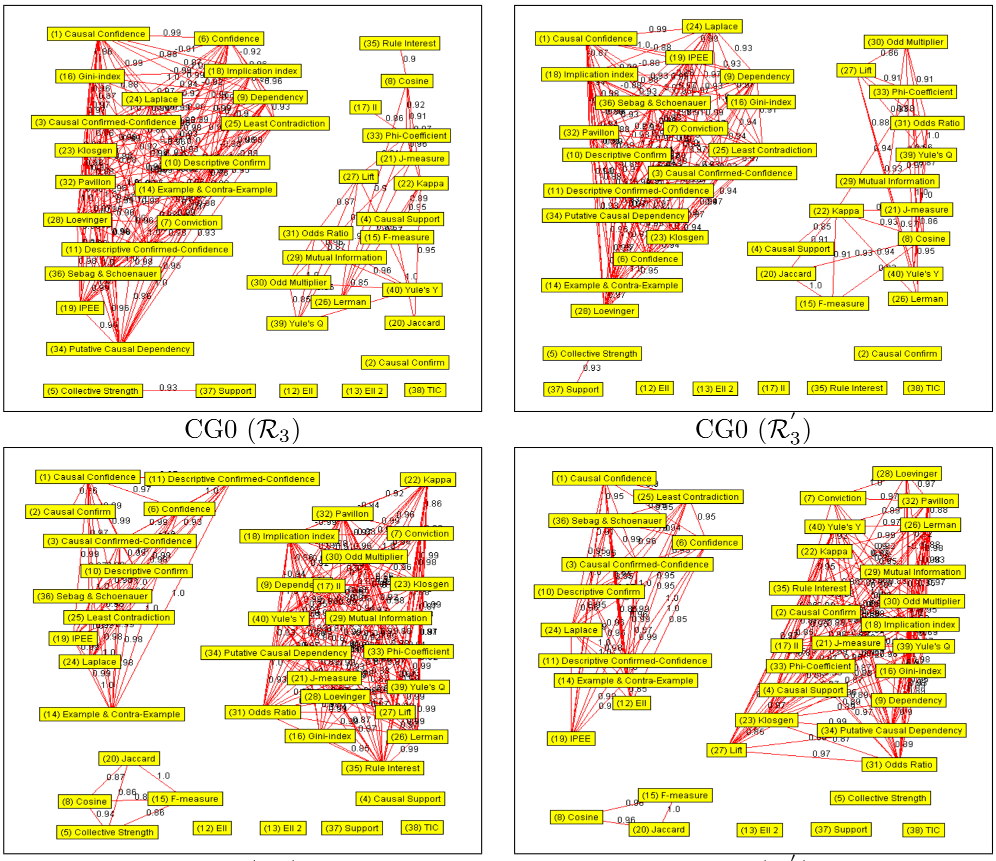
7.4.2 $CG+$ and $CG0$

Fig. 7.13 show the representative measures correspond to each cluster found with Kendall's coefficients in the $\mathcal{R}_1$ ruleset. Fig. 7.14 and Fig. 7.15 shows eight CG+ graphs obtained with Kendall's coefficients. As discussed before, there are fewer τ-correlated between the 40 objective measures.

We can see eight CG0 graphs on Fig. 7.16 and Fig. 7.17. The number of θ-uncorrelated arises strongly on the real-life rulesets ($\mathcal{R}_3$, $\mathcal{R}'_3$, $\mathcal{R}_4$, $\mathcal{R}'_4$) around the measure TIC as a center.

7.4.3 $\overline{CG0}$ graphs: uncorrelated stability

Although there are a lot of θ-uncorrelated in the last four CG0 graphs (Fig. 7.17) but the $\overline{CG0}$ graphs over the eight CG0 graphs on $\mathcal{R}_1$, $\mathcal{R}'_1$, $\mathcal{R}_2$, $\mathcal{R}'_2$, $\mathcal{R}_3$, $\mathcal{R}'_3$, $\mathcal{R}_4$, $\mathcal{R}'_4$ are empty.



CG0 ($\mathcal{R}_3$)

CG0 ($\mathcal{R}_3^{'}$)

CG0 ($\mathcal{R}_4$)

CG0 ($\mathcal{R}_4^{'}$)

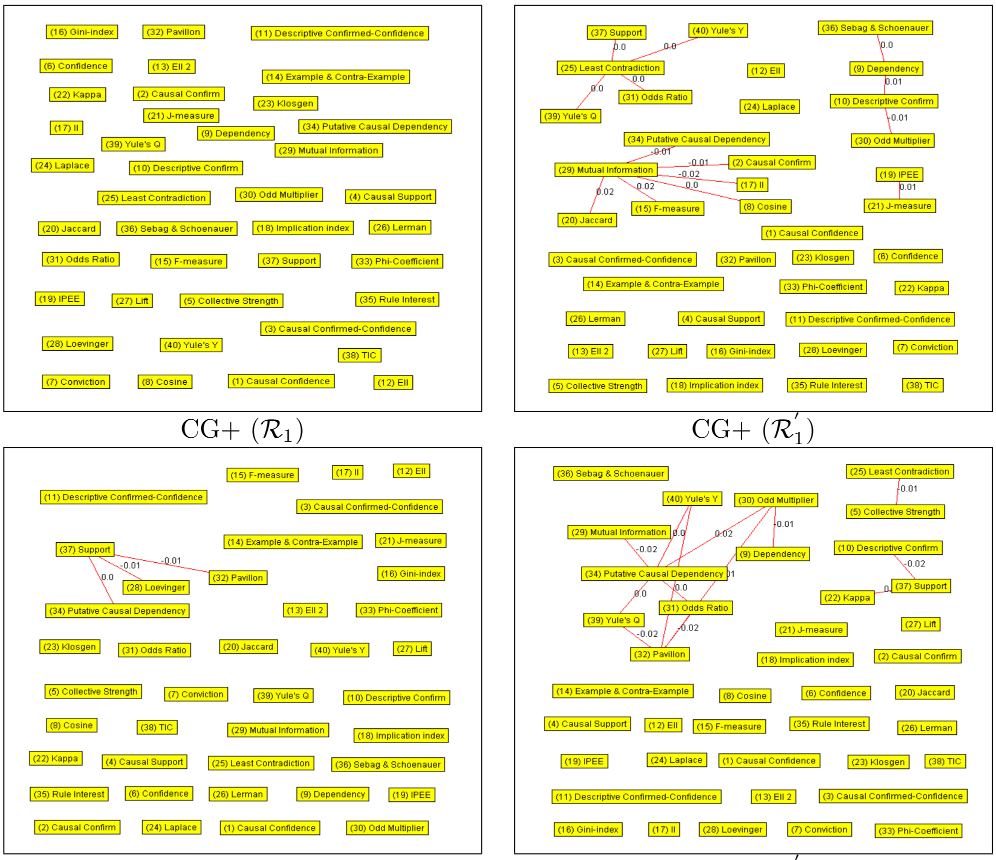
Figure 7.9: CG+ graphs with Spearman's coefficient (cont.).

7.4.4 $\overline{CG+}$ graph: correlated stability

With five τ-stable clusters, the $\overline{CG+}$ graph is the which one has the fewest of τ-stable clusters (see Fig. 7.18) in comparison with the other $\overline{CG+}$ graphs obtained with Pearson's and Spearman's coefficients (see Fig. 7.6 and Fig. 7.12). The number of τ-correlated (i.e., that is stable) decreases from 20 (in $\overline{CG+}$ graph with Pearson's coefficient) and 19 (in $\overline{CG+}$ graph with Spearman's coefficient) to 12 in the $\overline{CG+}$ graph with Kendall's coefficient. The clusters such as (EII, EII 2), (Yule's Q, Yule's Y) in the other two $\overline{CG+}$ graphs with Pearson's and Spearman's coefficients are disappeared. In Fig. 7.18 we can easily see the "Confidence" cluster from the last two $\overline{CG+}$ graphs is now divided into two "Confidence" clusters (Causal Confidence, Causal Confirmed-Confidence) and (Laplace, Confidence, Descriptive Confirmed-Confidence, Example & Contra-Example).

7.5 Comparative analysis: all coefficients

For each complement ruleset (*couple* or *multiple* rulesets) we also conduct the experimental results on $\overline{CG+}$ graphs with PS, PK, SK and PSK summaries (Sec. 5.4.2 [Chap. 5]).



CG+ ($\mathcal{R}_1$) CG+ ($\mathcal{R}_1'$)

CG+ ($\mathcal{R}_2$) CG+ ($\mathcal{R}_2'$)

Figure 7.10: CG0 graphs with Spearman's coefficient.

7.5.1 Discussion

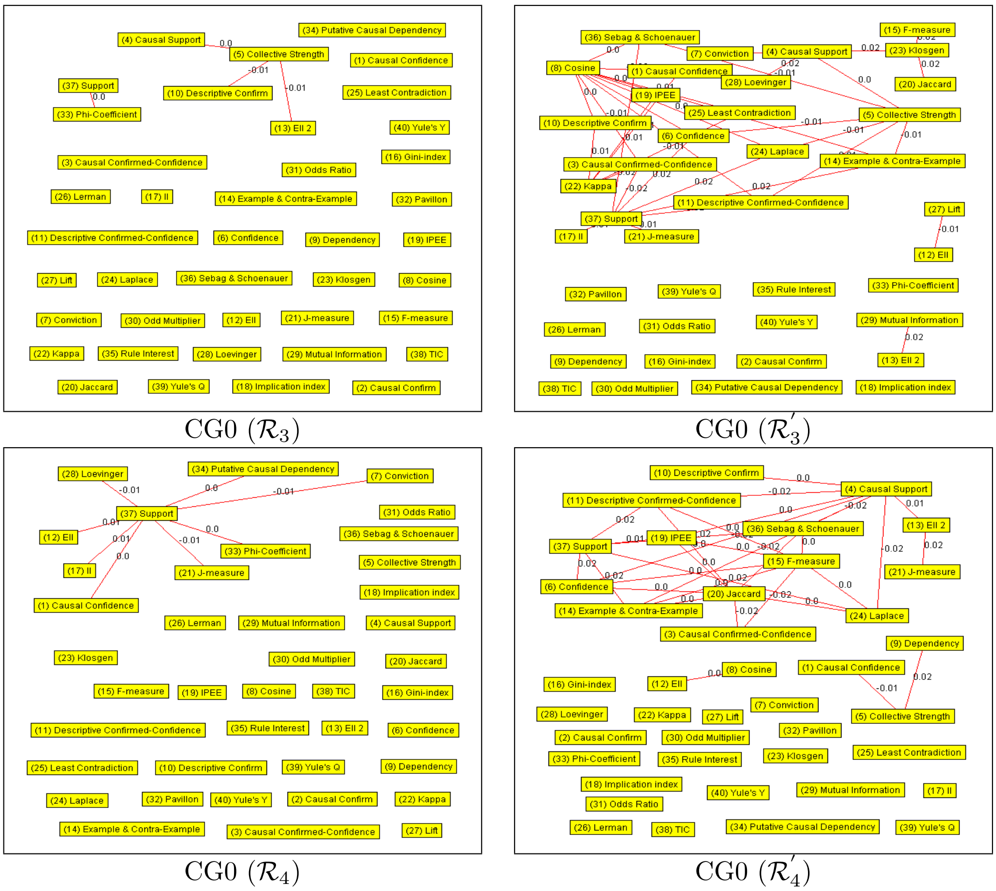
In this section we will give an example on the $\overline{CG+}$ graphs obtained from $\mathcal{R}_1$ on a summary combination S, P, K, PS, PK, SK, PSK. A first view on $\mathcal{R}_1$ is illustrated in Tab. 7.4;

7.5.2 $\overline{CG0}$ graphs: uncorrelated stability

As seen from the previous three sections, the $\overline{CG0}$ graph is often empty. It is also the same for this part.

7.5.3 $\overline{CG+}$ graph: correlated stability

Fig. 7.19 shows four $\overline{CG+}$ obtained from the $\mathcal{R}_1$ ruleset with PS, PK, SK and PSK respectively. The firs three CG+ graphs with P, S, and K are illustrated before in the three studied sections with Pearson's, Spearman's and Kendall's coefficients. In (PS) we can see a strong agreement between Pearson's and Spearman's coefficients in the $\overline{CG+}$ graph. The last three $\overline{CG+}$ graphs are closer with the results on Kendall's coefficients (on PK, SK and PSK).



CG0 ($\mathcal{R}_3$)

CG0 ($\mathcal{R}_3^{'}$)

CG0 ($\mathcal{R}_4$)

CG0 ($\mathcal{R}_4^{'}$)

Figure 7.11: CG0 graphs with Spearman's coefficient (cont.).

7.6 Comparative analysis: with AHC and PAM

Our aim is to discover the behaviors of the measures via two views the *strong relation* and the *relative distance* between measures when they are applied to the dissimilarity matrixes obtained from the rulesets studied. We use the two techniques AHC and PAM for each of these views respectively. The dissimilarity between each pair of measures is calculated from Pearson's coefficient.

7.6.1 On a ruleset

The first comparative study is conducted with the two techniques AHC and PAM on the same ruleset $\mathcal{R}_1$ [HGB05c]. As a first result, a subset of 35 measures are used for evaluating. Tab. 7.5 gives the clusters of measures obtained from the AHC technique on the $\mathcal{R}_1$ ruleset with a dissimilarity interval = 0.15. Tab. 7.6 represents the clusters obtained from the PAM technique. As discussed above, we can see the clusters that are agreed by both AHC and PAM in Tab. 7.7.

The subset of 35 measures are numbered from 0 to 34 as the following orders [HGB05c]: Causal Confidence (0), Causal Confirm (1), Causal Confirmed-Confidence (2), Causal Support (3), Collective Strength (4), Confidence (5), Conviction (6), Cosine (7), Dependency (8), Descriptive Confirm

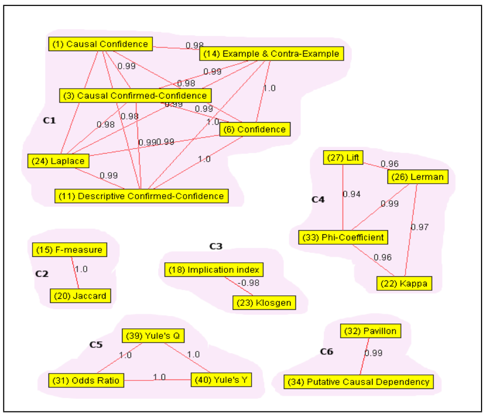


Figure 7.12: $\overline{CG+}$ graph with Spearman's coefficient.

Ruleset	Number of correlations		Number of clusters	
	τ-correlated	θ-uncorrelated	CG+	CG0
$\mathcal{R}_1$	37	2	24	38
$\mathcal{R}_2$	56	5	18	35
$\mathcal{R}_3$	88	46	19	1
$\mathcal{R}_4$	106	47	13	1
$\mathcal{R}_1^{'}$	27	28	24	18
$\mathcal{R}_2^{'}$	35	17	22	25
$\mathcal{R}_3^{'}$	80	87	19	1
$\mathcal{R}_4^{'}$	61	72	16	1

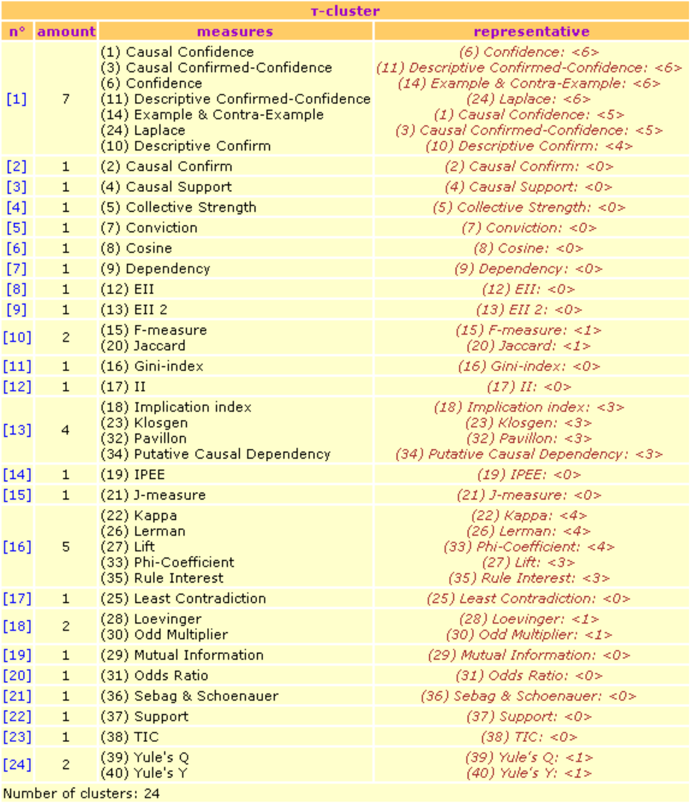
Table 7.3: Comparison of correlations with Kendall's coefficient.

(9), Descriptive Confirmed-Confidence (10), EII (11), EII 2(12), Example & Contra-Example (13), Gini-index (14), II (15), Jaccard (16), J-measure (17), Kappa (18), Klosgen (19), Laplace (20), Least Contradiction (21), Lift (22), Loevinger (23), Odds Ratio(24), Pavillon (25), Phi-Coefficient (26), Putative Causal Dependency (27), Rule Interest (28), Sebag & Schoenauer (29), Similarity Index[1] (30), Support (31), TIC (32), Yule's Q (33), Yule's Y (34).

7.6.2 On two opposite rulesets

We experiment on two opposite rulesets $\mathcal{R}_1$, $\mathcal{R}_2$ together with their samples $\mathcal{R}'_1$, $\mathcal{R}'_2$. Another subset of 36 objective measures are used this second comparative study [HGB06e]: Causal Confidence (0), Causal Confirm (1), Causal Confirmed-Confidence (2), Causal Support (3), Collective Strength (4), Confidence (5), Conviction (6), Cosine (7), Dependency (8), Descriptive Confirm (9), Descriptive Confirmed-Confidence (10), EII (11), EII 2 (12), Example & Contra-Example (13), Gini-index (14), II (15), IPEE (16), Jaccard (17), J-measure (18), Kappa (19), Klosgen (20), Laplace (21), Least Contradiction (22), Lerman (23), Lift(24), Loevinger (25), Odds Ratio (26), Pavillon (27), Phi-Coefficient (28), Putative Causal Dependency (29), Rule Interest (30), Sebag & Schoenauer (31),

[1] This is another name of the Lerman measure in Tab. 3.1 [Chap. 3].



τ-cluster			
n°	amount	measures	representative
[1]	7	(1) Causal Confidence (3) Causal Confirmed-Confidence (6) Confidence (11) Descriptive Confirmed-Confidence (14) Example & Contra-Example (24) Laplace (10) Descriptive Confirm	*(6) Confidence: <6>* *(11) Descriptive Confirmed-Confidence: <6>* *(14) Example & Contra-Example: <6>* *(24) Laplace: <6>* *(1) Causal Confidence: <5>* *(3) Causal Confirmed-Confidence: <5>* *(10) Descriptive Confirm: <4>*
[2]	1	(2) Causal Confirm	*(2) Causal Confirm: <0>*
[3]	1	(4) Causal Support	*(4) Causal Support: <0>*
[4]	1	(5) Collective Strength	*(5) Collective Strength: <0>*
[5]	1	(7) Conviction	*(7) Conviction: <0>*
[6]	1	(8) Cosine	*(8) Cosine: <0>*
[7]	1	(9) Dependency	*(9) Dependency: <0>*
[8]	1	(12) EII	*(12) EII: <0>*
[9]	1	(13) EII 2	*(13) EII 2: <0>*
[10]	2	(15) F-measure (20) Jaccard	*(15) F-measure: <1>* *(20) Jaccard: <1>*
[11]	1	(16) Gini-index	*(16) Gini-index: <0>*
[12]	1	(17) II	*(17) II: <0>*
[13]	4	(18) Implication index (23) Klosgen (32) Pavillon (34) Putative Causal Dependency	*(18) Implication index: <3>* *(23) Klosgen: <3>* *(32) Pavillon: <3>* *(34) Putative Causal Dependency: <3>*
[14]	1	(19) IPEE	*(19) IPEE: <0>*
[15]	1	(21) J-measure	*(21) J-measure: <0>*
[16]	5	(22) Kappa (26) Lerman (27) Lift (33) Phi-Coefficient (35) Rule Interest	*(22) Kappa: <4>* *(26) Lerman: <4>* *(33) Phi-Coefficient: <4>* *(27) Lift: <3>* *(35) Rule Interest: <3>*
[17]	1	(25) Least Contradiction	*(25) Least Contradiction: <0>*
[18]	2	(28) Loevinger (30) Odd Multiplier	*(28) Loevinger: <1>* *(30) Odd Multiplier: <1>*
[19]	1	(29) Mutual Information	*(29) Mutual Information: <0>*
[20]	1	(31) Odds Ratio	*(31) Odds Ratio: <0>*
[21]	1	(36) Sebag & Schoenauer	*(36) Sebag & Schoenauer: <0>*
[22]	1	(37) Support	*(37) Support: <0>*
[23]	1	(38) TIC	*(38) TIC: <0>*
[24]	2	(39) Yule's Q (40) Yule's Y	*(39) Yule's Q: <1>* *(40) Yule's Y: <1>*
Number of clusters: 24			

Figure 7.13: τ-cluster on $\mathcal{R}_1$ with Kendall's coefficient.

Support (32), TIC (33), Yule's Q (34), Yule's Y (35).

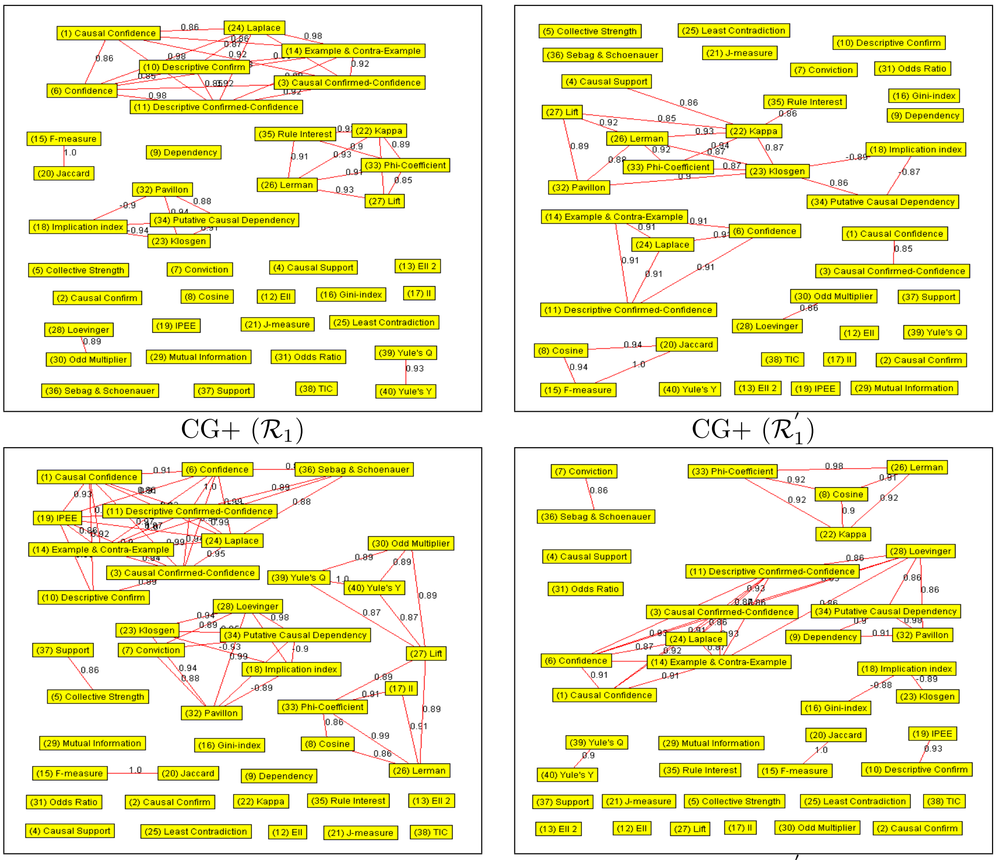
To have a complete evaluation with the two techniques AHC and PAM, three approaches are conducted: (i) evaluating with AHC on $\mathcal{R}_1$, $\mathcal{R}'_1$, $\mathcal{R}_2$, $\mathcal{R}'_2$; (ii) evaluating with PAM on $\mathcal{R}_1$, $\mathcal{R}'_1$, $\mathcal{R}_2$, $\mathcal{R}'_2$; and (iii) evaluating with AHC and PAM on $\mathcal{R}_1$, $\mathcal{R}'_1$, $\mathcal{R}_2$, $\mathcal{R}'_2$.

With AHC

Tab. 7.8 shows the clusters obtained from the four rulesets $\mathcal{R}_1$, $\mathcal{R}'_1$, $\mathcal{R}_2$, $\mathcal{R}'_2$ in columns respectively. Tab. 7.9 gives details on the agreed clusters between each ruleset pairs $\mathcal{R}_1 \cap \mathcal{R}_1^{'}$,$\mathcal{R}_2 \cap \mathcal{R}_2^{'}$, $\mathcal{R}_1^{'} \cap \mathcal{R}_2^{'}$, $\mathcal{R}_1 \cap \mathcal{R}_2$. The content in the last column is the agreed clusters over all the four rulesets $\mathcal{R}_1 \cap \mathcal{R}_1^{'} \cap \mathcal{R}_2 \cap \mathcal{R}_2^{'}$.

With PAM

The same purpose as the precedent section on evaluating with AHC on the four rulesets illustrated in Tab. 7.10 and Tab. 7.11.



CG+ $(\mathcal{R}_1)$

CG+ $(\mathcal{R}_1^{'})$

CG+ $(\mathcal{R}_2)$

CG+ $(\mathcal{R}_2^{'})$

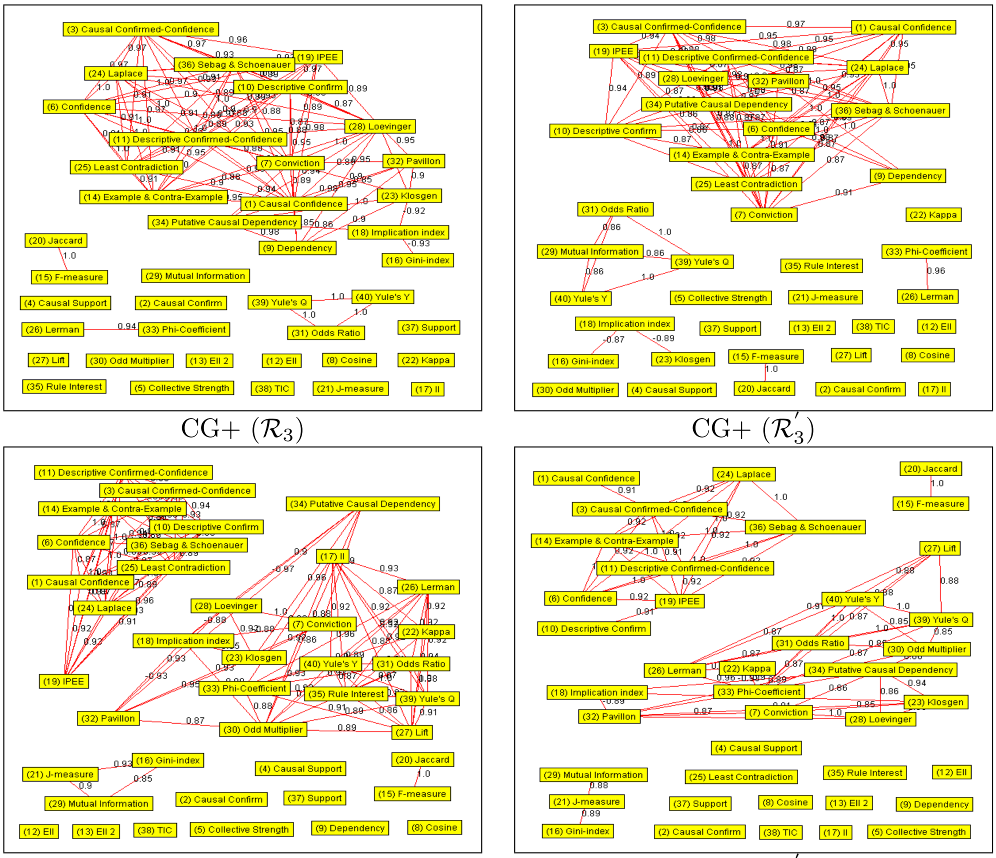
Figure 7.14: CG+ graphs with Kendall's coefficient.

Agreed clusters

Tab. 7.12 shows the clusters of measures that are agreed with both AHC and PAM on the four rulesets $\mathcal{R}_1$, $\mathcal{R}'_1$, $\mathcal{R}_2$, and $\mathcal{R}'_2$.

7.7 Summary

Discovering the behaviors of interestingness measures is a research challenge. The results obtained can help the user to understand different hidden aspect on a specific ruleset or a set of rulesets. We have conducted comparative studies with three techniques CG, AHC and PAM. The datasets that have opposite characteristics are chosen to evaluate. By analyzing the stable clusters or agreed clusters, some stable/agreed clusters are found indicating an invariance with the nature of the dataset!



CG+ ($\mathcal{R}_3$)

CG+ ($\mathcal{R}_3^{'}$)

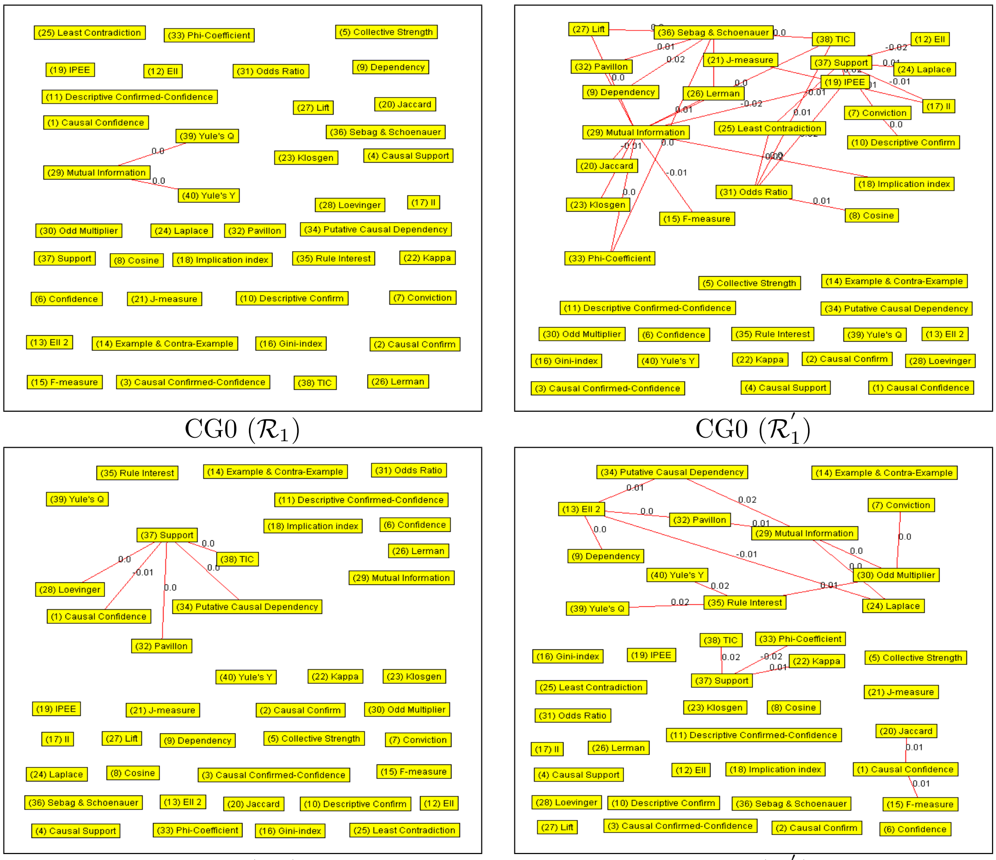
CG+ ($\mathcal{R}_4$)

CG+ ($\mathcal{R}_4^{'}$)

Figure 7.15: CG+ graphs with Kendall's coefficient (cont.).

	P	S	K	PS	PK	SK	PSK
τ-correlated	110	113	37	73	36	36	36
θ-uncorrelated	1	0	2	0	0	0	0
CG+	10	10	24	5	5	5	5
CG0	39	40	38	0	0	0	0

Table 7.4: Comparison of correlation with all coefficients on the $\mathcal{R}_1$ ruleset.



CG0 ($\mathcal{R}_1$)

CG0 ($\mathcal{R}_1^{'}$)

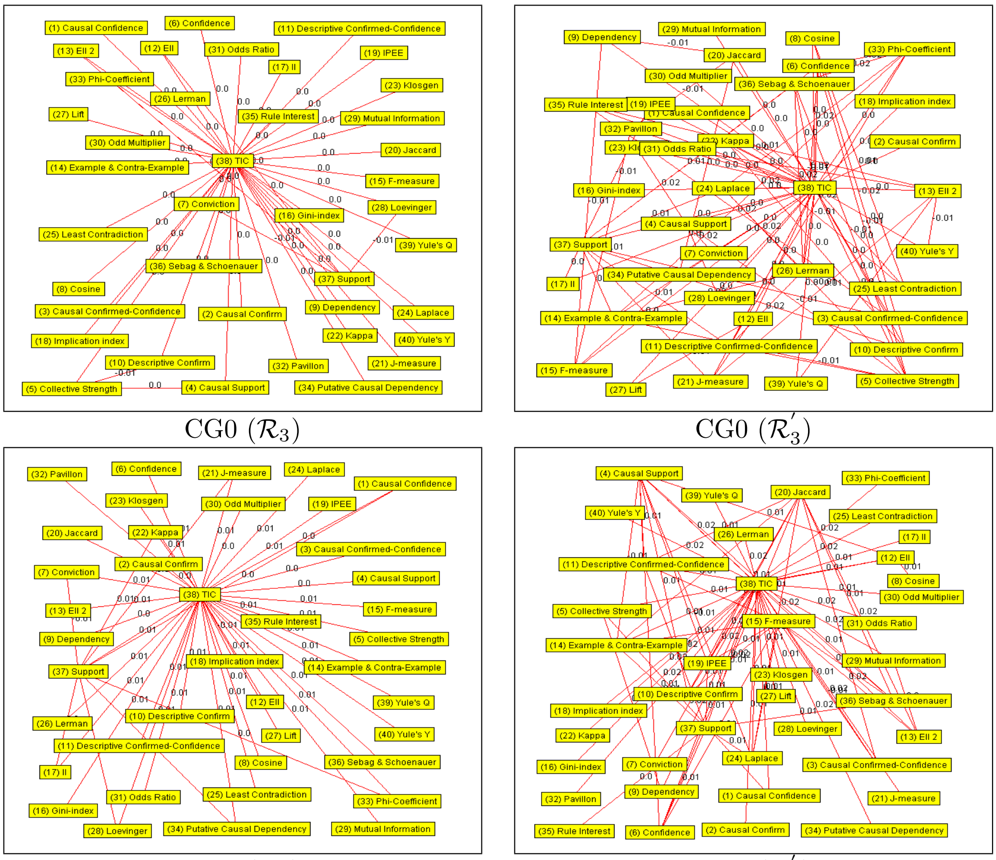
CG0 ($\mathcal{R}_2$)

CG0 ($\mathcal{R}_2^{'}$)

Figure 7.16: CG0 graphs with Kendall's coefficient.

Cluster	$\mathcal{R}_1$
1	Causal Confidence, Causal Confirmed-Confidence, Confidence, Descriptive Confirmed-Confidence, Laplace
2	Causal Confirm, Descriptive Confirm, Example & Contra-Example,Least Contradiction
3	Causal Support
4	Collective Strength
5	Conviction
6	Cosine, Jaccard
7	Dependency, Kappa, Klosgen, Lift, Pavillon, Phi-Coefficient, Putative Causal Dependency, Rule Interest, Similarity Index
8	EII, EII 2
9	Gini-index, J-measure
10	II
11	Loevinger
12	Odds Ratio
13	Sebag & Schoenauer
14	Support
15	TIC
16	Yule's Q, Yule's Y

Table 7.5: Clusters of measures with AHC on $\mathcal{R}_1$ (dissimilarity interval = 0.15).



CG0 ($\mathcal{R}_3$)

CG0 ($\mathcal{R}_3^{'}$)

CG0 ($\mathcal{R}_4$)

CG0 ($\mathcal{R}_4^{'}$)

Figure 7.17: CG0 graphs with Kendall's coefficient (cont.).

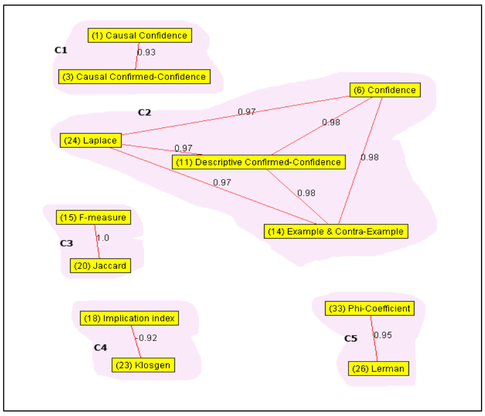
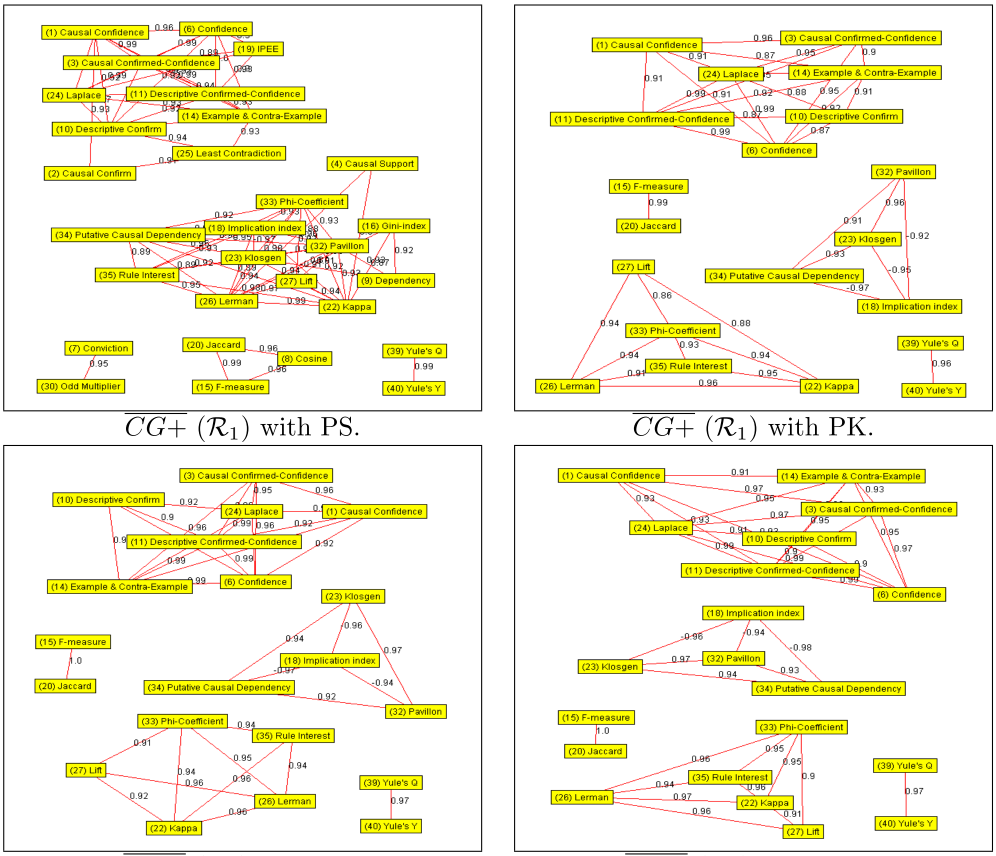


Figure 7.18: $\overline{CG+}$ graph with Kendall's coefficient.



$\overline{CG+}$ ($\mathcal{R}_1$) with PS.

$\overline{CG+}$ ($\mathcal{R}_1$) with PK.

$\overline{CG+}$ ($\mathcal{R}_1$) with SK.

$\overline{CG+}$ ($\mathcal{R}_1$) with PSK.

Figure 7.19: $\overline{CG+}$ graphs with PS, PK, SK and PSK summaries on $\mathcal{R}_1$.

Cluster	$\mathcal{R}_1$
1	Causal Confidence, Causal Confirmed-Confidence, Confidence, Descriptive Confirmed-Confidence, Laplace
2	Causal Confirm, Descriptive Confirm, Example & Contra-Example, Least Contradiction
3	Causal Support, Kappa, Lift, Phi-Coefficient, Rule Interest, Similarity Index
4	Collective Strength
5	Conviction
6	Cosine, Jaccard
7	Dependency, Klosgen, Pavillon,Putative Causal Dependency
8	EII, EII 2
9	Gini-index, J-measure
10	II
11	Loevinger
12	Odds Ratio
13	Sebag & Schoenauer
14	Support
15	TIC
16	Yule's Q, Yule's Y

Table 7.6: Clusters of measures with PAM on the $\mathcal{R}_1$ ruleset.

Cluster	$\mathcal{R}_1$
1	Causal Confidence, Causal Confirmed-Confidence, Confidence, Descriptive Confirmed-Confidence, Laplace
2	Causal Confirm, Descriptive Confirm, Example & Contra-Example, Least Contradiction
3	Collective Strength
4	Conviction
5	Cosine, Jaccard
6	Dependency, Klosgen, Pavillon,Putative Causal Dependency
7	EII, EII 2
8	Gini-index, J-measure
9	II
10	Kappa, Lift, Phi-Coefficient, Rule Interest, Similarity Index
11	Loevinger
12	Odds Ratio
13	Sebag & Schoenauer
14	Support
15	TIC
16	Yule's Q, Yule's Y

Table 7.7: Clusters agreed with both AHC and PAM on the $\mathcal{R}_1$ ruleset.

$\mathcal{R}_1$	$\mathcal{R}_1^{'}$	$\mathcal{R}_2$	$\mathcal{R}_2^{'}$
0,2,5,10,21	0,2,5,10,21	0,2,5,8,10,11,12,16,21,25,27,29	0,2,5,8,10,21,25,27,29
1,9,13,22	1,9,13,22	1,9	1,9
3	3,19,20,23,24,27,28,30	3	3
4	4	4	4,32
6	6	6,31	6,31
7,17	7,17	7,17,19,23,28	7,17,19,23,28,30,34,35
8,14,18	8,14,18		
11,12,16	11,12		11,12,16
		13	13
		14,18	14,18,20
15	15	15,34,35	15
	16		
19,20,23,27,28,29,30,34,35			
		20	
		22	22
24		24	24,26
25	25,29		
26	26	26	
		30	
31	31		
32	32	32	
33	33	33	33
	34,35		

Table 7.8: Clusters of measures with AHC (measures are represented by their orders).

$\mathcal{R}_1 \cap \mathcal{R}_1^{'}$	$\mathcal{R}_2 \cap \mathcal{R}_2^{'}$	$\mathcal{R}_1^{'} \cap \mathcal{R}_2^{'}$	$\mathcal{R}_1 \cap \mathcal{R}_2$	$\mathcal{R}_1 \cap \mathcal{R}_1^{'} \cap \mathcal{R}_2 \cap \mathcal{R}_2^{'}$
0,2,5,10,21	0,2,5,8,10,21,25,27,29	0,2,5,10,21	0,2,5,10,21	0,2,5,10,21
1,9,13,22	1,9	1,9	1,9	1,9
	3		3	
4			4	
6	6,31			
7,17	7,17,19,23,28	7,17	7,17	7,17
8,14,18				
11,12	11,12,16	11,12	11,12,16	11,12
	13			
	14,18	14,18	14,18	14,18
15		15		
19,20,23,27,28,30		19,23,28,30	19,23,28	19,23,28
	22			
			24	
		25,29		
26			26	
			27,29	
31				
32			32	
33	33	33	33	33
34,35	34,35	34,35	34,35	34,35

Table 7.9: Cluster comparison from AHC (measures are represented by their orders).

$\mathcal{R}_1$	$\mathcal{R}_1^{'}$	$\mathcal{R}_2$	$\mathcal{R}_2^{'}$
0,2,5,10,21	0,1,2,5,10,21	0,2,5,8,10,21,25,27,29	0,2,5,8,10,21,25,27,29
1,9,13,22		1,9	1,9
3,19,23,28,30,34,35	3,19,20,23,24,27,28,30	3	3
4	4	4	4,14,18,20
6	6	6,31	6,31
7,17	7,17	7,17,19,23,28	7,17,19,23,28,30
8,14,18	8,14,18		
	9,13,22		
11,12,16	11,12	11,12,16	11,12,16
		13	13
		14,18,30	
15	15	15,34,35	15,34,35
	16		
20,27,29		20	
		22	22
24		24	24,26
25	25,29		
26	26	26	
31	31		
32	32	32	32
33	33	33	33
	34,35		

Table 7.10: Clusters of measures with PAM (measures are represented by their orders).

$\mathcal{R}_1 \cap \mathcal{R}_1^{'}$	$\mathcal{R}_2 \cap \mathcal{R}_2^{'}$	$\mathcal{R}_1^{'} \cap \mathcal{R}_2^{'}$	$\mathcal{R}_1 \cap \mathcal{R}_2$	$\mathcal{R}_1 \cap \mathcal{R}_1^{'} \cap \mathcal{R}_2 \cap \mathcal{R}_2^{'}$
0,2,5,10,21	0,2,5,8,10,21,25,27,29	0,2,5,10,21	0,2,5,10,21	0,2,5,10,21
	1,9		1,9	
3,19,23,28,30	3			
4			4	
6	6,31			
7,17	7,17,19,23,28	7,17	7,17	7,17
8,14,18				
9,13,22				
11,12	11,12,16	11,12	11,12,16	11,12
	13			
	14,18	14,18	14,18	14,18
15	15,34,35			
		19,23,28,30	19,23,28	19,23,28
20,27				
	22			
			24	
		25,29		
26			26	
			27,29	
31				
32	32	32	32	32
33	33	33	33	33
34,35		34,35	34,35	34,35

Table 7.11: Cluster comparison from PAM (measures are represented by their orders).

Agreed cluster	$\mathcal{R}_1 \cap \mathcal{R}_1^{'} \cap \mathcal{R}_2 \cap \mathcal{R}_2^{'}$
1	Causal Confidence, Causal Confirmed-Confidence, Confidence, Descriptive Confirmed-Confidence, Laplace
2	Cosine, Jaccard
3	EII, EII 2
4	Gini-index, J-measure
5	Kappa, Lerman, Phi-Coefficient
6	Support
7	TIC
8	Yule's Q, Yule's Y

Table 7.12: Clusters agreed with both AHC and PAM on the $\mathcal{R}_1$, $\mathcal{R}'_1$, $\mathcal{R}_2$, and $\mathcal{R}'_2$ rulesets.

Chapter 8

Conclusions and perspectives

In the last decade, the designing of interestingness measure to evaluate association rules has become a challenge in the context of KDD. This is because association rule [AIS93a] [AS94] [AMS^{+}96] is one of the few models dedicated to unsupervised discovery of rule tendencies in data. It is unfortunately confronted to a major difficulty: the user must cope with a large amount of extracted rules in order to validate and select the best ones [PS91]. One way to reduce the cost of the user's task is to help him/her with a postprocessing task of association rules. Five postprocessing approaches are proposed in this task: constraints, pruning, grouping, summarizing and visualizing. In this chapter, we describe the main results of our work in the postprocessing of association rules and propose some future researches.

8.1 Main results

Postprocessing of association rules is an important task in the KDD process. The enormous number of rules discovered in the mining task requires not only an efficient postprocessing task but also an adapted results with the user's preferences. Depending on the user's point of view, each interestingness measure reflects his/her own interests on the data. An interestingness measure has its own ranking on the discovered rules, the most important rules have the highest ranks. As we known, it is difficult to have a common ranking on a set of association rules for all the interestingness measures. As much interestingness measures appeared, the problem is more difficult. As a solution, a set of interestingness measures having the same or next to a common ranking on a set of association rules will be classified into a group. In this group, a representative measure that has the strongest relations with the other measures in the group will be held. Now we can reduce all the measures in this group into a representative measure. The original set of interestingness measures will become a smaller set of representative measures. This is the principal approach of our thesis in the postprocessing task of association rules in KDD.

The main results of our thesis can be evaluated as the following:

- We have strengthened some important features to the IAS model [LHML99] in our approach, implemented in the ARQAT tool [Chap. 5].
- We have conducted a global survey on the principal properties of interestingness measures, particularly on objective interestingness measures. An important set of objective measures

(i.e., with 40 measures) is collected and each measure can have its own formula with four parameters $n, n_X, n_Y, n_{X\overline{Y}}$. The principal properties on objective measures considered in our study are: variation, particular situation, paradoxical phenomenon, countable, diversification, discriminative ability, interpretable, imbalance, attribute interestingness and quasi- [Chap. 3]. A classification of objective measures on some important properties are performed. Some interesting relation on mathematical formula are also introduced.

- A quick review on five principal approaches in the postprocessing task in the KDD process on association rules [Chap. 4].
- We have introduced a new approach in the postprocessing task, called representative measure. The experimentations are performed on three strong techniques AHC, PAM and correlation graph. Three common correlation coefficients of Pearson, Spearman and Kendall are implemented [Chap. 4].
- The results are performed on both real-life and synthetic datasets. The datasets are also studied both on correlated and weakly correlated datasets [Chap. 6]. The new exploratory approach (graphical and interactive) permits a quick interpretation of the results
- We have proposed a sample model reduced from a dataset. Some important features on the variation of interestingness values, joint-distribution matrix, correlation values, most interesting rules, sensitivity values and comparative studies are represented [Chap. 6].
- We have introduced an important notion of stable clusters between interestingness measures. As a result, another important notion of graph of stable clusters is also given [Chap. 7]. Some stable clusters issued from different datasets have shown unexpected stabilities.
- Complementary points of view and implemented measures allow an expert to better understand the correlations existing between the interestingness measures on his/her ruleset. A lot of tasks/views will give a fine/precise study of association rules.

8.1.1 ARQAT tool

We have designed and described the features of a new tool, ARQAT (Association Rule Quality Analysis Tool) [HGB05a] developed at the Polytechnic school of Nantes university, implementing an exploratory data-analysis approach for studying the behavior of interestingness measures on a specific dataset. Technically, ARQAT is a graphical tool, written in Java over 8000 lines of codes, and embeds a set of 14 graphical views. The user operates on an interface through a classical web-browser, using web technologies. The main features of the tool are: (1) ruleset analysis, (2) correlation and clustering analysis, (3) most interesting rules analysis, (4) sensitivity analysis, (5) comparative analysis. For exchange facilities, three common file formats are used for importing/exporting the rulesets: PMML (XML data-mining standard), CSV (Excel and SAS) and ARFF (used by WEKA). ARQAT will be freely available at http://www.polytech.univ-nantes.fr/arqat.

With ARQAT, we have shown the interest of such an exploratory approach, where the intensive use of graphical and complementary visualizations improves and facilitates data insight for the user. ARQAT is a first step toward a larger analysis platform in the postprocessing of association rules. Most of the important experimental results are all issued from this tool. Some other intermediate results, e.g. with the agglomerative hierarchical clustering (AHC) and partitioning around medoids (PAM) representations, are computed from the R[1] tool.

[1] http://www.r-project.org/

8.1.2 Correlation graph

The new approach called correlation graph is implemented by ARQAT. Two graph types CG+ and CG0 are proposed to evaluate measures by using graphs as a visual insight on the data. With this approach, the decision-maker has a few measures to decide. These graphical representations will help him/her to select the most interesting rules to examine.

We also improved our clustering results by using three common correlation coefficients: Pearson, Spearman and Kendall. Both the value and rank aspects are considered.

8.1.3 Representative measure

In order to improve the postprocessing of association rules, we have presented a new approach to finding a minimum set of suitable objective measures, called representative measures. The new approach finds the representative measures with the help of AHC, PAM or correlation graph (CG) particularly. These techniques allow the user to obtain not only the suitable measures representing the hidden aspects in the dataset, but also a graphical representation to evaluate the clustering results.

With a cluster of objective measures, we will find the "central of gravity" of this cluster. For example [HGB06b], when we applied to a rule-based dataset about 120000 association rules with 36 objective measures, we obtained a reduced set of sixteen representative measures. The result obtained also facilitates the validation of the most interesting rules.

8.1.4 Comparison study

To understand the behavior of the interestingness measures on a specific dataset, we have studied and compared the various interestingness measures described in the literature in order to help the decision-maker to better understand the behavior of the measures in the stage of postprocessing of association rules. Our approach is the first step towards the process of evaluating the knowledge issued in the form of association rules in the domain of knowledge quality research.

We use a data analysis approach based on the dissimilarity computed between interestingness measures in order to evaluate the behavior of 40 interestingness measures. We also determine some stable clusters with significant results: cluster that seems independent from the nature of data and the selection of rules. The stable clusters denote an interesting relations between measures because they remark the stable behaviors. We also evaluate the behavior of the measures on some important clusters agreed with each others. An interpretation of this clustering results has been proposed. With these results, the decision-maker will decide what measures are interesting to capture the best knowledge.

The evaluation of numerous measures on the datasets having opposite characteristics is an important method. Based on two types of correlation graphs CG+ and CG0, we have found such stable clusters, called τ-stable in the $\overline{CG+}$ graphs but not in the $\overline{CG0}$ graphs. We have introduced the notions of couple and multiple rulesets to examine the behaviors of interestingness measures on many rulesets efficiently. The stable clusters are evaluated on three situations with Pearson's, Spearman's and Kendall's coefficients. In particular, the stable clusters issued from these three coefficients are also observed. The results with the techniques AHC and PAM are also illustrated.

8.1.5 Cluster evaluation

Besides, we have also proposed a way to study the k_m most interesting rules of each cluster, the union and intersection of the ten most interesting rules of all the cluster in relation to the current cluster. The union of the k_m most interesting rules for all the clusters is also presented for the user's choice. For the first presentation of the results, we just use 40 measures for implementation.

Other functions such as ranking the measures by sensitivity values, ruleset characteristics, distribution of interestingness measures, smart views of scatterplot matrix both on interestingness values and rankings, summary tables will help the user to understand the different aspects in the datasets graphically and exploratory.

8.2 Perspectives

Here we propose some future approaches based on our results obtained to improve the strength of the postprocessing task of association rules.

8.2.1 Improving the sample model

Having a sample ruleset from an original ruleset is a completely new result from our approach. Each interestingness measure will have a set of the most interesting association rules from an original ruleset. This set of rules obtained by ranking the original ruleset with the corresponding interestingness values. The most interesting rules are the ones that have the highest values of interestingness. The sample ruleset is the created by collecting all the most interesting rules calculated from all the interestingness measures. The number k_m (see Sec. 5.4.3 [Chap. 5]) of rules that have highest ranks of each will be chosen by the user. To improve the efficiency of the sample model, we can proceed with the following directions:

- Evaluating the relation between the variation of the value of k_m with the characteristic values of each ruleset calculated from the sample ruleset.
- Continuing to develop the criteria on the characteristic types for each ruleset that have the close relation with the criteria on interestingness values [Chap. 3]. This work leads to a "better" sample ruleset and the sample ruleset will be considered as the most representative sample of the original ruleset. It will help the user to choose the value of k_m easier.
- With an association rule, each interestingness measure will give a different interestingness value. We will exchange this value into a common interval. Each association rule will be assigned an average value of interestingness calculated from all the interestingness measures. The sample ruleset is now considered as the set of association rules that have the highest values of the average interestingness. The value of k_m is given by the user normally.
- Using the sample ruleset in the study of aggregated rules in comparison with the original ruleset.

8.2.2 Improving the cluster evaluation

The cluster (i.e. of interestingness measures) evaluation can be improved by studying the relation between the size of value k_m with the most interesting rules found.

When we evaluate the relation between a target cluster with the other clusters, each rule in the target cluster can be assigned a value as the average rank calculated from the same rule in the other clusters. Another technique is to use the bagging predictors to get an aggregated predictor of rules [Bre96]. From this result, we can find a more small set of representative rules for each cluster. The integration with a or many domain knowledge will give an experience view in the process of evaluation. The most suitable measure in a cluster can be obtained with the technique of random forests [Bre01] (i.e., by choosing the best features).

8.2.3 Hierarchy view

Different types of the stable hierarchy such as original, sample, couple, multiple, all, complement, single rulesets issued from a set of rulesets will be evaluated to have a hierarchy view on the set of ruleset.

8.2.4 Aggregation technique

One of the means to reduce the number of rules consists to develop the numeric indicators called interestingness measures that allows to choose the most interesting rules. However, numerous interestingness measures are available in the literature (we have collected about 40 interestingness measures) and but they do not adapted to the user's demand. In reality, we have to combine numerous measures to create an aggregated measure that is more performance.

It is necessary to implement an aggregated technique by Choquet's or Sugeno's integral to form an aggregated measure [Koj04] [Mar00]. In order to experiment and validate the proposed solutions, this work can be applied on the experimental data on which the interestingness values have already calculated.

Improving the clustering analysis with aggregation techniques to facilitate the user's decision making from the most suitable interestingness measures is still an attractive challenge.

8.3 Publication

Book chapter

- Xuan-Hiep Huynh, Fabrice Guillet, Julien Blanchard, Pascale Kuntz, Régis Gras, and Henri Briand. A graph-based clustering approach to evaluate interestingness measures : a tool and a comparative study. (Chapter 2) Quality measures in data mining. Springer-Verlag, 2006 (To appear).

National journal

- Xuan-Hiep Huynh, Fabrice Guillet, and Henri Briand. Comparaison des mesures d'intérêt de règles d'association : une approche basée sur des graphes de corrélation. Revue des Nouvelles Technologies de l'Information, RNTI-E-6(2). Cépaduès Edition, France, pp. 549-560, 2006.
- Xuan-Hiep Huynh, Fabrice Guillet, and Henri Briand. ARQAT: plateforme exploratoire pour la qualité des règles d'association. Revue des Nouvelles Technologies de l'Information, Extraction des connaissances : état et perspectives, RNTI-E-5. Cépaduès Edition, France, 2006 (To appear).

- Xuan-Hiep Huynh, Fabrice Guillet, and Henri Briand. Une plateforme exploratoire pour la qualité des règles d'association : apports pour l'analyse implicative. Revue " Quaderni di Ricerca in Didattica", Italy, pp. 339-349, 2005.

International conference

- Xuan-Hiep Huynh, Fabrice Guillet, and Henri Briand. Clustering interestingness measures with positive correlation. ICEIS'05, Proceedings of the 7th International Conference on Enterprise Information Systems. Miami, USA, pp. 248-253, 2005.
- Xuan-Hiep Huynh, Fabrice Guillet, and Henri Briand. ARQAT: an exploratory analysis tool for interestingness measures. ASMDA'05, Proceedings of the 11th International Symposium on Applied Stochastic Models and Data Analysis. Brest, France, pp. 334-344, 2005.
- Xuan-Hiep Huynh, Fabrice Guillet, and Henri Briand. A data analysis approach for evaluating the behavior of interestingness measures. DS'05, Proceedings of the 8th International Conference on Discovery Science, LNAI 3735. Springer-Verlag, Singapore, pp. 330-337, 2005.
- Xuan-Hiep Huynh, Fabrice Guillet, and Henri Briand. Extracting representative measures for the post-processing of association rules. IEEE RIVF'06, Proceedings of the 4th IEEE International Conference on Computer Sciences: Research & Innovation - Vision for the Future. Ho Chi Minh city, Vietnam, pp. 99-105, 2006 (Best paper of track "Software Engineeing, Knowledge Engineering, Agents and Interfaces").
- Xuan-Hiep Huynh, Fabrice Guillet, and Henri Briand. Discovering the stable clusters between interestingness measures. ICEIS'06, Proceedings of the 8th International Conference on Enterprise Information Systems. Cyprus, pp. 196-201, 2006.
- Xuan-Hiep Huynh, Fabrice Guillet, and Henri Briand. Evaluating interestingness measures with linear correlation graph. IEA-AIE'06, Proceedings of the 19th International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems, LNCS 4031. Springer-Verlag, Annecy, France, pp. 312-321, 2006.
- Xuan-Hiep Huynh, Fabrice Guillet, and Henri Briand. A graph-based approach for comparing interestingness measures. IEEE ICEIS'06, Proceedings of the First IEEE International Conference on Engineering of Intelligent Systems. Islamabad, Pakistan, pp. 375-380, 2006.

National conference/National workshop

- Xuan-Hiep Huynh. Une approche exploratoire pour la qualié des règles d'association. JDOC'05, Actes de 5èmes Journées des Doctorants. Nantes, France, pp. 89-92, 2005.
- Xuan-Hiep Huynh, Fabrice Guillet, and Henri Briand. ARQAT: plateforme exploratoire pour la qualité des règles d'association. DKQ'05, Actes d'Atelier Qualité des Données et des Connaissances - Associé à EGC'05, 5èmes Journées Francophones d'Extraction et de Gestion des Connaissance. Paris, France, pp. 58-68, 2005.
- Xuan-Hiep Huynh, Fabrice Guillet, and Henri Briand. Extraction de mesures d'intérêt représentatives pour le post-traitement des règles d'association. DKQ'06, Actes d'Atelier Qualité des Données et des Connaissances - Associé à EGC'06, 6èmes Journées Francophones d'Extraction et de Gestion des Connaissances. Lille, France, pp. 45-54, 2006.

Bibliography

[AAFF05] A. Amir, Y. Aumann, R. Feldman & M. Fresko – "Maximal association rules: A tool for mining associations in text", *Journal of Intelligent Information Systems* **25(3)** (2005), p. 333–345.

[AAP01] R. C. Agarwal, C. C. Aggarwal & V. V. V. Prasad – "A tree projection algorithm for generation of frequent itemsets", *Journal of Parallel and Distributed Computing* **61(3)** (2001), p. 350–371.

[AIS93a] R. Agrawal, T. Imielinski & A. Swami – "Mining association rules between sets of items in large databases", *Proceedings of 1993 ACM-SIGMOD International Conference on Management of Data* (1993), p. 207–216.

[AIS93b] R. Agrawal, T. Imielinski & A. Swami – "Database mining: A performance perspective", *IEEE Transactions on Knowledge and Data Engineering* **5(6)** (1993), p. 914–925.

[AK02] J. Azé & Y. Kodratoff – "A study of the effect of noisy data in rule extraction systems", *EMCSR'02, Proceedings of the Sixteenth European Meeting on Cybernetics and Systems Research* (2002), p. 781–788.

[AMS+96] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen & A. I. Verkamo – "Fast discovery of association rules", *Advances in knowledge discovery and data mining* (1996), p. 307–328.

[AS94] R. Agrawal & R. Srikant – "Fast algorithms for mining association rules", *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases* (1994), p. 487–499.

[AS96] R. Agrawal & J. C. Shafer – "Parallel mining of association rules", *IEEE Transactions on Knowledge and Data Engineering* **8(6)** (1996), p. 962–969.

[AT97] G. Adomavicius & A. Tuzhilin – "Discovery of actionable patterns in databases: the action hierarchy approach", *KDD'97, Proceedings of the Third International Conference Data Mining and Knowledge Discovery* (1997), p. 111–114.

[AT01] G. Adomavicius & A. Tuzhilin – "Expert-driven validation of rule-based user models in personalization applications", *Data Mining and Knowledge Discovery* **5(1-2)** (2001), p. 33–58.

[BA96] R. J. Brachman & T. Anand – "The process of knowledge discovery in databases", *Advances in Knowledge Discovery and Data Mining* (1996), p. 37–57.

[BA99] R. J. J. Bayardo & R. Agrawal – "Mining the most interestingness rules", *KDD'99, Proceedings of the 5th ACM SIGKDD International Confeference On Knowledge Discovery and Data Mining* (1999), p. 145–154.

[BAG00] R. J. J. Bayardo, R. Agrawal & D. Gunopulos – "Constraint-based rule mining in large, dense databases", *Data Mining and Knowledge Discovery* **4(2-3)** (2000), p. 217–240.

[BB05] S. Bistarelli & F. Bonchi – "Interestingness is not a dichotomy: introducing softness in constrained pattern mining", *PKDD'05, 9th European Conference on Principles and Practice of Knowledge Discovery in Databases* **LNCS 3721** (2005), p. 22–33.

[BBJ00] J.-F. Boulicaut, A. Bykowski & B. Jeudy – "Towards the tractable discovery of association rules with negations", *FQAS'00, Proceedings of the Fourth International Conference on Flexible Query Answering Systems* (2000), p. 425–434.

[BCF$^+$05] D. Burdick, M. Calimlim, J. Flannick, J. Gehrke & T. Yiu – "Mafia: A maximal frequent itemset algorithm", *IEEE Transactions on Knowledge and Data Engineering* **17(11)** (2005), p. 1490–1504.

[BGB03a] J. Blanchard, F. Guillet & H. Briand – "A user-driven and quality oriented visualization for mining association rules", *ICDM'03, Proceedings of the Third IEEE International Conference on Data Mining* (2003), p. 493–497.

[BGB03b] J. Blanchard, F. Guillet & H. Briand – "Exploratory visualization for association rule rummaging", *MDM/KDD'03, Workshop on Multimedia Data Mining in conjunction with ACM KDD'03* (2003), p. 107–114.

[BGGB05a] J. Blanchard, F. Guillet, R. Gras & H. Briand – "Assessing rule interestingness with a probabilistic measure of deviation from equilibrium", *ASMDA'05, Proceedings of the 11th International Symposium on Applied Stochastic Models and Data Analysis* (2005), p. 191–200.

[BGGB05b] J. Blanchard, F. Guillet, R. Gras & H. Briand – "Using information-theoretic measures to assess association rule interestingness", *ICDM'05, Proceedings of the 5th IEEE International Conference on Data Mining* (2005), p. 66–73.

[BKGG03] J. Blanchard, P. Kuntz, F. Guillet & R. Gras – "Implication intensity: from the basic statistical definition to the entropic version (chapter 28)", *Statistical Data Mining and Knowledge Discovery* (2003), p. 475–493.

[Bla68] N. M. Blachman – "The amount of information that y gives about x", *IEEE Transactions on Information Theory* **14(1)** (1968), p. 27–31.

[Bla05] J. Blanchard – "A visualization system for interactive mining, assessment, and exploration of association rules (In French)", *Ph.D. Thesis, University of Nantes* (2005), p. 1–200.

[BMUT97] S. Brin, R. Motwani, J. D. Ullman & S. Tsur – "Dynamic itemset counting and implication rules for market basket data", *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data* (1997), p. 255–264.

[Bre96] L. Breiman – "Bagging predictors", *Machine Learning* **24** (1996), p. 123–140.

[Bre01] L. Breiman – "Random forests", *Machine Learning* **45** (2001), p. 5–32.

[BVV00] B. Baesens, S. Viaene & J. Vanthienen – "Post-processing of association rules", *SWP/KDD'00, Proceedings of The Special Workshop on Post-processing in conjunction with ACM KDD'00* (2000), p. 2–8.

[CAK05] D. H. Choi, B. S. Ahn & S. H. Kim – "Prioritization of association rules in data mining: Multiple criteria decision approach", *Expert Sytems with Applications* (2005), p. 1–12.

[CFE03] D. R. Carvalho, A. A. Freitas & N. F. F. Ebecken – "A critical review of rule surprisingness measures", *Proceedings of Data Mining IV - International Conference on Data Mining* (2003), p. 545–556.

[CFE05] D. R. Carvalho, A. A. Freitas & N. F. F. Ebecken – "Evaluating the correlation between objective rule interestingness measures and real human interest", *PKDD'05, the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases* **LNAI 3731** (2005), p. 453–461.

[CGL04] F. Coenen, G. Goulbourne & P. Leng – "Tree structures for mining association rules", *Data Mining and Knowledge Discovery* **8** (2004), p. 25–51.

[CHY96] M.-S. Chen, J. Han & P. S. Yu – "Data mining: an overview from database perspective", *IEEE Transactions on Knowledge and Data Engineering* **8(6)** (1996), p. 866–883.

[CP00] X. Chen & I. Petrounias – "An integrated query and mining system for temporal association rules", *Proceedings of the Second International Conference on Data Warehousing and Knowledge Discovery* **LNCS 1874** (2000), p. 327–336.

[CR06] A. Ceglar & J. F. Roddick – "Association mining", *ACM Computing Surveys* **38(2)** (2006), p. 1–42.

[CZ03] W. Cheung & O. R. Zaiane – "Incremental mining of frequent patterns without candidate generation or support constraint", *IDEAS'03, Proceedings of the Seventh International Database Engineering and Applications Symposium* (2003), p. 111–116.

[DHS01] R. O. Duda, P. E. Hart & D. G. Stork – *Pattern classification*, John Wiley & Sons, Inc, 2001.

[DL98] G. Dong & J. Li – "Interestingness of discovered association rules in terms of neighborhood-based unexpectedness", *PAKDD'98, Proceedings of the Second Pacific-Asia Conference on Research and Development in Knowledge Discovery and Data Mining* **LNCS 1394** (1998), p. 72–86.

[EHZ03] M. El-Hajj & O. R. Zaïane – "COFI-tree mining: a new approach to pattern growth with reduced candidacy generation", *FIMI'03, Workshop on Frequent Itemset Mining Implementations in conjunction with IEEE ICDM'03* (2003), p. 1–10.

[Fle96] L. Fleury – "Knowledge discovery in a human resource management database (In French)", *Ph.D. Thesis, University of Nantes* (1996).

[FPSM91] W. J. Frawley, G. Piatetsky-Shapiro & C. J. Matheus – "Knowledge discovery in databases: an overview", *Knowledge Discovery in Databases* (1991), p. 1–27.

[FPSS96] U. M. Fayyad, G. Piatetsky-Shapiro & P. Smyth – "From data mining to knowledge discovery", *Advances in Knowledge Discovery and Data Mining* (1996), p. 1–34.

[Fre99] A. A. Freitas – "On rule interestingness measures", *Knowledge-Based Systems Journal* **12(5-6)** (1999), p. 309–315.

[GAIM00] M. Gavrilov, D. Anguelov, P. Indyk & R. Motwani – "Mining the stock market: which measure is best?", *KDD'00, Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining* (2000), p. 487–496.

[GB98] P. Gago & C. Bentos – "A metric for selection of the most promising rules", *PKDD'98, Proceedings of the Second European Conference on the Principles of Data Mining and Knowledge Discovery* (1998), p. 19–27.

[GBPP96] R. Gras, H. Briand, P. Peter & J. Philippé – "Implicative statistical analysis", *IFCS'96, Proceedings of the Fifth Conference of the International Federation of Classification Societies* (1996), p. 412–419.

[GCB$^+$04] R. Gras, R. Couturier, J. Blanchard, H. Briand, P. Kuntz & P. Peter – "Quelques critères pour une mesure de qualité de règles d'association", *Mesures de Qualité pour la Fouille de Données* **RNTI-E-1** (2004), p. 3–31.

[GH06] L. Geng & H. J. Hamilton – "Interestingness measures for data mining: A survey", *ACM Computing Surveys* **38(3)** (2006), p. 1–32.

[GK06] R. Gras & P. Kuntz – "Discovering R-rules with a directed hierarchy", *Soft Computing - A Fusion of Foundations, Methodologies and Applications* **10(5)** (2006), p. 453–460.

[GSG99] G. K. Gupta, A. Strehl & J. Ghosh – "Distance based clustering of association rules", *ANNIE'99, Intelligent Engineering Systems Through Artificial Neural Networks* **9** (1999), p. 759–764.

[Gui02] S. Guillaume – "Discovery of ordinal association rules", *PAKDD'02, Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining* **LNCS 2336** (2002), p. 322–327.

[Gui04] F. Guillet – "Mesures de la qualité des connaissances en ECD", *EGC'04, Actes des tutoriels, 4ème Conférence francophone Extraction et Gestion des Connaissances, http://www.isima.fr/∼egc2004/* (2004), p. 1–60.

[HBV01] M. Halkidi, Y. Batistakis & M. Vazirgiannis – "On clustering validation techniques", *Journal of Intelligent Information Systems* **17(2-3)** (2001), p. 107–145.

[HF99] J. Han & Y. Fu – "Discovery of multiple-level association rules from large databases", *IEEE Transactions on Knowledge and Data Engineering* **11(5)** (1999), p. 798–805.

[HGB05a] H. X. Huynh, F. Guillet & H. Briand – "ARQAT: an exploratory analysis tool for interestingness measures", *ASMDA'05, Proceedings of the 11th International Symposium on Applied Stochastic Models and Data Analysis* (2005), p. 334–344.

[HGB05b] H. X. Huynh, F. Guillet & H. Briand – "Clustering interestingness measures with positive correlation", *ICEIS'05, Proceedings of the 7th International Conference on Enterprise Information Systems* **2** (2005), p. 248–253.

[HGB05c] H. X. Huynh, F. Guillet & H. Briand – "A data analysis approach for evaluating the behavior of interestingness measures", *DS'05, the 8th International Conference on Discovery Science* **LNAI 3735** (2005), p. 330–337.

[HGB$^+$06a] H. X. Huynh, F. Guillet, J. Blanchard, P. Kuntz, R. Gras & H. Briand – "A graph-based clustering approach to evaluate interestingness measures : a tool and a comparative study. (chapter 2)", *Quality measures in data mining* (2006).

[HGB06b] H. X. Huynh, F. Guillet & H. Briand – "Extracting representative measures for the post-processing of association rules", *RIVF'06, Proceedings of the 4th International IEEE Conference on Computer Sciences - Research & Innovation - Vision for the Future* (2006), p. 99–105.

[HGB06c] H. X. Huynh, F. Guillet & H. Briand – "A graph-based approach for comparing interestingness measures", *IEEE ICEIS'06, Proceedings of the First IEEE International Conference on Engineering of Intelligent Systems* (2006), p. 375–380.

[HGB06d] H. X. Huynh, F. Guillet & H. Briand – "Evaluating interestingness measures with linear correlation graph", *IEA-AIE'06, Proceedings of the 19th International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems* **LNCS 4031** (2006), p. 312–321.

[HGB06e] H. X. Huynh, F. Guillet & H. Briand – "Discovering the stable clusters between interestingness measures", *ICEIS'06, Proceedings of the 8th International Conference on Enterprise Information Systems* **2** (2006), p. 196–201.

[HH01] R. J. Hilderman & H. J. Hamilton – *Knowledge discovery and measures of interestingness*, Kluwer Academic Publishers, 2001.

[HLL00] F. Hussain, H. Liu & H. Lu – "Relative measure for mining interesting rules", *PKDD'00, Proceedings of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases* (2000), p. 117–132.

[HLSL00] F. Hussain, H. Liu, E. Suzuki & H. Lu – "Exception rule mining with a relative interestingness measure", *PAKDD'00, Knowledge Discovery and Data Mining, Current Issues and New Applications, 4th Pacific-Asia Conference* (2000), p. 86–97.

[HMS01] D. Hand, H. Mannila & P. Smyth – *Principles of data mining*, MIT Press, 2001.

[HPYM04] J. Han, J. Pei, Y. Yin & R. Mao – "Mining frequent patterns without candidate generation: a frequent-pattern tree approach", *Data Mining and Knowledge Discovery* **8** (2004), p. 53–87.

[JMF99] A. K. Jain, M. N. Murty & P. Flyn – "Data clustering: a review", *ACM Computing Surveys* **31(3)** (1999), p. 264–323.

[JS02] S. Jaroszewicz & D. A. Simovici – "Pruning redundant association rules using maximum entropy principle", *PAKDD'02, Proceedings of the Sixth Pacific-Asia Conference on Knowledge Discovery and Data Mining* **LNAI 2336** (2002), p. 135–147.

[JWTF05] Y. Jiang, K. Wang, A. Tuzhilin & A. W.-C. Fu – "Mining patterns that respond to actions", *ICDM'05, Proceedings of the Fifth IEEE International Conference on Data Mining* (2005), p. 669–672.

[KH95] K. Koperski & J. Han – "Discovery of spatial association rules in geographic information databases", *SASD'95, Proceedings of the 4th International Symposium on Advances in Spatial Databases* **LNCS 951** (1995), p. 47–66.

[KMR+94] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen & A. I. Verkamo – "Finding interesting rules from large sets of discovered association rules", *CIKM'94, Proceedings of the Third International Conference on Information and Knowledge Management* (1994), p. 401–407.

[Kod01] Y. Kodratoff – "Comparing machine learning and knowledge discovery in databases: An application to knowledge discovery in texts", *Machine Learning and Its Applications, Advanced Lectures* **LNCS 2049** (2001), p. 1–21.

[Koj04] I. Kojadinovic – "Unsupervised aggregation by the choquet integral based on entropy functionals: Application to the evaluation of students", *MDAI'04, Proceedings of the First International Conference on Modeling Decisions for Artificial Intelligence* **LNCS 3131** (2004), p. 163–175.

[Kon95] I. Kononenco – "On biases in estimating multi-valued attributes", *IJCAI'95, Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* (1995), p. 1034–1040.

[KR90] L. Kaufman & P. J. Rousseeuw – *Finding groups in data: an introduction to cluster analysis*, Wiley and Sons, 1990.

[KS96] M. Kamber & R. Shinghal – "Evaluating the interestingness of characteristic rules", *KDD'96, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (1996), p. 263–266.

[KZ96] W. Klösgen & J. M. Zytkow – "Knowledge discovery in databases terminology", *Advances in Knowledge Discovery and Data Mining* (1996), p. 573–592.

[KZ02] W. Klösgen & J. M. Zytkow – *Handbook of data mining and knowledge discovery*, Oxford University Press, 2002.

[LD98] E. Lehmann & H. D'Abrera – *Nonparametrics: Statistical methods based on rank*, Prentice-Hall, 1998.

[LGB04] R. Lehn, F. Guillet & H. Briand – "Qualité d'un ensemble de règles: élimination des règles redondantes", *Mesures de Qualité pour la Fouille de Données* **RNTI-E-1** (2004), p. 169–192.

[LHC97] B. Liu, W. Hsu & S. Chen – "Using general impressions to analyze discovered classification rules", *KDD'97, Proceeding of the Third International Conference on Knowledge Discovery and Data Mining* (1997), p. 31–36.

[LHF98] H. Lu, J. Han & L. Feng – "Stock movement prediction and n-dimensional intertransaction association rules", *ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery* (1998), p. 12:1–12:7.

[LHH00] B. Liu, M. Hu & W. Hsu – “Multi-level organization and summarization of the discovered rules”, *KDD'00, Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining* (2000), p. 208–217.

[LHM99] B. Liu, W. Hsu & Y. Ma – “Pruning and summarizing the discovered associations”, *KDD'99, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (1999), p. 125–134.

[LHM01] B. Liu, W. Hsu & Y. Ma – “Identifying non-actionable association rules”, *KDD'01, Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (2001), p. 329–334.

[LHML99] B. Liu, W. Hsu, L.-F. Mun & H.-Y. Lee – “Finding interesting patterns using user expectations”, *IEEE Transactions on Knowledge and Data Engineering* **11(6)** (1999), p. 817–832.

[LHWC99] B. Liu, W. Hsu, K. Wang & S. Chen – “Visually aided exploration of interesting association rules”, *PAKDD'99, Proceedings of the Pacific Asia Conference on Knowledge Discovery and Data Mining* **LNCS 1574** (1999), p. 380–389.

[LLV05] S. Lallich, P. Lenca & B. Vaillant – “Variations autour de l'intensité d'implication”, *ASI'05, Proceedings of the third International Conference Implicative Statistic Analysis* (2005), p. 237–246.

[LMF82] L. Lebart, A. Morineau & J.-P. Fénelon – *Traitement des données statistiques : méthodes et programmes*, Dunod, 1982.

[LMP+04] P. Lenca, P. Meyer, P. Picouet, B. Vaillant & S. Lallich – “Evaluation et analyse multi-critères des mesures de qualité des règles d'association”, *Mesures de Qualité pour la Fouille de Données* **RNTI-E-1** (2004), p. 219–246.

[Loe47] J. Loevinger – “A systematic approach to the construction and evaluation of tests of ability”, *Psychological Monographs* **61** (1947), p. 1–49.

[LT04] S. Lallich & O. Teytaud – “Evaluation et validation de l'intérêt des règles d'association”, *Mesures de Qualité pour la Fouille de Données* **RNTI-E-1** (2004), p. 193–217.

[LVL05] S. Lallich, B. Vaillant & P. Lenca – “Parametrised measures for the evaluation of association rule interestingnes”, *ASMDA'05, Proceedings of the 11th International Symposium on Applied Stochastic Models and Data Analysis* (2005), p. 220–229.

[Mar00] J.-L. Marichal – “On sugeno integral as an aggregation function”, *Fuzzy Sets and Systems* **114(3)** (2000), p. 347–365.

[McG05] K. McGarry – “A survey of interestingness measures for knowledge discovery”, *Knowledge Engineering Review Journal* **20(1)** (2005), p. 39–61.

[MCPS93] C. J. Matheus, P. K. Chan & G. Piatetsky-Shapiro – “Systems for knowledge discovery in databases”, *IEEE Transactions on Knowledge Discovery and Data Engineering (special issue on Learning & Discovery in Knowledge-Based Databases)* (1993), p. 1–21.

[McQ67] J. McQueen – "Some methods for classification and analysis of multivariate observations", *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* **1** (1967), p. 281–297.

[MM95] J. A. Major & J. J. Magano – "Selecting among rules induced from a hurricane database", *Journal of Intelligent Information Systems* **4(1)** (1995), p. 39–52.

[MT96] H. Mannila & H. Toivonen – "On an algorithm for finding all interesting sentences", *EMCSR'96, Proceedings of the 13th European Meeting on Cybernetics and Systems* (1996), p. 973–978.

[MY97] R. Miller & Y. Yang – "Association rules over interval data", *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data* (1997), p. 452–461.

[NHBM98] D. J. Newman, S. Hettich, C. L. Blake & C. J. Merz – "[UCI] Repository of machine learning databases, http://www.ics.uci.edu/∼mlearn/mlrepository.html", *University of California, Irvine, Department of Information and Computer Sciences* (1998).

[NLHP98] R. Ng, L. Lakshmanan, J. Han & A. Pang – "Exploratory mining and pruning optimazations of constrained association rules", *Proceedings of the ACM SIGMOD International Conference on Management of Data* (1998), p. 13–24.

[NS05] R. Natarajan & B. Shekar – "Interestingness of association rules in data mining: issues relevant to e-commerce", *SADHANA* **30(2-3)** (2005), p. 291–309.

[Omi03] A. R. Omiecinski – "Alternative interest measures for mining associations in databases", *IEEE Transactions on Knowledge and Data Engineering* **15(1)** (2003), p. 57–69.

[ORS98] B. Ozden, S. Ramaswamy & A. Silberschatz – "Cyclic association rules", *ICDE'98, Proceedings of the Fourteenth International Conference on Data Engineering* (1998), p. 412–421.

[PBTL99] N. Pasquier, Y. Bastide, R. Taouil & L. Lakhal – "Discovering frequent closed itemsets for association rules", *Proceeding of the 7th International Conference on Database Theory (ICDT'99)* **LNCS 1540** (1999), p. 398–416.

[PCY95] J. S. Park, M.-S. Chen & P. S. Yu – "Efficient parallel data mining for association rules", *ICKM'95, Proceedings of the Fourth International Conference on Information and Knowledge Management* (1995), p. 31–36.

[PS91] G. Piatetsky-Shapiro – "Discovery, analysis, and presentation of strong rules", *Knowledge Discovery in Databases. In G. Piatesky-Shapiro and W. Frawley (editors)* (1991), p. 229–248.

[PSM94] G. Piatetsky-Shapiro & C. J. Matheus – "The interestingness of deviations", *AAAI'94, Knowledge Discovery in Databases Workshop* (1994), p. 25–36.

[PSS00] G. Piatetsky-Shapiro & S. Steingold – "Measuring lift quality in database marketing", *ACM SIGKDD Explorations Newsletter* **2(2)** (2000), p. 76–80.

[PT98] B. Padmanabhan & A. Tuzhilin – "A belief-driven method for discovering unexpected patterns", *KDD'98, Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* (1998), p. 94–100.

[PT99] B. Padmanabhan & A. Tuzhilin – "Unexpectedness as a measure of interestingness in knowledge discovery", *Decision Support Systems* **27(3)** (1999), p. 303–318.

[PT00] B. Padmanabhan & A. Tuzhilin – "Small is beautiful: discovering the minimal set of unexpected patterns", *KDD'00, Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining* (2000), p. 54–63.

[PT06] B. Padmanabhan & A. Tuzhilin – "On characterization and discovery of minimal unexpected patterns in rule discovery", *IEEE Transactions on Knowledge and Data Engineering* **18(2)** (2006), p. 202–216.

[Rat97] A. Ratnaparkhi – "A simple introduction to maximum entropy models for natural language processing", *IRCS Report 97-08, University of Pennsylvania* (1997), p. 1–13.

[RMS98] S. Ramaswamy, S. Mahajan & A. Silberschatz – "On the discovery of interesting patterns in association rules", *VLDB'98, Proceedings of the 24rd International Conference on Very Large Data Bases* (1998), p. 368–379.

[Ros87] S. M. Ross – *Introduction to probability and statistics for engineers and scientists*, Wiley, 1987.

[Rym92] R. Rymon – "Search through systematic set enumeration", *Proceedings of the Third International Conference Principles of Knowledge Representation and Reasoning* (1992), p. 539–550.

[RZ06] G. Ritschard & D. A. Zighed – "Implication strength of classification rules", *ISMIS'06, Proceedings of the 16th International Symposium on Methodologies for Intelligent Systems* **LNAI 4203** (2006), p. 463–472.

[SA95] R. Srikant & R. Agrawal – "Mining generalized association rules", *VLDB'95, Proceedings of the 21st International Conference on Very Large Databases* (1995), p. 407–419.

[SA96] R. Srikant & R. Agrawal – "Mining quantitative association rules in large relational tables", *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data* (1996), p. 1–12.

[Sah99] S. Sahar – "Interestingness via what is not interesting", *KDD'99, Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (1999), p. 332–336.

[Sah01] S. Sahar – "Interestingness preprocessing", *ICDM'01, Proceedings of the First IEEE International Conference on Data Mining* (2001), p. 489–496.

[Sah02a] S. Sahar – "Exploring interestingness through clustering: a framework", *ICDM'02, Proceedings of the Second IEEE International Conference on Data Mining* (2002), p. 677–680.

[Sah02b] S. Sahar – "On incorporating subjective interestingness into the mining process", *ICDM'02, Proceedings of the Second IEEE International Conference on Data Mining* (2002), p. 681–684.

[Sap90] G. Saporta – *Probabilité, analyse des données et statistiques*, Edition Technip, 1990.

[SM99] S. Sahar & Y. Mansour – "Empirical evaluation of interest-level evaluation", *SPIE'99, Proceedings of SPIE - DMKD: Theory, Tools and Technology* **3695** (1999), p. 63–74.

[SON95] A. Savasere, E. Omiecinski & S. Navathe – "An efficient algorithm for mining association rules in large databases", *VLDB'95, Proceedings of the 21th International Conference on Very Large Databases* (1995), p. 432–444.

[SON98] A. Savasere, E. Omiecinski & S. B. Navathe – "Mining for strong negative associations in a large database of customer transactions", *ICDE'98, Proceedings of the Fourteenth International Conference on Data Engineering* (1998), p. 494–502.

[SS88] M. Sebag & M. Schoenauer – "Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases", *EKAW'88, Proceedings of the European Knowledge Acquisition Workshop. Gesellschaft für Mathematik und Datenverarbeitung mbH* (1988), p. 28.1–28.20.

[ST95] A. Silberschatz & A. Tuzhilin – "On subjective measures of interestingness in knowledge discovery", *KDD'95, Proceedings of the First International Conference on Knowledge Discovery and Data Mining* (1995), p. 275–281.

[ST96] A. Silberschatz & A. Tuzhilin – "What makes patterns interesting in knowledge discovery systems", *IEEE Transactions on Knowledge and Data Engineering* **5(6)** (1996), p. 970–974.

[Sub98] R. Subramonian – "Defining diff as a data mining primitive", *KDD'98, Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* (1998), p. 334–338.

[Suz97] E. Suzuki – "Autonomous discovery of reliable exception rules", *KDD'97, Proceedings of the Third International Conference on Knowledge Discovery and Data Mining* (1997), p. 259–262.

[SVA97] R. Srikant, Q. Vu & R. Agrawal – "Mining association rules with item constraints", *KDD'97, Proceedings of the Third International Conference on Knowledge Discovery and Data Mining* (1997), p. 67–73.

[THC02] W.-G. Teng, M.-J. Hsieh & M.-S. Chen – "On the mining of substitution rules for statistically dependent items", *ICDM'02, Proceedings of the 2002 IEEE International Conference on Data Mining* (2002), p. 442–449.

[THC05] W.-G. Teng, M.-J. Hsieh & M.-S. Chen – "A statistical framework for mining substitution rules", *Knowledge and Information Systems* **7(2)** (2005), p. 158–178.

[TKR$^+$95] H. Toivonen, M. Klemettinen, P. Ronkainen, K. Hätönen & H. Mannila – "Pruning and grouping discovered association rules", *Proceedings of MLnet Workshop on Statistics, Machine Learning and Discovery in Databases* (1995), p. 47–52.

[TKS02] P.-N. Tan, V. Kumar & J. Srivastava – "Selecting the right interestingness measure for association patterns", *KDD'02, Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (2002), p. 32–41.

[TKS04] P.-N. Tan, V. Kumar & J. Srivastava – "Selecting the right objective measure for association analysis", *Information Systems* **29(4)** (2004), p. 293–313.

[Toi96] H. Toivonen – "Sampling large databases for finding association rules", *VLDB'96, Proceedings of the 22nd International Conference on Very Large Databases* (1996), p. 134–145.

[TS06] R. Tamir & Y. Singer – "On a confidence gain measure for association rule discovery and scoring", *The International Journal on Very Large Data Bases* **15(1)** (2006), p. 40–52.

[TSK06] P.-N. Tan, M. Steinbach & V. Kumar – *Introduction to data mining*, Pearson/Addison Wesley, 2006.

[VLL04] B. Vaillant, P. Lenca & S. Lallich – "A clustering of interestingness measures", *DS'04, the 7th International Conference on Discovery Science* **LNAI 3245** (2004), p. 290–297.

[VPL03] B. Vaillant, P. Picouet & P. Lenca – "An extensible platform for rule quality measure benchmarking", *HCP'03, Human Centered Processes* (2003), p. 187–191.

[vR79] C. J. van Rijsbergen – *Information retrieval (2nd edition)*, Butterworths, 1979.

[WHP06] J. Wang, J. Han & J. Pei – "Closed constrained gradient mining in retail databases", *IEEE Transactions on Knowledge and Data Engineering* **18(6)** (2006), p. 764–769.

[WTL98] K. Wang, S. H. W. Tay & B. Liu – "Interestingness-based interval merger for numeric association rules", *KDD'98, Proceedings of the 4th International Conference Knowledge Discovery and Data Mining* (1998), p. 121–128.

[Zak99] M. J. Zaki – "Parallel and distributed association mining: A survey", *IEEE Concurrency* (1999), p. 14–25.

[ZB03] Q. Zhao & S. S. Bhowmick – "Association rule mining: A survey", *Technical Report, Center for Advanced Information Systems, School of Computer Engineering, Nanyang Technological University* (2003), p. 1–20.

[ZHZ00] O. Zaiane, J. Han & H. Zhu – "Mining recurrent items in multimedia with progressive resolution refinement", *ICDE'00, Proceedings of the 16th International Conference on Data Engineering* (2000), p. 461–470.

Résumé :

Ce travail s'insère dans le cadre de l'extraction de connaissances dans les données (ECD), souvent dénommé "fouille de données". Ce domaine de recherche multidisciplinaire offre également de nombreuses applications en entreprises. L'ECD s'attache à la découverte de connaissances cachées au sein de grandes masses de données. Parmi les modèles d'extraction de connaissances disponibles, celui des règles d'association est fréquemment utilisé. Il offre l'avantage de permettre une découverte non supervisée de tendances implicatives dans les données, mais, en retour, délivre malheureusement de grandes quantités de règles. Son usage nécessite donc la mise en place d'une phase de post-traitement pour aide l'utilisateur final, un décideur expert des données, à réduire la masse de règles produites. Une manière de réduire la quantité de règles consiste à utiliser des indicateurs numériques de la qualité des règles, appelés "mesures d'intérêts". La littérature propose de nombreuses mesures de ce type, et étudie leurs propriétés.

Cette thèse se propose d'étudier la panoplie de mesures d'intérêts disponibles, afin d'évaluer leur comportement en fonction d'une part de la nature des données et d'autre part des préférences du décideur. L'objectif final étant de guider le choix de l'utilisateur vers les mesures les mieux adaptées à ses besoins et in fine de sélectionner les meilleures règles.

A cette fin, nous proposons une nouvelle approche implémentée dans un nouvel outil, ARQAT (Association Rule Quality Analysis Tool), afin de faciliter l'analyse du comportement des 40 mesures d'intérêt recensées. En plus de statistiques élémentaires, l'outil permet une analyse poussée des corrélations entre mesures à l'aide de graphes de corrélation s'appuyant sur les coefficients proposés par Pearson, Spearman et Kendall. Ces graphes sont également utilisés pour l'identification de clusters de mesures similaires.

En outre, nous avons proposé une série d'études comparatives sur les corrélations entre les mesures d'intérêt sur plusieurs jeux de données. A l'issue de ces études, nous avons découvert un ensemble de corrélations peu sensibles à la nature des données utilisées, et que nous avons appelées corrélations stables.

Enfin, 14 graphiques et vues complémentaires structurées à 5 niveaux d'analyse : l'analyse de jeu de règles, l'analyse de corrélation et de clustering, l'analyse des meilleures règles, l'analyse de sensibilité, et l'analyse comparative sont illustrées afin de montrer l'intérêt de l'approche exploratoire et de l'utilisation des vues complémentaires.

Mots-clés :
Extraction des Connaissances à partir de Données (ECD), mesures d'intérêt, post-traitement de règles d'association, clustering, graphe de corrélation, analyse de stabilité.

Abstract:

This work takes place in the framework of Knowledge Discovery in Databases (KDD), often called "Data Mining". This domain is both a main research topic and an application field in companies. KDD aims at discovering previously unknown and useful knowledge in large databases. In the last decade many researches have been published about association rules, which are frequently used in data mining. Association rules, which are implicative tendencies in data, have the advantage to be an unsupervised model. But, in counter part, they often deliver a large number of rules. As a consequence, a postprocessing task is required by the user to help him understand the results. One way to reduce the number of rules - to validate or to select the most interesting ones - is to use interestingness measures adapted to both his/her goals and the dataset studied. Selecting the right interestingness measures is an open problem in KDD. A lot of measures have been proposed to extract the knowledge from large databases and many authors have introduced the interestingness properties for selecting a suitable measure for a given application. Some measures are adequate for some applications but the others are not.

In our thesis, we propose to study the set of interestingness measure available in the literature, in order to evaluate their behavior according to the nature of data and the preferences of the user. The final objective is to guide the user's choice towards the measures best adapted to its needs and *in fine* to select the most interesting rules.

For this purpose, we propose a new approach implemented in a new tool, ARQAT (Association Rule Quality Analysis Tool), in order to facilitate the analysis of the behavior about 40 interestingness measures. In addition to elementary statistics, the tool allows a thorough analysis of the correlations between measures using correlation graphs based on the coefficients suggested by Pearson, Spearman and Kendall. These graphs are also used for identifying the clusters of similar measures.

Moreover, we proposed a series of comparative studies on the correlations between interestingness measures on several datasets. We discovered a set of correlations not very sensitive to the nature of the data used, and which we called stable correlations.

Finally, 14 graphical and complementary views structured on 5 levels of analysis: ruleset analysis, correlation and clustering analysis, most interesting rules analysis, sensitivity analysis, and comparative analysis are illustrated in order to show the interest of both the exploratory approach and the use of complementary views.

Keywords:
Knowledge Discovery in Databases (KDD), interestingness measures, postprocessing of association rules, clustering, correlation graph, stability analysis.