Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 96 (2016) 345 – 354

Procedia Computer Science

20th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2016, 5-7 September 2016, York, United Kingdom

Ontology Knowledge mining based Association Rules Ranking

Rihab Idoudi[a,b*] , Karim Saheb Ettabaa [b], Basel Solaiman [b],Kamel Hamrouni [a]

[a] Université Tunis ElManar, Ecole Nationale d'Ingénieurs de Tunis, Tunis, 1200, Tunisia
[b] ITI Laboratoire, Telecom Bretagne, Brest, 29238, France

Abstract:

Medical association rules induction is used to discover useful correlations between pertinent concepts from large medical databases. Nevertheless, ARs algorithms produce huge amount of delivered rules and do not guarantee the usefulness and interestingness of the generated knowledge. To overcome this drawback, we propose an ontology based interestingness measure for ARs ranking. According to domain expert, the goal of the use of ARs is to discover implicit relationships between items of different categories such as 'clinical features and disorders','clinical features and radiological observations', etc. That's to say, the itemsets which are composed of "similar" items are uninteresting. Therefore, the dissimilarity between the rule's items can be used to judge the interestingness of association rules; the more different are the items, the more interesting the rule is. In this paper, we design a distinct approach for ranking semantically interesting association rules involving the use of an ontology knowledge mining approach. The basic idea is to organize the ontology's concepts into a hierarchical structure of conceptual clusters of targeted subjects, where each cluster encapsulates "similar" concepts suggesting a specific category of the domain knowledge. The interestingness of association rules is, then, defined as the dissimilarity between corresponding clusters. That's to say, the further are the clusters of the items in the AR, the more interesting the rule is. We apply the method in our domain of interest – mammographic domain- using an existing mammographic ontology called Mammo*, with the goal of deriving interesting rules from past experiences, to discover implicit relationships between concepts modeling the domain.

Peer-review under responsibility of KES International

Keywords: Association Rule, ontology knowledge mining, conceptual clusters, interestingness measures, Ranking

* http://sourceforge.net/p/gimimammography/code/HEAD/tree/trunk/owl

Peer-review under responsibility of KES International
doi:10.1016/j.procs.2016.08.147

1. Introduction

Association rule mining is, actually, one of the most important tasks in knowledge extraction in databases [1]. This technique aims to discover implicit correlations between items in databases that can be of great interest to domain experts. A typical and widely used application of AR mining is the medical domain. In fact, the exponential increase of the volume as well as the complexity of the radiological data raises crucial needs, for data management and knowledge discovery. However, the large number of extracted association rules makes the task of processing and interpreting them, very complicated for domain experts. Addressing this issue, researchers have proposed to rank generated rules according to their potential interest and enables highly ranked rules to be straightaway presented to decision makers. Methods for interestingness rules examination can be divided into objective and subjective methods. In subjective analysis, the ARs evaluation is based on prior domain knowledge that can be modeled within different supports (such as Rule schemas, ontologies, etc.). While, the objective methods are based on the use of statistical information in the database. Although the latter method can bring useful information regarding the dataset structure such as the support and the confidence measures, the rule interestingness strongly depends on the domain knowledge. Indeed, the more the knowledge is represented in an expressive and formal way, the more the rule evaluation is efficient. To this end, some approaches have proposed to involve ontologies in the post-processing task, to model domain knowledge since they provide a semantic support for vocabularies structuring. Recently, in medical domains, ontologies have become the cornerstone in knowledge acquisition and formalizing [2]. For example, in the mammographic domain, the mammographic ontology might be used to describe: the radiological observations associated with feature descriptors, mammogram Bi-Rads classification, clinical observation, etc. Surprisingly, while intense researches exist on applying association rule to mine ontologies, few approaches have, so far, exploited the benefits brought by these ontologies to compute rule interestingness measures [2]. Such measure represents a function that takes ontology concepts as input and returns a numerical value reflecting the dissimilarity between these concepts. Generally, this conceptual dissimilarity (distance) is based on the use of path based measure. That's to say, the furthest the concepts are in the ontological hierarchy, the more interesting the rule is [3] [2].

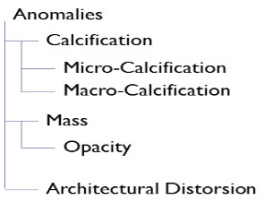

Figure 1: Extract of mammographic ontology

For example, the distance between the ontological concepts 'Calcification' and 'Micro-calcification' would be interpreted differently from the distance between 'Calcification' and 'Opacity' (see Figure 1), because of their different localization regarding the root. Moreover, the computed distance is strongly limited to the subjective construction of the ontology. Basing on multiple interviews, domain expert have revealed that these concepts like any two other entities regarding the anomalies category (see Figure1) are to a certain extent semantically similar since they are a part of a single subject. In order to meet the constraints imposed by the domain experts, we propose in this paper to increase the abstraction level of the conceptual knowledge encoded in the ontology, through the ontology knowledge mining process. The proposed idea consists on mining a novel structure of conceptual clusters organized hierarchically [24]. Each cluster groups similar objects encoding a specific topic regarding the ontology knowledge. That's to say, we aim to investigate the mined hierarchical conceptual clusters to determine so-called 'semantic interestingness measures' by computing the dissimilarity between the items of a given rule. Replacing exact similarity measure between concepts with semantic similarity between their correspondent clusters enables novel ways of interpreting AR, and hence may lead to the right identification of the interestingness of the rules. Therefore, the more clusters are far away, the more dissimilar are their respective concepts. The use of efficient data mining methods and appropriate interestingness metrics enables the identification of high quality relationships [3].

As already mentioned, our work focuses on mammographic domain. The overall approach starts, first by extracting ARs from the database. The latter comprises transactions of patients' medical records (previously diagnosed). After mining a hierarchy of conceptual clusters basing on mammography ontology, the generated ARs are ranked by measuring the interestingness rate based on a computed dissimilarity over the rule's items. The output results are straightaway presented to decision makers to be validated. The mammographic ontology used in this paper is the 'Mammo' ontology which is open source and available in the net. This paper is organized as following: In the first section, we advance the state of the art by reconnoitering the application of interestingness measures in the context of association rule post processing (ranking). Second, we introduce the general definitions of basic notions used in this manuscript. Thirdly, we introduce our proposed method which consists first on mining the ontology knowledge, then ranking the derived ARs to be validated by domain experts.

2. Related Work

2.1. *Objective and subjective methods for ARs processing*

Association Rules, ARs, reveal the relationships among items in a transaction of a database. However, the large number of generated ARs makes it difficult for decision makers to process, and interpret their utility. To tackle this limitation, several studies have been proposed to process generated rules using objective as subjective methods. In [3], Shaharanee et al. proposed the application of objective analysis to assess the generated rules. The approach consists in combining data mining and statistical measurement techniques (such as redundancy analysis, sampling and multivariate statistical analysis) to discard the insignificant rules. Alcala-Fdez et al. proposed a method based on rule covers to prune ARs [4]. The method defines subsets of rules describing the same transaction row. Then the rule set is reduced to its rule cover. In [5], D. Franke introduced the notion of subsumed rules which consist in a set of rules having the same conclusion part and several additional conditions in the condition part. A similar approach was introduced in [6] where the authors proposed to combine two different algorithms of data mining applied on the same data set. By comparing condition and conclusion of both kind extracted rules, these rules are then categorized into robust, consistent, and noteworthy rules. Other methods have proposed a rule-like formalism to model the user's expectations such as in [7]. Discovered rules are pruned/ filtered by comparing them to the user's expectations. In order to prune non-pertinent ARs, Concaro et al. have proposed in [8] a novel measure, called the Minimum Improvement measure. This measure describes the difference between the confidences of two rules. In fact, the rule can be pruned when the measure value is less than a fixed threshold. In [9], the authors have proposed an iterative rule validation system based on several operators, including rule grouping, filtering, browsing, and redundant rule elimination. An original method was proposed in [10] to prune and organize rules with the same consequent. First, the algorithm transforms the database in an ARs base, and then meta-rules r1 → r2 are extracted. These latter express the relations between two ARs and allow pruning/grouping of the discovered rules. Other researchers have been interested on the use of ontologies [11] [12] [13] [14] [15] in first steps of data mining process. The first application of domain ontologies was introduced in [16] with the concept of Generalized ARs. In order to generalize/specify rules, the authors proposed taxonomies of mined data (is-a hierarchy). In [17], the data set is first preprocessed according to the constraints extracted from the ontology and then, the rules generation step takes place. The main difference with our approach is that the role of ontology is integrated in the preprocessing task [17], whereas, in this paper, we are mainly interested in the post processing task.

2.2. Ontologies in ARs post processing

Another set of existing methods applying ontologies in ARs post processing task have been proposed. In [18], the authors have focused on ontology-based ARs post processing to improve the integration of user's knowledge. This method is based on the use of a rule schema reflecting the user's expectation and an ontology involving concept constraints. The ARs evaluation is carried over the defined rule schemas in order to prune and filter rules.

For medical ARs filtering, authors in [19] proposed a hybrid pruning method involving the use of both objective and subjective analysis, with the latter involving the use of an ontology. The proposed method was applied using general medical-domain ontology constructed using the Unified Medical Language System [20] with the goal of pruning already known rules. In [21], the authors used ARs to point out dependence relationships between Gene Ontology terms [25] using an annotation dataset and background knowledge. In [22], authors have proposed an additional gene ontology layer via discovering cross-ontology association rules from GO annotations. In [29], authors have proposed to group AR based on whether the rule items share relationships within a domain ontology. To this end, the method, uses vector space modeling of rule elements and an ontology based semantic similarity measure. The latter is based on measuring the depth of the least common ancestor node of two concepts in the ontology. In the perspective of computing ARs interestingness using domain ontology, the approach in [23] consists on calculating the conceptual distance by computing the number of edges of the shortest path between two concepts. The shorter the path is (from one concept to the other), the more similar the concepts are. In [2], the authors proposed to combine concept similarity metrics, formulated using the domain ontology with traditional interestingness measures (support and confidence). The domain specific semantic similarity between two items i_1 and i_2 is defined as well:

$$Sim(i_1, i_2) = \frac{Dist(LCA(i_1, i_2), Root)}{Dist(i_1, i_2) + Dist(LCA(i_1, i_2), Root)} \quad (1)$$

Where; LCA (i_1 , i_2) is the lowest common ancestor of the concepts i_1 and i_2, Dist(LCA(i_1 , i_2), Root) is the length of path from LCA(i_1 , i_2) to the root and Dist(i_1 , i_2) is a distance measure between i1 and i2.

3. Basic notions

3.1. Association rule mining

In data mining process, ARs are used for discovering important relations between items in a database D, where $D = \{t_1, t_2 \dots t_m\}$ is a set of transactions over $I = \{i, i_2 \dots i_n\}$ which is a set of items. A non-empty subset of I, $X = \{i_1, i_2 \dots i_k\}$ is called an itemset. Each transaction t_i in D is defined as an itemset $i_1, i_2 \dots i_k$ of length k.
An AR is an implication between two itemsets X and Y, in the form of $X \rightarrow Y$ where $X \cap Y = \emptyset$, In other words X and Y have no items in common. X is called the antecedent and Y is the conclusion, or the consequent, of the association rule. Each AR $R: X \rightarrow Y$ may be characterized by two measures Support and confidence. They are used for selecting ARs according to their potential interest to the user:

- The Support (sup):the occurrences of a specific event containing $X \cup Y$, $\sup(X \rightarrow Y) = s$;
- The Confidence (conf) is defined as:$\text{conf}(X \rightarrow Y) = \sup(X \rightarrow Y) / \sup(X) = \sup(X \cup Y) / \sup(X) = c$.

The algorithm Apriori [18] is the most widely used algorithm to discover ARs in databases. It gets as input the database to be mined and two thresholds *minConf* and *minSup* which represent respectively the minimum values of confidence and support that an AR must hold. The algorithm consists of two main steps: First all the item sets where the support is greater than *minSup* are generated; second, rules with support and confidence greater than *minConf* and *minSup* are extracted from the item sets (generated in the first step).

3.2. Ontology Definition

An ontology is defined as a formal specification of a shared conceptualization [26]. Ontologies are used to capture and formalize knowledge by modeling concepts and relations associated to the domain of interest. Concepts represent the pertinent entities[28] for example, *mammogram* is a concept within the mammographic domain, whereas relations designate the interactions between revealed concepts, for example, mammogram *classified_as* Bi-Rads3(Breast Imaging Reporting and Data System), that's to say, the concepts *mammogram* and *Bi-Rads3* are related via the relationship *classified_a.* Relations in the ontology are categorized into: (i)*Subsumption*: used to define the taxonomy which refers to the hierarchical concept tree, for example Opacity is a type of anomalies, (ii)*Associative relations*: relate the different concepts of the hierarchy (e.g. *classified_as*). Ontologies have been

widely used in medical domains to capture knowledge and formalize medical lexicons. In the mammographic domain, several ontologies have been proposed such as BCGO[†], Mammo[‡], Radlex[§]. Each ontology has been developed by different communities and for specific intended task. In this paper, we propose to use the Mammo ontology since it is considered as well structured and rich of vocabularies (concepts and relationships) that are relevant to the mammographic domain such as anomalies description, diagnosis and mammogram classifications.

4. Ontology Knowledge mining

In this section, we propose a new approach for ontology knowledge mining which aims at increasing the abstraction level of the conceptual knowledge encoded in the ontology. Our method consists on extracting a hierarchy of groups of correlated concepts reflecting the different knowledge granularity levels. Thus, the k-medoid clustering algorithm is applied iteratively (following a top-down strategy). In the following, we define first the k-medoid algorithm, second we propose a novel conceptual distance based on the concept's context. Finally, we describe the hierarchical clustering algorithm.

4.1. K-medoid for ontology's concepts clustering

The conceptual clustering procedure implemented in our method is based on the use of the k-medoid algorithm [27]. The latter is applied over the ontology's concepts to create k flat clusters. The produced clusters are introduced with robust representative data called medoids; they represent the cluster's centers. The latter can help significantly the domain expert to visualize the extracted knowledge topics in the generated hierarchy of clusters. The cluster's medoid represents the concept with the lowest average distance (the used distance is introduced in) with respect to the other concepts in the cluster. The algorithm gets as input parameters k: the number of clusters, the set of medoids (initialized randomly). Then, two main steps are performed iteartively; first, the algorithm computes the distances between concepts and the current medoids to assign each concept to the closest one of the k clusters. Second, we update the medoids set according to the new repartition. The algorithm converges when a maximum number of iteration have been achieved or when the set of medoids become stable. The medoid of a cluster $\mathrm{C} = \{\mathrm{x}_1, \mathrm{x}_2, \ldots \mathrm{x}_\mathrm{n}\}$ is defined as well: $v = \operatorname{argmin}_{x \epsilon C} \sum_{j=1}^{n} d(x, x_j)$ (2)

4.2. Semantic conceptual distance

We propose to use, in this section, a distance measure based on the concept's context. This measure has been proposed in previous work [30]. Semantically, similar entities should have in common concepts to which are related. On the ground of such an intuition, we introduce the notion of context as the set of concepts to which the concept of interest is related through both type of relations: *Subsumption* and *Associative*. The context $Cont(c)$ of a given concept c is defined as: $Cont(c) = \{c_i | (c, c_i) \in R \cup \{c\}\}$ .Where R is the set of relations including the subsumption and associative relations. The rationale of the new measure is to compare the concepts on the grounds of their contexts which stand as a group features. Based on the notion of the concept's context, we have proposed a semantic distance measure which is based on the idea of comparing their semantics along their contexts. This measure can be defined as follows: Given two concepts $\mathrm{c_i}$ and $\mathrm{c_j}$, the distance $\mathrm{d}(\mathrm{c_i}, \mathrm{c_j})$ based on the relational context is given as well:

$$\mathrm{d}(\mathrm{c_i}, \mathrm{c_j}) = 1 - \left(2.\frac{|\mathrm{Cont}(\mathrm{c_i}) \cap \mathrm{Cont}(\mathrm{c_j})|}{|\mathrm{Cont}(\mathrm{c_i})| + |\mathrm{Cont}(\mathrm{c_j})|}\right) \quad (3)$$

4.3. Hierarchical divisive clustering algorithm

The process starts with one cluster grouping the ontology's concepts, and then it repeatedly breaks clusters into more specific and smaller sub-clusters until stopping criterion is satisfied.

[†] https://bioportal.bioontology.org/ontologies/BCGO
[‡] http://sourceforge.net/p/gimimammography/code/HEAD/tree/trunk/owl
[§] http://bioportal.bioontology.org/ontologies/RADLEX

This algorithm can be thought as producing a dendrogram reflecting different levels of abstraction of the encoded knowledge. In particular, in the first level of the hierarchy, the number of desired cluster is selected by the user. This number depends, generally, on the domain of interest. Through this level, we aim to produce clusters as general as possible, which reflect the user's interests or/and the domain's main categories. For example, the mammographic domain is characterized by four main categories: 'Anatomical_entities', 'Conceptual_entities', 'Descriptors', and 'Diagnosis'. Those sorts represent as well the top-level classes in the mammographic ontologies ('BCGO', 'Mammo'). Once the clusters of the first level are produced, the iterative clustering algorithm is launched; each candidate cluster is verified if it can be further split according to its cohesiveness measure. The non-dense cluster is candidate to being divided into two-sub-clusters. The cohesiveness of the cluster is computed in order to verify the cluster's density. Our objective is to maximize the similarity of the concepts into a cluster to the medoid concept. The cohesiveness of a cluster C_i is computed as well:

$$Conhesiveness(C_i) = \sum_{i=1}^{K} \sum_{x \epsilon C_i} d(x - v_i)^2 \qquad (3)$$

Where v_i is the medoids of C_i; d() is the conceptual distance. That's to say, if the cluster's density is greater than a predefined threshold, the cluster is partitioned and the partition is constructed around two medoids chosen as the most dissimilar elements in the cluster and then iteratively adjusted in the inner loop (as described in algorithm 1). The advantage of our method is that it allows automatically determining the optimal number of main clusters. In the structure of the generated hierarchy, each node designates a cluster introduced with a medoid which characterizes the concepts in the given cluster while discriminating those in the twin cluster at the same level. A cluster of the hierarchical level L is presented by C_{iL}; v_{iL} designates the medoid of the ${C_{iL}}^m$ including m concepts: ${C_{iL}}^m = \{v_{iL}/\, x_k, k = 1, m\}$. At the end of this step, the items (concepts) in same cluster are similar to each other, while two items from two different clusters are dissimilar, and the dissimilarity between them can be judged with the dissimilarity between the two clusters.

5. Ontology Knowledge mining based ARs Ranking

5.1. Semantic dissimilarity measures

In our approach, we are making a distinction between the clusters to which belong the rule's items to measure the interestingness rate of the AR. In the following, we describe a semantic distance for the rules ranking. This distance measure is based on the hierarchical structure of the conceptual clusters resulted from the ontology knowledge mining process. That's to say, the more the clusters are distant, the more the rule is considered as interesting. If i_1 and i_2 are two items of an AR, we define the semantic similarity between them as:

$$SemDist(i_1, i_2) = 1 - SemSim(i_1, i_2) \ (4); \qquad SemSim(i_1, i_2) = \frac{2*D(LCC(C(i_1),C(i_2)),root)}{D(C(i_1),root)+D(C(i_2),root)} \ (5);$$

Where: $\mathrm{LCC}(i_1, i_2)$ is the lowest common cluster of (i_1, i_2); $\mathrm{C}(i_1)$ is the cluster to which i_1 belong; $\mathrm{D}(\mathrm{LCC}, \mathrm{root})$ is the length of path from LCC (i_1, i_2) to the root; $\mathrm{Dist}(\mathrm{C}(i_1), \mathrm{root})$ is the length of path from $\mathrm{C}(i_1)$ to the root; $\mathrm{Dist}(\mathrm{C}(i_2), \mathrm{root})$ is the length of path from $\mathrm{C}(i_2)$ to the root. This formula determines the semantic similarity of two items based on both the distance between their correspondent clusters and the location of their LCC in the structure. The value of the semantics similarity $\mathrm{SemSim}(i_1, i_2)$ ranges from 0 to 1 and so does the value of the Semantic distance. It can be observed that if two items belong to the same cluster the value of the SemSim is 1.

5.2. *R*ule *interestingness measures*

To determine the AR interestingness measure, we compute the distance between items of the given rule $\mathrm{R}: \mathrm{X} \rightarrow \mathrm{Y}$, where R is made of $\{a_1, a_2 .. a_n\}$ its interestingness is defined as following; $SemDist(a_i, a_j)$ is the semantic distance between each couple of items of the AR. Similar to the semantic distance, the value falls in [0,1]:

$$\mathrm{Interest(R)} = \frac{\sum_{1 \le i,j \le n;\ i \ne j} SemDist(a_i, a_j)}{\sum_{k=1}^{n-1} k} \qquad (6)$$

5.3. AR Ranking

Once the interestingness measure is computed for each rule in the AR set according to Equation (6), the rules with high interestingness can be output to the domain decision. An alternate way it to output the top-k rules with high interestingness and the domain expert can choose the number of interesting rules to be generated.

6. Results

6.1. Data set

Automated radiological systems are accumulating, daily, large quantities of information about patients and their medical conditions. Therefore, we propose to explore in this work, the medical data-repository which includes the patients' medical records previously diagnosed. The overall database includes 1000 patient records. Those data have been collected from the hospital Charles Nicolle in Tunisia. Each record encloses a textual description about, patient's information (such as the age, menopause, etc.), clinical observation (breast nature, skin change, etc.), radiological observation (tumor description) and the evaluation and/or mammogram classification (according to Bi-Rads classification). The main objective of the paper is to find out relations between those classes such as correlations between clinical features and disorders, clinical features and radiological observations, clinical features and mammogram classification, etc. In data processing, we have eliminated information that is not relevant to our context, and converted the past experiences into transactions with readable format (arff file). including 30 attributes values.

6.2. Mammo ontology knowledge mining

As we have mentioned, we have selected the mammographic ontology 'Mammo' to be used as a semantic support for ARs interestingness measure measuring. This is due to the fact that it is rich of vocabularies and it well-structured. This ontology has 692 concepts and 73 properties. To ensure the ontology knowledge mining, we have conceived an application that gets as input: the ontology to be mined, a reference file with information about concepts contexts, and the similarity matrix of concepts. The output is a hierarchical structure of conceptual clusters where the first level handles four general categories (fixed by the radiological experts) reflecting the main suitable seeds 'Anatomical_entities', 'Conceptual_entities', 'Descriptors', 'Diagnosis'.

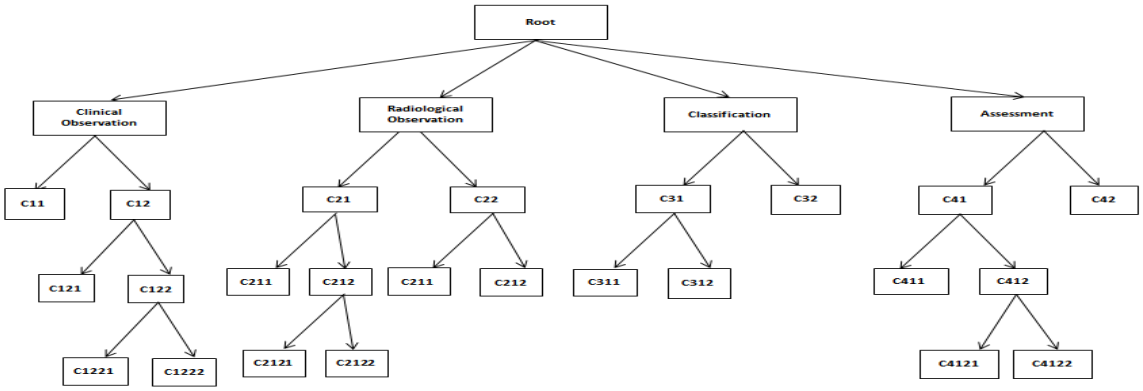

Figure 2:Mammo ontology knowledge mining

We have obtained 5 hierarchical levels: 4 clusters at the top level 16 clusters at the main level (at the most specific level). Note that clusters of higher levels are more general than the deeper ones. Figure 2 presents an extract of the generated hierarchical clusters. To simplify the visualization, each cluster is presented with its respective medoid (see Table 1). The leaf nodes (clusters) constitute the clusters to be used in computing the semantic distance between items (which are mapped to concepts in the ontology) of a rule.

Table 1. Medoids' codes

Code	Medoid	Code	Medoid
C11	Risk Factors	C212	Spiculate
C12	Anatomical Entity	C31	Histology
C121	Macro-Anatomy	C32	BiRads
C122	Micro-Anatomy	C311	Histopathological Grading
C21	Mass	C312	Histopathological Scoring
C22	Architectural Distorsion	C41	Diagnosis
C211	Mass density	C42	Recommendation
C212	Mass shape	C411	Benin Diagnosis
C2121	Mass Size	C412	Malign Diagnosis
C2122	Mass Margin	C4121	carcinoma
C211	Focal	C4122	Invasive_carcinoma

6.3. Evaluation of the clustering algorithm

We have assessed the hierarchy quality; whether sub-clusters of a given class in the hierarchy are well linked. Therefore, the contents of clusters at each level are compared with the content of corresponding reference clusters (with manual clustering). This evaluation is realized by the means of precision and recall metrics. Figure 3 shows the average clustering precision and recall of the proposed algorithm per level, it can be noticed that as clusters are becoming more specific, semantically related concepts remain clustered together.

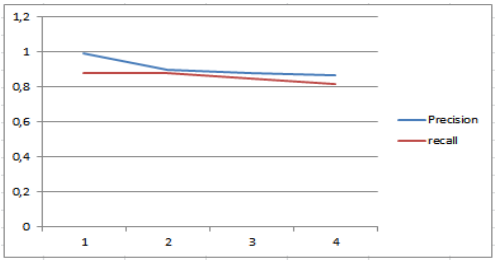

Figure 3: Precision and recall of each hierarchical level for the Mammo

6.4. ARs Extraction

The Apriori Algorithm is used here for AR Mining to extract rules that satisfy the predefined minimum support and confidence. We have chosen the Weka3.6.1 [**] software which is an open source data mining software. In order to observe the influence of support and confidence thresholds on the generated rules, we experimented with different parameter thresholds settings. Table 2 shows the numbers of generated rules with variation of MinSup and Minconf.

Table 2. Numbers of rules generated for various minimum supports

MinSup	0.7	0.3	0.1	0.7	0.3	0.1	0.1	0.09	0.07	0. 05
MinConf	1	1	1	0.9	0.9	0.9	0.8	0.7	0.7	0.7
RA	0	0	74	2	23	536	1177	2167	3395	6552

It can be seen that high minConf and minSup values lead to fewer but more robust rules, i.e. rules with a high conditional probability, antecedent and consequent being almost always correlated, while decreasing the minimum support as well as confidence values, can conspicuously increase the number of rules. As a matter of fact, if the parameters of interest, i.e. the threshold values, are fixed too high, then too few rules are generated with omitting useful information. Otherwise, if the parameters are fixed too low, then, the algorithms can generate an extremely large amount of rules with unsuitable or uninteresting knowledge. For finding a good compromise between number of rules and robustness, we have empirically chosen minSup= 0.1 and minConf=0.8 leading to the extraction of 1177 rules. Table 3 shows an extract of the generated rules with their interestingness measures.

Table 3. Extract of the generated ARs

[**] http://sourceforge.net/projects/weka/files/weka-3-7-windows-x64/

AR	Semantic interpretation by expert	Sem Dist
mass_oval→ benign_diagnosis	A well oval mass is highly predictive of benign lesion.	1
irregular_mass→ malign_diagnosis	Irregular mass shape is highly predictive malignant mass.	1
parallel_mass→benign_diagnosis	Parallel orientation with respect to the skin surface is highly predictive benign mass.	1
fibrocystic_breast→ benign_diagnosis	Fibrocystic breast lumps are completely benign, and are not associated with any risk for the future development of breast cancer.	1
Bi-rads3→ follow_up	BI-RADS category 3 lesions are recommended for 6-month follow-up	1
Age[60-69],mass_oval→ benign_diagnosis	Oval mass is predictive of benign lesion (This rule is a specification of the first rule)	1
Age [60-69], mass_circumscribed→ Bi-rads4	The association of circumscribed mass and Age [60-69] induce a classification of Bi-rads4.	1
Age[60-69], mass_oval,mass_circumscribed→ benign_diagnosis	Oval and well circumscribed mass is highly predictive of benign lesion (This rule is as well a specification of the first rule)	0.83
mass_low_density, mass_circumscribed→ Bi-rads4	Circumscribed and low density masses induce a classification of Bi-rads4.	0.66
mass_round, mass_circumscribed→ Bi-rads4	Round and Circumscribed masses induce a classification of Bi-rads4.	0.66

6.5. AR Interestingness results

All together 1177 rules are ranked according to the proposed algorithm, where the dissimilarities between clusters are computed as the interestingness of corresponding rules. To select highly interesting rules, we have fixed a threshold value σ, where ARs with interestingness measure greater than this parameter will be selected. After filtering, rules of low interestingness are unconcerned. The outstanding rules of high interestingness can then be presented to domain experts. To assess the ARs quality, we have varied σ and observe the count of ARs which have been judged as interesting (see Figure 4). The value of interestingness ranges from 0 to 1, with 1 denoting the higher interestingness of the AR. For threshold equals to 0.2, 1025 ARs are judged interesting. If σ is set to 1, then 402 rules are generated. An example of uninteresting rule is 'Breast_pain→ mastalgia', this is due to the fact that both items belong to the same cluster (designating clinical observations). An example of interesting rule: Age [60-69], mass_circumscribed→ Bi-rads4. In fact, this rule involves items of different clusters of the hierarchy.

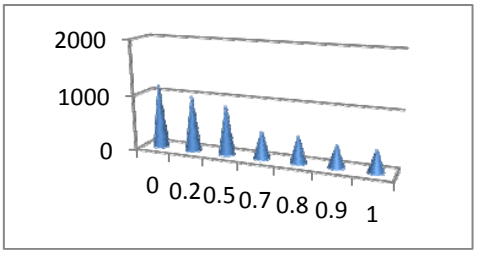

Figure 4: Numbers of interesting rules generated for various interestingness thresholds σ

6.6. Discussion

The path based similarity measure ranges between 0 and 1, with 1 denoting an exact match, while the rest of the values being dependent on the depth in the hierarchy and the distance between the concepts. In other words, it only uses the structural distances instead of semantic content. The use of the knowledge mining concept provides more semantic meaning to the distance between items. Even, the domain experts found that the different partitions helped them to further define their research objectives. Since the semantic similarity relies on the structure of conceptual clusters, this improvement is heavily dependent on the evidence provided by the domain ontology over the real domain knowledge. Consequently, in order to apply our method in different domains, an investigation is required to determine well-structured ontology of the domain of interest.

7. Conclusion

In this paper, an ontology knowledge mining based interestingness measures is proposed. According to doctors, an AR is judged interesting if its items are "dissimilar" that's to say the more different are the items, the more interesting the rule is. Existing ontology based methods compute the path between concepts (items) in the

ontological hierarchy, or concepts of the same category should be considered as “similar”. To tackle this problem, we have proposed to raise the abstraction level of the ontology knowledge base and extract a hierarchy of conceptual clusters of different levels of knowledge granularity. Thus, ARs interestingness is computed according to clusters to which belong the items. As scenario of application, we have selected the mammo ontology since it is well structured and rich of vocabularies. The ranking of ARs has been performed over their interestingness measures. Preliminary results have proved the usefulness of the proposed approach in determining real and precise interestingness of the ARs.

References

[1] I. N. Mohd «Interestingness measures for association rules based on statistical validity,» *Knowledge-Based Systems,* p. 386–392, 2011.

[2] P.Razan, T.Groza, J.Hunter et A.Zankl, «Semantic interestingness measures for discovering association rules in the skeletal dysplasia domain,» *Journal of Biomedical Semantics,* vol. 5, n18, pp. 1-13, 2014.

[3] S.Hassanpour, M.O’Connor et A.K.Das, «Clustering rule bases using ontology-based similarity measures,» *Web Semantics: Science, Services and Agents on the World Wide Web,* vol. 25, pp. 1-8, 2014.

[4] P. Manda, F. McCarthy, B. Nanduri et M. Bridges, «Information Theoretic Interestingness Measures for Cross-Ontology Data Mining in the Mouse Anatomy Ontology and the Gene Ontology,» *Computational Engineering, Finance, and Science (cs.CE),* pp. 1-16, 2015.

[5] M. Shaharanee, I. Nizal, H. Fedja et D. Tharam, «Interestingness measures for association rules based on statistical validity,» *Knowledge-Based Systems,* vol. 24, n° 13, pp. 386-392, 2011.

[6] J. Alcala-Fdez, R. Alcala et F. Herrera, «A fuzzy association rule based classification model for high-dimensional problems with genetic rule selection and lateral tuning,» *IEEE Transactions on Fuzzy systems,* vol. 19, n° 15, pp. 857-872, 2011.

[7] D. Franke, «System and method for efficiently generating association rules,» *U.S. Patent 8,401,986,,* n° %18, 2013.

[8] S. Anthony, R. Karthik, R. Kulathur «Finding Persistent Strong Rules» *Data Mining: Concept, Methodologis* vol. 28, pp. 85-103, 2012.

[9] B. Liu, W. Hsu«Visually Aided Exploration of Interesting Association Rules,» *Knowledge Discovery,* vol. 1574, pp. 26-28, 1999.

[10] S. Concaro, L. Sacchi, C. Cerra, P. Fratino et R. Bellazzi, «Mining healthcare data with temporal association rules: Improvements and assessment for a practical use.,» *Artificial Intelligence in Medicine,* pp. 16-25, 2009.

[11] G. Adomavicius et A. Tuzhilin., «Expert-Driven Validation of Rule» *Data Mining and Knowledge Discovery,* pp. 33-58, 2001.

[12] A. Berrado et G. Runger., «Using metarules to organize discovered AR» *Data Mining and Knowledge Discovery,* pp. 409-431, 2007.

[13] A. Bellandi «Ontology-driven association rules extraction: a case study,»*Representation and Reasoning,* pp. 1-10, 2007.

[14] Y. Kuo, A. Lonie, L. Sonenberg «Domain ontology driven data mining» *Knowledge Discovery and Data Mining,* p. 11–17, 2007.

[15] J. Phillips et B G Buchanan, «Ontology guided knowledge discovery in databases,» *Knowledge Capture,* p. 123–130, 2001.

[16] S. Sharma «Organization-ontology based framework for implementing the business understanding phase» *System Science,* 2008.

[17] J. Wang, J. Han, Y. Lu et P. Tzvetkov, «an efficient algorithm for mining top-k frequent closed itemsets,,» *IEEE Transaction on Knowledge and Data Engineering,* vol. 17, p. 652–664, 2005.

[18] R. Shrikant et R. Agrawal, «Mining Generalized Association Rules,» *Proc.21st Int conf Very Large Database,* pp. 407-419, 1995.

[19] G. Mansingh «The Role of Ontologies in Developing Knowledge Technologies,» *KM for Development,* pp. 145-156, 2015.

[20] C. Marinica et F. Guillet, «Knowledge-based interactive postmining of association rules using ontologies,» *IEEE Transactions on knowledge and data engineering,* vol. 22, p. 784–797, 2010.

[21] G. Mansingh «Using ontologies to facilitate post-processing of association rules,» *Information Sciences,* vol. 181, pp. 419-434, 2011.

[22] « UMLS,» National Library of Medicine, [En ligne]. Available: http://www.nlm.nih.gov/research/umls/.

[23] A.Kumar, B.Smith et C.Borgelt, «Dependence Relationships between Gene Ontology Terms based on TIGR Gene Product Annotations,» *3rd International Workshop on Computational Terminology,* pp. 31-38, 2004.

[24] R.Idoudi, Ettabaa, K., K.Hamrouni, & B.Solaiman, «Ontology Knowledge Mining for Ontology Alignment,» *Internation Journal of computational intelligence systems*, 2016.

[25] The Gene Ontology Consortium, « The Gene Ontology project in 2008,» *Nucleic Acids Research*, 36, pp. 440-444, 2008.

[26] T.Gruber, « A translation approach to portable ontology specifications,» W *Knowledge Acquisition*, 5(2), pp. 199-220, 1993.

[27] J.B.MacQueen, « Some methods for classification of multivariate observations,» *Statistics and Probability*, pp. 281-297, 1976.

[28] M.Uschold, & M.King, « Towards a methodology for building ontologies,» *Ontological Issues in Knowledge Sharing IJCAI*-95, 1995.

[29] S.Myhre «Additional gene ontology structure for improved biological reasoning,» *Bioinformatics*, vol. 22, n116, pp. 2020-2027, 2006.

[30] R.Idoudi, K.Saheb Ettabaa, B.Solaiman, K.Hamrouni, «Fuzzy Clustering based Approach for Ontology Alignment»*International Conference Entreprise Information systels, 24-28 April, 2016.*