http://www.ftsm.ukm.my/apjitm
Asia-Pacific Journal of Information Technology and Multimedia
*Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik*
Vol. 3 No. 1, June 2014 : 1 - 13
e-ISSN: 2289-2192

# EVALUATION AND OPTIMIZATION OF FREQUENT ASSOCIATION RULE BASED CLASSIFICATION

IZWAN NIZAL MOHD SHAHARANEE
JASTINI JAMIL

ABSTRACT

*Deriving useful and interesting rules from a data mining system is an essential and important task. Problems such as the discovery of random and coincidental patterns or patterns with no significant values, and the generation of a large volume of rules from a database commonly occur. Works on sustaining the interestingness of rules generated by data mining algorithms are actively and constantly being examined and developed. In this paper, a systematic way to evaluate the association rules discovered from frequent itemset mining algorithms, combining common data mining and statistical interestingness measures, and outline an appropriated sequence of usage is presented. The experiments are performed using a number of real-world datasets that represent diverse characteristics of data/items, and detailed evaluation of rule sets is provided. Empirical results show that with a proper combination of data mining and statistical analysis, the framework is capable of eliminating a large number of non-significant, redundant and contradictive rules while preserving relatively valuable high accuracy and coverage rules when used in the classification problem. Moreover, the results reveal the important characteristics of mining frequent itemsets, and the impact of confidence measure for the classification task*

*Keywords: rule optimization, interestingness measures, association rules*

## INTRODUCTION

Discovering useful and interesting patterns is one of the main tasks in data mining applications. A pattern is considered interesting and useful if it is comprehensible, valid on both test data and new unseen data, potentially useful, actionable and novel. However, Han and Kamber (2001) claim that, while patterns discovered from the data mining approach are considered strong, but not all are interesting. There are two main problems in dealing with pattern selection, namely the quantity and the quality of the rules (Lenca, Vaillant and Lallich, 2008). Quantity of the rules refers to the problem of generating a large volume of output whilst the quality issues are concerned with the rules potentially reflecting real, significant associations in the domain under investigation.

Various objective interestingness criteria have been used to limit the nature of rules extracted. However, assessing whether a rule satisfies a particular constraint is accompanied by a risk that the rule will satisfy the constraint with respect to the sample data, but not with respect to the whole data distribution (Webb, 2007). As such, the rules may not reflect the real association between the underlying attributes. Since the nature of data mining techniques is data-driven, the generated rules can often be effectively validated by a statistical methodology in order for them to be useful in practice (Goodman, Kamath and Kumar, 2008). Interesting rules are those rules that have a sound statistical basis and are neither redundant nor contradictive. Such an approach requires additional measures based on statistical independence and correlation analysis techniques to verify and evaluate the usefulness and quality of the rules discovered. This will filter out the redundant, misleading, random and coincidentally occurring rules, while at the same time sustaining the accuracy of the rule set and retaining valuable rules.

Although many interestingness and constraint-based measures have been successfully utilized in previous works, there is still a need to understand the roles these parameters play and the way they should be utilized. Thus, in the previous works, the problem addressed by Shaharanee, Hadzic and Dillon (2009) and Shaharanee, Hadzic and Dillon (2011) was developing systematic ways to verify the usefulness of rules obtained from association rule mining. A unified framework, that combines several techniques to assess the quality and remove any redundant and unnecessary rules, has been proposed. This framework shows how the interestingness and constraint based parameters can be utilized and the sequencing of their usage. However, the implication of different confidence values and the time at which the constraint is applied was not investigated. In addition, while confidence measures are often used to reduce the rule set size to only those reflecting highly confident association, no study has been performed on the implication of using different confidence values and the differences of applying this constraint at different stages of the rule verification process. This paper extends the previously developed framework by Shaharanee, Hadzic and Dillon (2009) and Shaharanee Hadzic and Dillon (2011) which seeks to evaluate the impact on classification accuracy, generalization power, and rule coverage rate, when rules are generated using frequent itemset mining algorithms, as well as when different confidence measures are used and applied at different stages of the verification process.

This paper provides an empirical analysis of the usefulness and implications behind using frequent itemset mining for classification tasks, with respect to their classification accuracy and coverage rate. Additionally, the role that the confidence measure plays in the process is highlighted by studying the implications of using high/or low confidence measures.

## PROBLEM DEFINITION

The problem of finding association rules $x \rightarrow y$ was first introduced in (Agrawal, Imieliński and Swami, 1993) as a data mining task for finding frequently co-occurring items in large databases. Let $I = \{i_1, i_2, ..., i_m\}$ be a set of items. Let *D*, be a transactions database for which each transaction $T$ is a set of items, such that $T \subseteq I$. An association rule is an implication of the form of $x \rightarrow y$ where $x \subseteq I$ and $y \subseteq I$ and $x \cap y = \phi$. The *support* of a rule $x \rightarrow y$ is the number of transactions that contain both *x* and *y*. Let the *support* (or *support* ratio) of rule $x \rightarrow y$ (denoted as $\sigma(x \rightarrow y)$) be *s*%. This implies that there are *s*% transactions in *D* that contain items (itemsets) $x$ and *y*. In other words, the probability $P(x \cup y)$ = *s*%. Sometimes, it is expressed as *support count* or *frequency*, that is, it reflects the actual frequency count of the number of transactions in *D* that contain the items that are in the rules. An itemset is *frequent* if it satisfies the user-specified *minimum support* threshold. The *confidence* of a rule $x \rightarrow y$ is the number of transactions containing $x$, that also contain $y$. The *confidence* of a rule $x \rightarrow y$, in other words, is the conditional probability of a transaction containing the *consequent* $(y)$ if the transaction contains the *antecedent* $(x)$. Hence, the *confidence* of a rule $x \rightarrow y$ is calculated as $\sigma(x \rightarrow y) / \sigma(x)$.

Association rule discovery finds all rules that satisfy specific constraints such as the minimum *support* and *confidence* threshold, as is the case in the Apriori algorithm (Agrawal, Imieliński and Swami, 1993). It consists of two main phases: frequent itemsets discovery and association rule generation, of which the former task is the pre-requisite and the most complex. The Apriori-based algorithm has been favorable for frequent itemsets generation as it performs well on sparse data in discovering frequent patterns that are often comprised of rather smaller itemsets.

Let $F_k$ denote the set of frequent *k*–itemsets and $FI$ the set of all frequent itemsets. In the rest of the paper, the focus is on evaluating the rules discovered using the Apriori approach. The datasets with a predefined class label is utilized, where one of the attributes from the dataset is considered as a class to be predicted. Thus, only $FI$ that contain this class attribute is considered.

Let the frequent itemsets $FI$ that have a class label (value) be denoted as $FIC$. The problem can be stated as: Given $FIC$ with accuracy $ac(FIC)$, reduce $FIC$ into $\tilde{FIC}$ such that $\tilde{FIC} \supset FIC$ and $ac(\tilde{FIC}) \geq (ac(FIC) - \varepsilon)$, where $\varepsilon$ is an arbitrary user defined small value ($\varepsilon$ is used to reflect the noise that is often present in real world data).

## RELATED WORK

Previous works on discovering and measuring the interestingness of rules from data are extensive. Such rules may be extracted using specific data mining methods for characterization, classification and prediction, cluster analysis and association rule mining (Han and Kamber, 2001). The patterns and rules generated from these data mining systems can often be large and complex, hindering the analysis process. This motivated another research branch in the data mining field, that of finding interesting, useful and significant patterns.

Many algorithms exist for association rule mining which can be classified as algorithms for association rule improvement, mining well defined subsets of the rule (e.g. closed/maximal) and mining dense datasets. Such classification of the association rule algorithms mentioned earlier refers to a "*classic association rule problem*" (Hipp, Güntzer and Nakhaeizadeh, 2000). Apriori-based association rule mining algorithms have been studied extensively in the classic association rule problem for dense datasets. Patterns from the frequent pattern set are often redundant and unrelated (Wei, Yi and Wang, 2006). Webb (2007) defines redundant rule constraints that are capable of discarding redundant rules. Furthermore, Bayardo, Agrawal and Gunopulos (1999) define a more dominant minimum improvement constraint in order to discard the redundant rules with the development of the Dense-Miner. Han and Kamber (2001) assert that from the rules generated using these data mining systems, often only small subsets of the discovered rules are considered interesting and significant. Several rule/pattern interestingness measures have been developed in order to tackle these issues. Objective, subjective and semantic measures are the three main categories of methods used for discovering interestingness of rules (Aydın and Güvenir, 2009; Geng, and Hamilton, 2006; Han and Kamber, 2001; McGarry, 2005; Simon, Kumar and Li, 2011). While these criteria offer some constraints in discovering strong patterns/rules, many spurious, misleading, uninteresting and insignificant rules may still be produced for many domains (Han and Kamber, 2001). This problem arises because some association rules are discovered due to pure coincidence resulting from certain randomness in the particular dataset being analyzed. Association rule mining frameworks may provide either a true discovery or instances of random behaviors (Lallich, Teytaud and Prudhomme, 2007). To date several works have addressed this rule interestingness issue. The capabilities of statistical analysis in addressing the random effects of patterns from data mining systems have been raised by Hamalainen and Nykanen (2008), Kirsch et al. (2012), Lallich, Teytaud and Prudhomme (2007), Piatetsky-Shapiro (1991), Webb (2007) and Weiß (2008). Statistical independence and correlation analysis are two approaches applied by Han and Kamber (2001) in weeding out uninteresting and misleading data mining patterns. Chi Squared test (Han and Kamber, 2001), Log Linear analysis (Agresti, 2007) and Regression Analysis (Hosmer and Lemeshow,

1989) are several well-known statistical techniques capable of capturing statistical dependence among data items.

The evaluation of the interestingness of rules is essential in many applications. While a substantial number of interestingness and constraint-based measures have been proposed and successfully applied, there is still a need to understand the roles that these parameters play and the way in which they should be utilized. An understanding of the various implications of applying each parameter and providing a systematic, sequential procedure would ensure that one will arrive at a more reliable and interesting set of rules.

## PROPOSED METHOD

Figure 1 shows the proposed framework. The first partition is used for frequent itemsets generation and statistical evaluation, while the second partition acts as sample data drawn from the dataset used to verify the accuracy and coverage rate of the discovered rules. The pre-processing technique is applied to the selected data, to ensure clean and consistent data. The relevance of the input attributes in predicting the class attributes is calculated based on the Symmetrical Tau technique (Zhou and Dillon, 1991) which removes any irrelevant attributes from the initial dataset. The rules are then generated based on frequent itemset mining algorithms. The discovered rules are then evaluated using the statistical analysis, and any rules determined to be statistically insignificant are discarded. Additionally, constraint measurement techniques are employed to discard redundant and contradictive rules.

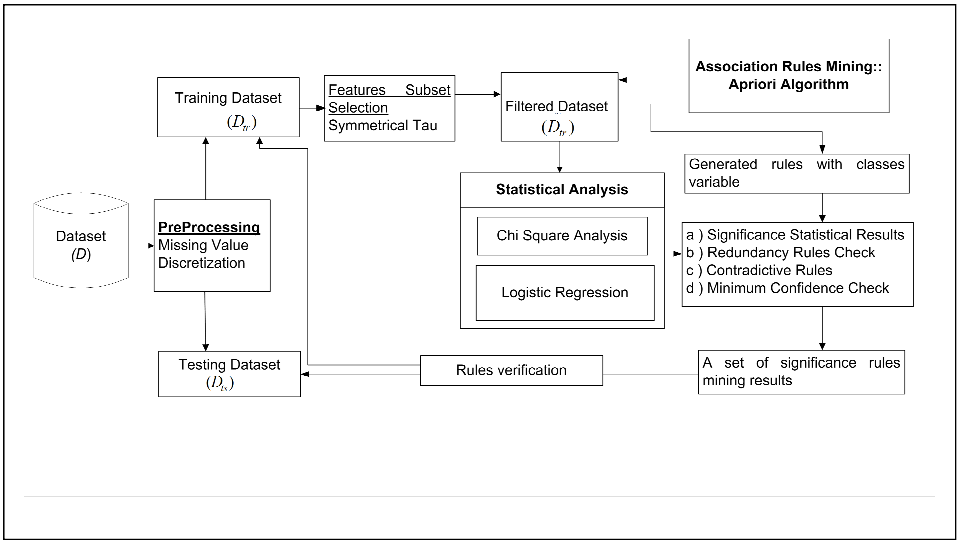



FIGURE 1. Framework for rule interestingness analysis.

A formal description of the conceptual framework follows. Given a relational database $D$, $I = \{i_1, i_2, ..., i_{|D|}\}$ the set of distinct items in $D$, $AT = \{at_1, at_2, ..., at_{|AT|}\}$ the set of input attributes in $D$, and $Y = \{y_1, y_2, ..., y_{|Y|}\}$ the class attribute with a set of class label in $D$. Assume that $D$ contains a set of $n$ records $D = \{x_r, y_r\}_{r=1}^{n}$, where $x_r \subseteq I$ is an item or a set of items and $y_r \in Y$ is a class label, then $|x_r| = |AT|$ and $x_r = \{at_1val_r, at_2val_r, ..., at_{|AT|}val_r\}$ contains the attribute names and corresponding values for record $r$ in $D$ for each attribute $at$ in $AT$. The training dataset is denoted as $D_{tr} \subseteq D$ and the test dataset as $D_{ts} \subseteq D$.

**Pre-processing:** The pre-processing is applied to each $at_i$ in $D$, where $at_i \in AT, (i = (1,...,|AT|))$ in order to obtain clean and consistent data.

**Features Subset Selection:** The relevance of each $at_i$ by determining their importance towards predicting the value of the class attribute $Y$ in $D_{tr}$, where $at_i \in AT, (i = (1,...,|AT|))$ using a statistical-heuristic measure, namely the Symmetrical Tau (Zhou, & Dillon 1991). It measures the capability of an attribute to predict the class of another attribute. Any irrelevant attributes are removed from the dataset, and the filtered database as $\tilde{D}_{tr}$, $\tilde{I} \subseteq I$ is represented.

**Rules Generation (Apriori$^{(S,C)}$):** For a given $\tilde{D}_{tr}$, the association rules were generated based on Apriori framework using minimum *support* and *confidence* thresholds, and the set of obtained association rules are denoted as $F(A)$.

**Rules Generation (Apriori$^{(S)}$):** For a given $\tilde{D}_{tr}$, the association rules were generated based on Apriori framework using only the minimum *support* threshold, and the set of obtained association rules are denoted as $F(B)$.

**Chi Square Test:** For a given $\tilde{D}_{tr}$, the occurrence of $at_i$ where $at_i \in AT, (i = (1,...,|AT|))$ is independent of the occurrence of $Y$ if $P(at_i \cup Y) = P(at_i)P(Y)$; otherwise $at_i$ and $Y$ are dependent and correlated (Han, & Kamber 2001). Hence, Chi Square test discards any $fA_k \in F(A)$ and $fB_k \in F(B)$ for which $\exists at_i$ contained in $x$ of $x \rightarrow y$, the $\chi^2$ value is not significant towards $Y$ (class attribute).

**Logistic Regression Analysis:** For a given $\tilde{D}_{tr}$, several logistic regression models were developed. The model that fits the data well and has the highest predictive capability is selected. The co-efficient $\beta$ of an input attribute $at_i$ where $at_i \in AT, (i = (1,...,|AT|))$, is determined based on the log likelihood value. Hence, logistic regression $\ln(\tilde{Y})$ discards any $fA_k \in F(A)$ and $fB_k \in F(B)$ for which $\exists at_i$ contained in $x$ of $x \rightarrow y$, the $\beta_i at_i$ value is not significant towards the class attribute $Y$. From the initial set of frequent rules, $F(A)$ and $F(B)$ the resulting sets that have been reduced according to the statistical analysis are denoted as $FS(A) = \{fsA_1, fsA_2, ..., fsA_{|FS(A)|}\}$ and $FS(B) = \{fsB_1, fsB_2, ..., fsB_{|FS(B)|}\}$ respectively.

**Redundant and Contradictive Removal:** *Productive* rules based on minimum improvement redundant rule constraint (Bayardo, Agrawal and Gunopulos, 1999), discards any $fsA_k \in FS(A)$ and $fsB_k \in FS(B)$ if $confidence\ (x \rightarrow y) \leq \max_{z \subset x}(confidence\,(z \rightarrow y))$.

In other words, a rule $x \rightarrow y$ with confidence value $c_1$ is considered as redundant if there exists another rule $z \rightarrow y$ with confidence value $c_2$, where $z \subset x$ and $c_1 \leq c_2$.

From the set of statistically reduced frequent rules, $FS(A)$ and $FS(B)$, the resulting sets that have been reduced according to the minimum improvement redundant rule constraints are denoted as $FR(A) = \{frA_1, frA_2, ..., frA_{|FR(A)|}\}$ and $FR(B) = \{frB_1, frB_2, ..., frB_{|FR(B)|}\}$ respectively.

Contradictive rule constraint (Zhang & Zhang 2001), discards any two rules $frX_j, frX_k \in FR(X)$ if $frX_j = x \rightarrow y$ and $frX_k = x \rightarrow \neg y$, where $j,k = (1,...,|FR(X)|)$, $X = (A,B)$ and $j \neq k$. From the rule sets $FR(A)$ and $FR(B)$, the resulting sets that have been reduced according to contradictive rule constraints are denoted as $\tilde{F}(A)$ and $\tilde{F}(B)$, respectively.

**Rules Accuracy and Coverage:** Determining the accuracy and coverage rate of rule sets. For each of the resulting rule sets, $(F(A)$ and $F(B))$, $(FS(A)$ and $FS(B))$, $(FR(A)$ and $FR(B))$, and ( $\tilde{F}(A)$ and $\tilde{F}(B)$), the accuracy rate and the coverage rate in both $D_{tr}$ and $D_{ts}$ are calculated. The combination of these rule evaluating strategies will facilitate the association rule mining framework to determine the right and high quality rules which remain in sets ( $\tilde{F}(A)$ and $\tilde{F}(B)$).

## RESULTS

The evaluation of the unification framework is performed using the Wine, Mushroom, Iris and Adult datasets, real-world datasets of varying complexity obtained from the UCI Machine Learning Repository (Asuncion and Newman, 2007). Since all the datasets used are supervised, which reflects a classification problem, the target variables have been chosen to be the right hand side/consequence of the association rules discovered during association rule mining analysis. An equal depth binning approach is applied to all continuous attributes in the Adult, Iris and Wine datasets. This equal depth binning approach will ensure manageable data sizes by reducing the number of distinct values per attribute (Han and Kamber, 2001). Other discrete attributes in the Adult and Mushroom datasets were preserved in their original state.

TABLE 1. Dataset Characteristics.

| Dataset | #Records | #Attributes | # Selected Attributes. (Sym. Tau) | # of Rules with Target Variable Apriori $^{(S,C)}$ | Apriori $^{(S)}$ |
|---|---|---|---|---|---|
| Wine | 178 | 13 | 12 | 234 | 272 |
| Adult | 45222 | 15 | 10 | 1680 | 2192 |
| Mushroom | 8124 | 23 | 11 | 75237 | 77815 |
| Iris | 150 | 4 | 4 | 51 | 58 |

Table 1 indicates the characteristics of the aforementioned datasets used in our evaluation. The Apriori$^{(S,C)}$ in Column 5 will act as the initial benchmark having both the minimum *support* of 10% and the minimum *confidence* of 60% in generating the rule set. The Apriori$^{(S)}$ in Column 6 will discover only the rules based on the minimum *support* of 10%.

## APRIORI$^{(S,C)}$ VS APRIORI$^{(S)}$

Apriori algorithms have demonstrated a good performance in generating frequent patterns (Han and Kamber, 2001). However, the patterns generated need to be evaluated in order to arrive at significant and useful patterns. A unification framework for evaluating the interestingness of frequent itemsets obtained by the Apriori algorithm was previously developed and reported in Shaharanee, Dillon and Hadzic (2009) and Shaharanee, Hadzic and Dillon (2011). It was found that the rules generated from the Apriori algorithm were large and contaminated with useless patterns. With appropriate statistical analysis, and redundancy and contradictive assessment methods, the unification framework managed to discard a large number of rules while still preserving high accuracy and coverage rate of the final reduced rule set.

In this section, the usefulness of the rules generated from both variants is compared. Table 2 reveals the progressive difference in the number of rules, the Accuracy Rate (AR) and the Coverage Rate (CR) values, as the Symmetrical Tau (ST) feature selection application, statistical analysis, redundancy and contradictive assessment methods are utilized. For most of the discovered rules in Table 2, the AR in the training set was consistently higher than the testing set. This is due to the fact that the discovered rules were generated from the training set, and as a consequence, the rules mostly fit well the characteristics of the data objects that exist predominantly in the training set.

TABLE 2. Comparison between Apriori $^{(S,C)}$and Apriori $^{(S)}$ in Wine Dataset.

| Type of analysis | Data Partition | Apriori $^{(S,C)}$ | | | Apriori $^{(S)}$ | | |
|---|---|---|---|---|---|---|---|
| | | # Of Rules | AR % | CR% | # Of Rules | AR % | CR % |
| Initial # of Rules | Training | 234 | 87.58 | 100.00 | 272 | 76.83 | 100.00 |
| | Testing | | 79.84 | 100.00 | | 69.68 | 100.00 |
| # of Rules after ST | Training | 195 | 87.53 | 100.00 | 217 | 74.26 | 100.00 |
| | Testing | | 79.44 | 100.00 | | 68.00 | 100.00 |
| Statistics Analysis | Training | 17 | 85.07 | 100.00 | 24 | 64.16 | 100.00 |
| | Testing | | 81.98 | 100.00 | | 60.46 | 100.00 |
| Redundant Removal | Training | 16 | 85.07 | 100.00 | 23 | 63.52 | 100.00 |
| | Testing | | 81.98 | 100.00 | | 60.05 | 100.00 |
| Contradictive Removal | Training | 16 | 85.07 | 100.00 | 16 | 85.63 | 100.00 |
| | Testing | | 81.98 | 100.00 | | 81.94 | 100.00 |
| Confidence 60% | Training | | | | 15 | 87.84 | 100.00 |
| | Testing | | | | | 84.77 | 100.00 |

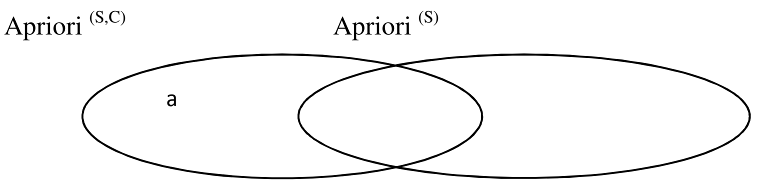
The initial number of rules from Apriori constrained with *min_sup* is larger compared to the initial number of rules in Apriori constrained with both *min_sup* and *min_conf* due to the removal of the minimum *confidence* threshold. As application of the Symmetrical Tau, statistical analysis and redundancy assessment were progressively applied to the initial set of rules, at least 90% of the rules in the rule set have been discarded. Both AR values for the testing dataset in Apriori $^{(S,C)}$ and Apriori $^{(S)}$ increased while the CR of the rules was still preserved at 100%. As an extension of our previous work in Shaharanee, Dillon and Hadzic (2009), another method of analysis to discard contradictive rules (Zhang and Zhang, 2001) was included. Contradictive rules exist in Apriori $^{(S)}$ because they are constrained by only a minimum *support* threshold, because at the set *confidence* threshold of 60% in Apriori $^{(S,C)}$, they do not exist. However, this also points to the important difference. The rules with confidence higher than 60% that are contradictive to other frequent rules in the data, which cannot be present in the rule set as they cannot have 60% confidence at the same time, will remain in the rule set, but will have a higher misclassification rate. Hence, their contradictive nature would not be captured, which essentially would negatively affect the accuracy of the rule set as a whole. An example of this scenario is provided later. The contradictive rules detected in Apriori $^{(S)}$ rule set are shown in Table 3.

TABLE 3. List of contradictive rules in Wine dataset for Apriori$^{(S)}$.

| Confidence (%) | Support (%) | Rules |
|---|---|---|
| 64.10 | 23.36 | Flavanoids(2.24 - 3.18) ==> Class(Low) |
| 35.90 | 13.08 | Flavanoids(2.24 - 3.18) ==> Class(Middle) |
| 57.50 | 21.50 | ColorIntensity(3.62 - 5.97) ==> Class(Low) |
| 30.00 | 11.21 | ColorIntensity(3.62 - 5.97) ==> Class(High) |
| 38.78 | 17.76 | Magnesium(88.4 - 106.8) ==> Class(Low) |
| 34.69 | 15.89 | Magnesium(88.4 - 106.8) ==> Class(High) |
| 26.53 | 12.15 | Magnesium(88.4 - 106.8) ==> Class(Middle) |

With the removal of the contradictive rules in Apriori $^{(S)}$, both approaches now contain the same number of rules (16) with only a modest difference in AR% as shown in Table 2. Even though both contain the same number of rules, there are still differences as shown in

Figure 2. These differences are due to the sequence of the evaluation process in both approaches. Rule (b) does not appear in Apriori $^{(S,C)}$ due to the *confidence* value being lower than the minimum threshold of 60%, while rule (a) does not exist in Apriori $^{(S)}$ because the rule contradicts another rule (see Table 3 row 3).




| | Confidence. (%) | Support (%) | Rules |
|---|---|---|---|
| a | 64.10 | 23.36 | Flavanoids(2.24 - 3.18) ==> Class(Low) |
| b | 58.33 | 13.08 | Magnesium(88.4 - 106.8) & ColorIntensity(3.62 - 5.97) ==> Class(Low) |

FIGURE 2. Rule differences between Apriori $^{(S,C)}$ and Apriori $^{(S)}$ after contradictive rule removal.

Finally, the minimum *confidence* constraint was utilized on the Apriori $^{(S)}$ rule set and 15 rules were obtained as our final significant rule set (i.e. Rule (b) from Figure 2 was removed). As for the final 15 rules, the AR value in Apriori $^{(S)}$ is higher than Apriori $^{(S,C)}$, while the CR value remained the same (see Table 2). When the individual accuracy of each rule was checked, it was exactly rules (a) and (b) (Figure 2) causing lower AR in the rules from Apriori $^{(S,C)}$ and Apriori $^{(S)}$, respectively. Rule (a) was discarded in Apriori $^{(S)}$ because it contradicted another rule as shown in Table 3.

This knowledge of rule (a) being contradictive to another rule (frequent association to another class value) was not available in Apriori $^{(S,C)}$ because the minimum *confidence* constraint was applied at the start. This approach missed the fact that association "Flavanoids(2.24 - 3.18) ==> Class(Middle)" occurred frequently enough to know that the rule "Flavanoids(2.24 - 3.18) ==> Class(Low)" is not reliable enough to be used for prediction. This is supported by the fact that the AR of the final 15 rules is higher than the AR of the 16 rules from Apriori $^{(S,C)}$ containing the contradictive rule. In Apriori Apriori $^{(S,C)}$, the contradictive rule "Flavanoids(2.24 - 3.18) ==> Class(Low)" has misclassified 14 instances from the training set and 10 instances from the testing set. By removing this rule, a portion of the misclassified instances is captured by other rule(s) that are based on different attribute constraints, and there is an increase in accuracy as seen in Table 2.

These results suggest that it may be advantageous to not apply the confidence constraints at the start of the process but rather at the end or after any contradictive frequent rules/patterns have been removed. Another option would be to start with a lower confidence threshold to still discard those patterns where the confidence is not high enough for them to be considered as a significant contradiction to another rule with much higher confidence. One can then increase the threshold, and the effects of progressively increasing the confidence threshold are shown in Section 5.2. This relationship between contradictive rules and the application of a confidence threshold was not discussed in Zhang and Zhang (2001), where the contradictive assessment was introduced.

The comparison of the rules generated from the Apriori $^{(S,C)}$ and Apriori $^{(S)}$ of the Iris, Mushroom and Adult datasets is fairly similar to the rules extracted from the Wine dataset. The initial rule set from the Apriori $^{(S)}$ algorithm is naturally always larger than the rule set of the Apriori $^{(S,C)}$ algorithm as depicted in Tables 4, 5 and 6.

The Symmetrical Tau (ST) application, statistical analysis, redundancy and contradictive assessment methods, and a specific minimum confidence threshold (Apriori $^{(S)}$) are progressively applied to each rule set. As the number of rules for each dataset and each variant was reduced dramatically, the AR for the training and the testing dataset increased gradually except for the rule set from Apriori $^{(S,C)}$ for Iris and Mushroom dataset as there are slight decreases in their AR. While the CR for each of the Mushroom and Iris datasets was well preserved at 100%, the CR in Adult marginally decreased. The Adult dataset is characterized by imbalanced target data, as discussed in Liu, Ma and Wong (2000) and Shaharanee, Hadzic and Dillon (2011), and many rules were discarded so there were no rules left to cover the rarely occurring class value '>50K'.

TABLE 4. Comparison between Apriori $^{(S,C)}$ and Apriori $^{(S)}$ in Iris Dataset.

| Type of analysis | Data Partition | Apriori $^{(S,C)}$ | | | Apriori $^{(S)}$ | | |
|---|---|---|---|---|---|---|---|
| | | # Of Rules | AR % | CR% | # Of Rules | AR % | CR % |
| Initial # of Rules | Training | 51 | 92.86 | 100.00 | 58 | 81.77 | 100.00 |
| | Testing | | 90.99 | 100.00 | | 78.46 | 100.00 |
| # of Rules after ST | Training | 51 | 92.86 | 100.00 | 58 | 81.77 | 100.00 |
| | Testing | | 90.99 | 100.00 | | 78.46 | 100.00 |
| Statistics Analysis | Training | 22 | 88.15 | 100.00 | 29 | 71.60 | 100.00 |
| | Testing | | 85.29 | 100.00 | | 68.07 | 100.00 |
| Redundancy Removal | Training | 22 | 88.15 | 100.00 | 29 | 71.60 | 100.00 |
| | Testing | | 85.29 | 100.00 | | 68.07 | 100.00 |
| Contradictive Removal | Training | 22 | 88.15 | 100.00 | 21 | 89.79 | 100.00 |
| | Testing | | 85.29 | 100.00 | | 86.43 | 100.00 |
| Confidence 60% | Training | | | | 21 | 89.79 | 100.00 |
| | Testing | | | | | 86.43 | 100.00 |

TABLE 5. Comparison between Apriori $^{(S,C)}$ and Apriori $^{(S)}$ in Mushroom Dataset.

| Type of analysis | Data Partition | Apriori $^{(S,C)}$ | | | Apriori $^{(S)}$ | | |
|---|---|---|---|---|---|---|---|
| | | # Of Rules | AR % | CR% | # Of Rules | AR % | CR % |
| Initial # of Rules | Training | 75237 | 94.27 | 100.00 | 77815 | 91.79 | 100.00 |
| | Testing | | 94.34 | 100.00 | | 83.20 | 100.00 |
| # of Rules after ST | Training | 653 | 91.63 | 100.00 | 669 | 89.97 | 100.00 |
| | Testing | | 91.75 | 100.00 | | 90.08 | 100.00 |
| Statistics Analysis | Training | 44 | 92.43 | 100.00 | 48 | 81.20 | 100.00 |
| | Testing | | 92.51 | 100.00 | | 81.06 | 100.00 |
| Redundancy Removal | Training | 21 | 91.33 | 100.00 | 24 | 76.97 | 100.00 |
| | Testing | | 91.28 | 100.00 | | 76.88 | 100.00 |
| Contradictive Removal | Training | 21 | 91.33 | 100.00 | 20 | 94.62 | 100.00 |
| | Testing | | 91.28 | 100.00 | | 94.24 | 100.00 |
| Confidence 60% | Training | | | | 20 | 94.62 | 100.00 |
| | Testing | | | | | 94.24 | 100.00 |

The differences between the final number of rules for both Apriori $^{(S,C)}$ and Apriori $^{(S)}$, in each of the Iris, Mushroom and Adult datasets are due to the sequence of the evaluation process as mentioned earlier. For the final rule sets obtained from the Iris, Mushroom and Adult datasets, the Apriori $^{(S)}$ approach achieved higher accuracy, which again confirms our earlier suggestion to apply the *confidence* constraint after the contradictive rules have been removed. In all cases, the Apriori $^{(S)}$ approach removed a contradictive rule that remained in Apriori $^{(S,C)}$.

TABLE 6. Comparison between Apriori $^{(S,C)}$ and Apriori $^{(S)}$ in Adult Dataset.

| Type of analysis | Data Partition | Apriori $^{(S,C)}$ | | | Apriori $^{(S)}$ | | |
|---|---|---|---|---|---|---|---|
| | | # Of Rules | AR % | CR% | # Of Rules | AR % | CR % |
| Initial # of Rules | Training | 1680 | 81.23 | 100.00 | 2192 | 68.98 | 100.00 |
| | Testing | | 81.35 | 100.00 | | 69.05 | 100.00 |
| # of Rules after ST | Training | 233 | 80.46 | 100.00 | 303 | 67.46 | 100.00 |
| | Testing | | 80.50 | 100.00 | | 67.45 | 100.00 |
| Statistics Analysis | Training | 71 | 81.49 | 100.00 | 107 | 63.83 | 100.00 |
| | Testing | | 81.65 | 100.00 | | 63.87 | 100.00 |
| Redundancy Removal | Training | 46 | 85.46 | 100.00 | 58 | 69.65 | 100.00 |
| | Testing | | 85.61 | 100.00 | | 69.72 | 100.00 |
| Contradictive Removal | Training | 46 | 85.46 | 100.00 | 48 | 81.79 | 99.98 |
| | Testing | | 85.61 | 100.00 | | 81.91 | 99.95 |
| Confidence 60% | Training | | | | 43 | 88.31 | 96.38 |
| | Testing | | | | | 88.41 | 96.12 |

## MINIMUM *CONFIDENCE* EFFECT

While conducting experiments on the Wine dataset (Refer to Table 2), it has been observed that the performance of the AR and CR can vary by altering the value of minimum confidence. By increasing the minimum *confidence* from 60% to 70%, the CR values in the training set remained stable at 100%. While there was an increase in the AR values for the test set, the CR values decreased. Such a condition occurs because the 13 rules failed to capture all of the instances in this dataset. As the confidence thresholds are gradually increased to 70%, 80%, 90% and 100%, the number of rules in the rule sets became smaller and identical, which lead to the increase in AR but at the cost of decreasing the number of instances covered by the rules. The changes in confidence values have a direct impact on the size of the rule set, AR, and CR values. Progressively increasing the minimum confidence threshold results in an even smaller set of rules which are more accurate but then the CR suffers (Table 7). Thus, determining the tradeoff between finding a rule set with optimal values of AR and CR is essential (Novak, Lavrač and Webb, 2009). This agrees with Wang, Dillon and Chang (2002), who assert the need for balancing these conflicting regularization parameters.

Table 7 show the effect of altering the minimum confidence of rules obtained from all datasets. Such results are in agreement with Do, Hui and Fong (2005), who state that a rule with a high confidence value implies an accurate prediction. However, as shown in Table 7, even though the AR increased simultaneously with the increment of minimum confidence values, the CR values decreased as a result. This depicted the trade-off in choosing the suitable minimum confidence threshold for each dataset or domain considered. For example, in the Mushroom dataset, it appears that for best results, the confidence could have been safely set up to 80% without a loss in coverage rate.

Restricting the rule sets according to the minimum *confidence* values impacts on the trade-off between accuracy and coverage rates. Experiments show that, the AR increase simultaneously with the increase of the confidence values. However at some stages, too many rules will be discarded which significantly make the coverage rate suffer. It is important in this framework to monitor the CR in reducing the number of rules and to identify the break point/right time at which to stop reducing the number of rules (increasing the confidence values).

TABLE 7. Minimum *Confidence* Effect for Wine, Iris, Mushroom and Adult Dataset.

| Type of analysis | Data Partition | Wine | | | Iris | | | Mushroom | | | Adult | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | #Rules | AR % | CR % | #Rules | AR % | CR % | #Rules | AR % | CR % | #Rules | AR % | CR % |
| Conf. 60% | Training | 15 | 87.84 | 100 | 21 | 89.79 | 100 | 20 | 94.62 | 100 | 43 | 88.31 | 96.38 |
| | Testing | | 84.77 | 100 | | 86.43 | 100 | | 94.24 | 100 | | 88.41 | 96.12 |
| Conf. 70% | Training | 13 | 92.03 | 100 | 19 | 92.91 | 100 | 20 | 94.62 | 100 | 41 | 89.63 | 93.78 |
| | Testing | | 89.6 | 98.59 | | 93.23 | 100 | | 94.24 | 100 | | 89.75 | 93.44 |
| Conf. 80% | Training | 11 | 95.19 | 99.07 | 17 | 94.76 | 100 | 19 | 95.84 | 100 | 38 | 90.61 | 90.45 |
| | Testing | | 90.14 | 97.18 | | 95.98 | 100 | | 95.51 | 100 | | 90.72 | 90.05 |
| Conf. 90% | Training | 9 | 98.04 | 85.98 | 14 | 97.25 | 94.44 | 15 | 98.15 | 99.47 | 21 | 96.16 | 53.5 |
| | Testing | | 91.26 | 83.10 | | 97.89 | 88.33 | | 97.67 | 99.69 | | 96.00 | 53.9 |
| Conf. 100% | Training | 6 | 100 | 58.88 | 9 | 100 | 74.44 | 8 | 100 | 85.86 | 0 | - | - |
| | Testing | | 92.98 | 53.52 | | 100 | 71.67 | | 100 | 85.88 | | - | - |

## CONCLUSIONS AND FUTURE WORKS

This paper has presented an empirical analysis of the usefulness and implication behind using frequent patterns for classification tasks, with respect to their classification accuracy and coverage rate. The quality of the rules discovered are measured based on a statistical, redundancy and contradictive assessment methods.

Initially, two variants of the Apriori algorithm were evaluated. The first variant corresponded to the standard Apriori algorithm with both *support* and *confidence* threshold, while the second variant was constrained using only the minimum *support* threshold. The result demonstrated that the Apriori algorithm with a minimum s*upport* variant produced more rules in comparison with the first variant, due to no constraint being imposed regarding the *confidence* of the rules. Rules were then verified in order to determine their validity and interestingness. The results show that it is more advantageous to remove the rules that failed the statistical test, the redundant rules, and the contradictive rules in the initial evaluating process and utilize the confidence constraint only at the end of the process. This will result in a relatively small number of rules and at the same time any detected contradictive rules will be removed. As demonstrated in the experiments, a drawback of applying the minimum *confidence* threshold at the start of the process is the existence of a contradictive rule that has relatively low confidence will go unnoticed. This lack of knowledge can cause an unreliable association rule to become part of the final rule set which, as demonstrated, reduces the accuracy of the rule set in comparison to when the rule was removed. Alternatively, in the second variant (Apriori with minimum support*)* approach, initially the two or more contradictive rules exist so all of the contradictive rules will be discarded, as the contradiction implies that they are unreliable for prediction purposes. An alternative approach would be to start with a lower confidence threshold to still discard those patterns where the confidence is not high enough for them to be considered as a significant contradiction to another rule with much higher confidence. One can then progressively increase the threshold after the statistical heuristic rule validation techniques have been applied. Based on the proper rule evaluating steps in the proposed framework, the final rules from the Wine, Iris, Mushroom, and Adult datasets generated using the second variant are fewer in number and achieve a better classification and prediction accuracy for both the training and the test datasets.

In the second experiment, the minimum confidence effects on the proposed framework are demonstrated. Increasing the confidence threshold will gradually reduce the number of rules to those that have high accuracy because of large confidence. However, as the rule sets have been reduced, more instances will not be captured by the rule set; hence, typically there is deterioration in the CR. Choosing smaller confidence thresholds will result in larger sets of rules that may lack in generalization power, thereby weakening the AR performance but are capable of covering more instances. Alternatively, choosing relatively high confidence thresholds will result in a smaller set of rules thereby achieving higher AR with the tradeoff of capturing fewer instances. Thus, it is important to balance the trade-off between AR and CR in order to determine the optimal value for the minimum confidence threshold, which may differ depending on the sensitivity of the domain at hand.

The experimental results have demonstrated that the proposed framework managed to reduce a large number of non-significant and redundant rules while simultaneously preserving a relatively high level of accuracy. As part of the ongoing works (Shaharanee and Hadzic, 2013), the proposed framework is intended to be used to evaluate the differences between frequent, maximal and close patterns when used for classification tasks, and the effect of the confidence threshold.

## REFERENCES


Agrawal, R., Imieliński, T., and Swami, A. 1993. Mining association rules between sets of items in large databases. In Buneman, P., and Jajodia, S. (eds.) *Proceedings of the ACM SIGMOD International Conference on Management of Data.* New York: ACM, 207-216.

Agresti, A. 2007. *An introduction to categorical data analysis. Wiley series in probability and mathematical statistics.* 2nd ed. New Jersey:Wiley-Interscience.

Asuncion, A. & Newman, D. J. 2007. *UCI Machine Learning Repository*. University of California, School of Information and Computer Sciences. http://www.ics.uci.edu/~mlearn/MLRepository.html [5 March 2013]

Aydın, T., and Güvenir, H. A. 2009. Modeling interestingness of streaming association rules as a benefit-maximizing classification problem. *Knowledge-Based Systems*, 22(1): 85–99.

Bayardo, R. J. J., Agrawal, R., and Gunopulos, D. 1999. Constraint-based rule mining in large, dense databases. *Proceedings of the 15th International Conference on Data Engineering*, California: IEEE Computer Society, 188-197.

Do, T. D., Hui, S. C., and Fong, A.C.M. 2005. Prediction confidence for associative classification. *Proceedings of the 4th International Conference on Machine Learning and Cybernetics*, California: IEEE Computer Society, 4: 1993-1998.

Geng, L., and Hamilton, H. J. 2006. Interestingness measures for data mining. *ACM Computing Surveys*, 38(3): 1-32

Goodman, A., Kamath, C., and Kumar, V. 2008. Data Analysis in the 21st Century. *Statistical Analysis and Data Mining*, 1(1):1–3.

Hamalainen, W., and Nykanen, M. 2008. Efficient Discovery of Statistically Significant Association Rules. *Proceeding of the 8th IEEE International Conference on Data Mining*. California: IEEE Computer Society, 203–212.

Han, J., and Kamber, M. 2001. *Data mining: concepts and techniques. The Morgan Kaufmann series in data management systems*. San Francisco:Morgan Kaufmann.

Hipp, J., Güntzer, U., and Nakhaeizadeh, G. 2000. Algorithms for association rule mining - a general survey and comparison. *ACM SIGKDD Explorations Newsletter*, 2(1):58–64.

Hosmer, D. W., and Lemeshow, S. 1989. *Applied logistic regression*. New York: Wiley.

Kirsch, A., Mitzenmacher, M., Pietracaprina, A., Pucci, G., Upfal, E., and Vandin, F. 2012. An efficient rigorous approach for identifying statistically significant frequent itemsets. *Journal of the ACM (JACM)*, 59(3):12.

Lallich, S., Teytaud, O., and Prudhomme, E. 2007. Association rule interestingness: measure and statistical validation. In Guillet, F.J. & Hamilton, H.J. (eds.) *Quality measures in data mining*, 251–275. Berlin/Heidelberg:Springer.

Lenca, P., Meyer, P., Vaillant, B., and Lallich, S. 2008. On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. *European Journal of Operational Research*, 184(2):610–626.

Liu, B., Ma, Y., and Wong, C. 2000. Improving an Association Rule Based Classifier. In Zighed, Komorowski, and Zytkow (eds.) *Principles of Data Mining and Knowledge Discovery*, Vol. 1910: 293–317. Berlin/Heidelberg:Springer.

McGarry, K. 2005. A survey of interestingness measures for knowledge discovery. *Knowl. Eng. Rev.*, 20(1): 39–61.

Novak, P.K., Lavrač, N., and Webb, G.I. 2009. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *The Journal of Machine Learning Research*, 10: 377–403.

Piatetsky-Shapiro, G. 1991. Discovery, analysis and presentation of strong rules. In Piatetsky-Shapiro, G., & Frawley, W.J. (eds.) *Knowledge Discovery in Databases*, 229-248. Menlo Park, California:AAAI Press.

Shaharanee, I.N.M., and Hadzic, F. 2013. Evaluation and optimization of frequent, closed and maximal association rule based classification. *Statistics and Computing*, 1–23.

Shaharanee, I.N.M., Hadzic, F., and Dillon, T.S. 2009. Interestingness of association rules using Symmetrical Tau and Logistic Regression. In Nicholson, A., and Li, X. (eds.) *AI 2009 Advances in Artificial Intelligence*, 5866:442-431. Berlin/Heidelberg:Springer.

Shaharanee, I.N.M., Hadzic, F., and Dillon, T.S. 2011. Interestingness measures for association rules based on statistical validity. *Knowledge-Based Systems*, *24*(3): 386–392.

Simon, G. J., Kumar, V., and Li, P.W. 2011. A simple statistical model and association rule filtering for classification. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM, 823–831.

Wang, D., Dillon, T., and Chang, E. 2002. Trading off between Misclassification, Recognition and Generalization in Data Mining with Continuous Features. In Hendtlass, T., and Ali, M. (eds.) *Developments in Applied Artificial Intelligence*, 2358: 121–130. Berlin/Heidelberg:Springer.

Webb, G. I. 2007. Discovering Significant Patterns. *Machine Learning*, 68(1): 1–33.

Wei, J.M., Yi, W.G., and Wang, M.Y. 2006. Novel measurement for mining effective association rules. *Knowledge-Based Systems*, 19(8):739–743.

Weiß, C.H. 2008. Statistical mining of interesting association rules. *Statistics and Computing*, 18(2):185–194.

Zhang, C., and Zhang, S. 2001. Collecting quality data for database mining. In Stumptner, M., Corbett, D., and Brooks, M. (eds.) *AI 2001: Advances in Artificial Intelligence*, 2256: 593–604. Berlin/Heidelberg: Springer.

Zhou, X. J., and Dillon, T.S. 1991. A statistical-heuristic feature selection criterion for decision tree induction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8):834-841.

Izwan Nizal Mohd Shaharanee,
Jastini Jamil
School of Quantitative Sciences,
Universiti Utara Malaysia,
06010 Sintok, Kedah, Malaysia.
nizal@uum.edu.my, jastini@uum.edu.my

Received: 13 August 2013
Accepted: 22 October 2013