Journal of ICT, 14, 2015, pp: 95–110

IRRELEVANT FEATURE AND RULE REMOVAL FOR STRUCTURAL ASSOCIATIVE CLASSIFICATION

[1]Izwan Nizal Mohd Shaharanee & [2]Jastini Mohd Jamil
[1&2]School of Quantitative Sciences, Universiti Utara Malaysia, Malaysia

nizal@uum.edu.my; jastini@uum.edu.my

ABSTRACT

In the classification task, the presence of irrelevant features can significantly degrade the performance of classification algorithms, in terms of additional processing time, more complex models and the likelihood that the models have poor generalization power due to the over fitting problem. Practical applications of association rule mining often suffer from overwhelming number of rules that are generated, many of which are not interesting or not useful for the application in question. Removing rules comprised of irrelevant features can significantly improve the overall performance. In this paper, we explore and compare the use of a feature selection measure to filter out unnecessary and irrelevant features/attributes prior to association rules generation. The experiments are performed using a number of real-world datasets that represent diverse characteristics of data items. Empirical results confirm that by utilizing feature subset selection prior to association rule generation, a large number of rules with irrelevant features can be eliminated. More importantly, the results reveal that removing rules that hold irrelevant features improve the accuracy rate and capability to retain the rule coverage rate of structural associative association.

Keywords: Features selection, rules removal, frequent item set mining.

INTRODUCTION

Irrelevant and redundant attributes can easily contaminate a real word dataset. These features can degrade the performance and interfere with any data mining processes typically resulting in reduction on the quality of the discovered rules/ patterns. Generally, the feature subset selection tasks is to find the necessary

and sufficient subset of features or attributes which results in simplification of the discovered knowledge model, better generalization power, while at the same time not compromising the accuracy for classification tasks.

Association rule mining is a useful data mining technique capable of discovering interesting relationships hidden in large datasets. It has also been utilized in the classification task, where it can discover strong associations between occurring attributes and class values (Shaharanee & Hadzic, 2013; Tan, Steinbach, & Kumar, 2014). One important property of the frequent pattern-based classifier is that it generates frequent patterns without considering their predictive power (Cheng, Yan, Han, & Hsu, 2007). This property will result in a huge feature space for possible frequent patterns. Feature subset selection is one of the steps performed in the pre-processing stage of the data mining process to remove any irrelevant attributes. If the whole dataset were used as input, this would produce a large number of rules, many of which are created or made unnecessarily complex by the presence of irrelevant and/or redundant attributes. Determining the relevant and irrelevant attributes poses a great challenge to many data mining algorithms (Roiger & Geatz, 2003). If the irrelevant attributes are left in the dataset, they can interfere with the data mining process and the quality of the discovered patterns may deteriorate (Cheng et al., 2007). Furthermore, if a large volume of attributes is present in a dataset, this will slow down the data mining process.

To overcome these problems, it is important to find the necessary and sufficient subset of features so that the application of association rules mining will be optimal and no irrelevant features will be present within the discovered rules. This would prevent the generation of rules that include any irrelevant and/or redundant attributes. In this work, we explore the application of feature subset selection measures to filter out unnecessary and irrelevant features/attributes for the associative classification task. A feature subset selection method is used prior to association rule generation. Once the initial set of rules is obtained, irrelevant rules are determined as those that are comprised of attributes not determined to be statistically significant for the classification task. The experiments are conducted using real-world datasets of varying complexity obtained from the UCI Machine Learning Repository (Asuncion & Newman, 2007). The results indicate that feature subset selection discards a large number of insignificant attributes/features thus eliminating a large number of non-significant rules while preserving relatively valuable high accuracy and coverage rules when used in the classification problem.

LITERATURE REVIEW

The feature subset selection as describes in Han, Kamber and Pei (2011) is a way to minimize the number of features within the dataset by removing irrelevant or redundant features/attributes. In general, the objective of feature subset selection as defined in Han et al. (2011) is "to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes". Tan, Stcinbach & Kumar, (2014) asserted that domain expertise can be employed in order to pick up useful attributes. Nevertheless, this task is often exhaustive and involves a great deal of time. This is due to the data mining application which involves complex data that is massive in size.

The test of statistical significance has been renowned as the better way in evaluating the usefulness of attributes/features. Han et al. (2011) have identified three commonly used heuristic techniques utilized in statistical significance tests namely the stepwise forward selection, stepwise backward selection and a combination of both for the development of regression model. However the capability of the software used for building the regression model is limited (Shaharanee and Hadzic, 2013). For example, there is a limit on the number of unique values each attribute can have. Hence, an additional assumption that was included is that, any input attributes exceeding a certain limit of its possible values will be omitted. The removal of these attributes might be useful in reducing the number of attributes for logistic regression models; however, they might also contain important information for the classification task, and thus their removal may result in deterioration of the model. Moreover, the application of correlation analysis such as the chi-squared test utilized by Brin, Motwani and Silverstein (1997) and Han et al. (2011) is also valuable in identifying redundant variables for features subset selection. This statistical-based test offers a way to determine the closeness of two probability distributions and is capable of accessing the statistical significance level of dependence between the antecedent and consequent in association rules. Another powerful statistical technique for this purpose is the Symmetrical Tau (Zhou & Dillon, 1991) as described in Hadzic and Dillon (2006), which is a statistical-heuristic feature selection criterion. This measure was derived from the Goodman and Kruskal Asymmetrical Tau measure of association for cross-classification task in the statistical area. It measures the capability of an attribute in predicting the class of another attribute. The Symmetrical Tau measure has been proven to be useful for feature subset selection problems in decision tree learning.

Another type of feature selection technique is based on the information theory procedure. Information theory refers to the interpretation of information from patterns. A feature within a structural associative classification rule is considered useful when the feature provides a great deal of information about the class (Blanchard, Guillet, Gras, & Briand, 2005). There are various measures for evaluating the features based on the information theory approach. The mutual information as characterized by Geng and Hamilton (2006), Jaroszewicz and Simovici (2001), and Ke, Cheng and Ng (2008) is a measure that describes how much information one random variable imparts about another one. Blanchard et al. (2005) and Geng and Hamilton (2006) described the Shannon conditional entropy as an information theory that calculates the average amount of information of the consequent given that the antecedent is true. The J-measure as proposed by Smyth and Goodman in Smyth and Goodman (1992) is an information measure capable of quantifying the information content of a rule or a hypothesis. The Gini index as reported in Bayardo Jr. and Agrawal (1999), Blanchard et al. (2005), and Jaroszewicz and Simovici (2001) is a measure based on distribution divergence. ID3 (Quinlan, 1986) and C4.5 (Quinlan, 1993) as described in Han et al. (2011) are capable in selecting the most prominent class distinguishing attributes as split nodes in the decision tree. Furthermore, Bolón-Canedo, Sánchez-Maroño and Alonso-Betanzos (2013), Dash and Liu (1997), Kudo and Sklansky (2000), Molina, Belanche and Nebot (2002) have provided an extensive overview and comparison of different approaches to the feature selection problems.

The original purpose of features subset selection is to reduce the number of attributes to only those that are relevant for a certain data mining task as presented and proved in Olanweraju, Aburas, Omran and Abdalla (2010), and Yusof, Paulraj and Yaacob (2008). They nevertheless can be utilized to measure the interestingness of rules/pattern generated. For example, if the rules/pattern themselves consist of irrelevant attributes, the aforementioned measure can also give some indication that the rules/pattern is not interesting (Shaharanee & Hadzic, 2013). Since frequent patterns are generated based solely on frequency without considering their predictive power, the use of frequent patterns without selecting appropriate features will still result in a huge feature space which leads to larger volume and complexity of rules. This might not only slow down the model learning process, but even worse, the classification accuracy deteriorates (Cheng et al., 2007).

In the datasets where there is a predefined class label (i.e. classification tasks), a structural associative classification can contribute to discovering strong associations between occurring attribute and class values (Li, Shen, &

Topor, 2002). A combination of a frequent-pattern-based framework with a feature selection algorithm, namely the Maximal Marginal Relevance Feature Selection (MMRFS) as proposed by Cheng et al. (2007), is a discriminative frequent-pattern-based classification that is capable of overcoming the over fitting in a classification problem and has proven to be scalable and highly accurate. The aforementioned framework involves a two-step process, firstly to mine the frequent pattern, and secondly, to perform feature selection or rule ranking. An improvement of this work that is capable of directly mining the discriminative pattern is proposed by Cheng, Yan, Han and Yu (2008). This approach, namely the Direct Discriminative Pattern Mining (DDPMine), is capable of discovering classification rules by incorporating the feature selection method into the mining framework by directly mining the most discriminative patterns, and then incrementally eliminating the training instances which are covered by those patterns.

The problem focused in this research work is the evaluation of feature selection technique for association rules based classification. Each of the aforementioned techniques has their own strengths and weaknesses (Hilderman & Hamilton, 2001). While (Cheng et al., 2007) and (Cheng et al., 2008) has examined the latest developments in the feature selection for structural associative classification. Nevertheless, an understanding of the various implications including issues concerning the optimal cut-off value between features with the datasets, selection of a suitable feature selection and assessment of the rules performances will ensure that one will arrive at a more reliable and interesting set of rules. The common properties and advantages of using the feature selection in determining the relevant attribute for association rules based classification is given in the next section.

FEATURE SUBSET SELECTION TO DETERMINE RELEVANT ATTRIBUTES

The feature subset selection problem to be addressed in this work can be more formally described as follows: Given a relational database D, with $AT = AT = \{at_1, at_2, ..., at_{|AT|}\}$ as the set of input attributes in D, and $Y = \{y_1, y_2, ..., y_{|Y|}\}$ the class attribute with a set of class labels in. The feature subset selection is utilized in this work to reduce the size of frequent rules, thus reducing the complexity of the rules.

Let an association rule mining algorithm be denoted as AR_{AL}. This set of association rules is used for predicting the value of a class attribute Y from D extracted using AR_{AL} as $AR(D)$, and accuracy of $AR(D)$ as $ac(AR(D))$. The

problem of feature subset selection is to reduce D into D' such that $AT \leq AT$ and $ac(AR(D') \leq ac(AR(D') - \varepsilon$, where ε is an arbitrary user defined as small value to reflect noise present in real-world data. In other words, the task is to find the optimal set of attributes, $AT_{OPT} \leq AT$, such that the accuracy of the association rule set using AR_{AL} is maximized.

Feature subset selection is an important pre-processing step in the data mining process. The feature subset selection task is utilized in this research purposely to determine irrelevant attributes in predicting the class variable. The removal of these attributes will result in a much smaller dataset, thereby reducing the number of rules that need to be generated from the association rule mining algorithm, while closely maintaining the integrity of the original data (Han et al., 2011). Additionally, rules described with fewer attributes are also expected to perform better when classifying future cases; hence, they will have better generalization power than do the more specific rules that take many attributes into account. Besides, the patterns extracted will also be simpler and easier to analyze and understand.

FEATURE SUBSET SELECTION UTILIZATION

This section is devoted on describing the feature subset selection process by applying the Symmetrical Tau and comparing its results with those obtained by the Mutual Information approaches. Based on the proposed framework, this feature subset selection is needed in order to determine the relevance of attributes by classifying their importance to characterize an association. Both techniques are capable of measuring the capability of an input attribute in predicting the class of another attribute. This step is defined as follows:

Determine the relevance of each at_i by determining its importance in predicting the value of the class attribute Y in D_{tr}, where $at_i \in AT$, (i=(1,....,$|AT|$)) using a statistical-heuristic measure. Any irrelevant attributes are removed from the dataset and are represented in the filtered database as D_{tr} , $I \subseteq I$.

Symmetrical Tau Utilization

Let there be R rows and C columns in the contingency table for attributes at_i and Y. The probability that an individual belongs to row category r and column category c is represented as $P(rc)$, and $P(r+)$ and $P(+c)$ are the marginal probabilities in row category r and column category c respectively. The measure is based on the probability of one attribute value occurring together with the value of the second attribute. In this sense, the Y attribute can be seen

as a representative of the class attribute, and the Symmetrical Tau measure for the capability of input attribute at_i in predicting the class attribute Y is defined by Zhou and Dillon (1991) as follows:

$$\tau(at_i, Y) = \frac{\sum_{c=1}^{C}\sum_{r=1}^{R}\frac{P(rc)^2}{P(+c)} + \sum_{r=1}^{R}\sum_{c=1}^{C}\frac{P(rc)^2}{P(r+)} - \sum_{r=1}^{R}P(r+)^2 - \sum_{c=1}^{C}P(+c)^2}{2 - \sum_{r=1}^{R}P(r+)^2 - \sum_{c=1}^{C}P(+c)^2} \quad (1)$$

Higher values of the Symmetrical Tau measure would indicate better discriminating criteria (feature) for the class that is to be predicted in the domain. Symmetrical Tau has many more desirable properties in comparison to other feature subset selection techniques, as reported in Zhou and Dillon (1991). It is utilized here to indicate the relative usefulness of attributes in predicting the value of the class attribute, and to discard any of the attributes whose relevance value is fairly low. This would prevent the generation of rules which then would need to be discarded anyway once it was found that they include irrelevant attributes.

Mutual Information

In this research, the capabilities of Symmetrical Tau as the determinant of the relevance of attributes are evaluated by comparing it with an information-theoretic measure, namely the Mutual Information. The information-theoretic measures are principally comprehensible and useful since they can be interpreted in terms of information. For a rule interestingness measure, the relation is interesting when the antecedent provides a great deal of information about the consequent (Blanchard et al., 2005). Although several information-theoretic measures exist (Peng, Long, & Ding, 2005), the Symmetrical Tau is only compared with the Mutual Information measurement technique which is the most well-known among these techniques. The Mutual Information measure (Ke, Cheng, & Ng, 2008; Tan, Kumar, & Srivastava, 2002) is calculated based on the following formula:

$$M(at_i, Y) = \frac{\sum_r \sum_c P(at_i, Y)\log\frac{P(at_i, Y)}{P(at_i)P(Y)}}{\min(-\sum_r P(at_i)\log P(at_i) - \sum_c P(Y)\log P(Y))} \quad (2)$$

The information that at_i gives us about Y is the reduction in uncertainty about Y due to knowledge of at_i and similarly for the information that Y tells about at_i. The greater the values of M, the more information at_i and Y contain about each other (Ke et al., 2008).

FEATURE SUBSET SELECTION PROCESS AND COMPARISON OF SYMMETRICAL TAU (ST) AND MUTUAL INFORMATION (MI)

The comparison of Symmetrical Tau (ST) and Mutual Information (MI) for feature selection process is performed using the Wine, Mushroom, Iris and Adult datasets obtained from the UCI Machine Learning Repository (Asuncion & 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. For all continuous attributes in the Adult, Iris and Wine datasets, we apply an equal depth binning approach method. This equal depth binning approach will ensure we have manageable data sizes by reducing the number of distinct values per attribute (Han et al., 2011). Other discrete attributes in the Adult and Mushroom datasets are preserved in their original state. The selected attributes are measured according to their capabilities in predicting the values of attribute class in each dataset.

ST and MI are capable of measuring the relevance of attributes in predicting a class value, but they are different from each other in terms of their approach as aforementioned in Shaharanee, Hadzic and Dillon (2011) and Shaharanee and Hadzic (2013). They can both be used as a means of selecting a feature subset to be used for rule generation, and in this section the two approaches are compared in terms of their general properties and utilization for the feature subset selection process. At the end of the section, the feature subsets used for each of the datasets considered in the experimental evaluation is indicated.

The ST and MI measures for all the attributes in the Mushroom, Adult, Wine and Iris datasets are shown in Table 1, 2, 3 and 4 respectively (Shaharanee & Jamil, 2013). The attributes were ranked according to their decreasing ST and MI values. Based on the experiment with the Adult dataset, the MI approach seems to favor variables with more values. This can be observed in Table 1 for the Adult dataset as variables with more values have all been ranked in the top 7 based on the MI measure (i.e. Education(16), Occupation(14), Education Number(8), Age(10) and Hour Per Week(10)), while each one of these is ranked lower based on ST, with attribute Capital Gain(6) occurring higher than all these attributes with more values. Similarly, for the Mushroom dataset, variables with more values such as Gcolor(12), Scabovering(9), Scbelowring(9), are all ranked higher based on MI in contrast to ST ranking. For example, the ST measure has ranked the attribute Gsize with only two values as third in the ranking, higher than all these multi-valued attributes, whereas in the MI ranking the Gsize is seventh in the ranking after all those multi-valued attributes.

Table 1

Comparison between ST and MI for Adult Dataset

# of Values	Variables	ST Values	# of Values	Variables	MI Values
7	Marital Status	0.14	6	Relationships	0.17
6	Relationship	0.12	7	Marital Status	0.16
6	Capital Gain	0.07	16	Education	0.09
8	Education Number	0.07	14	Occupation	0.09
16	Education	0.05	8	Education Number	0.09
2	Sex	0.05	10	Age	0.08
14	Occupation	0.05	10	Hours Per Week	0.05
10	Age	0.04	6	Capital Gain	0.05
5	Capital Loss	0.04	2	Sex	0.04
10	Hours Per Week	0.03	5	Capital Loss	0.02
7	Work Class	0.01	7	Work Class	0.02
5	Race	0.01	41	Native Country	0.01
41	Native Country	0.01	5	Race	0.01
10	FNLWGT	0.00	10	FNLWGT	0.00

Table 2

Comparison between ST and MI for Mushroom Dataset

# of Values	Variables	ST Values	# of Values	Variables	MI Values
9	Odor	0.59	9	Odor	0.91
9	SporePrintColor	0.32	9	SporePrintColor	0.48
2	Gsize	0.29	12	Gcolor	0.41
5	Ringtype	0.24	5	Ringtype	0.32
2	Bruises	0.24	9	Scabovering	0.25
12	Gcolor	0.22	9	Scbelowring	0.24
9	Scabovering	0.15	2	Gsize	0.23
6	Pop	0.15	6	Pop	0.20
9	Scbelowring	0.14	2	Bruises	0.19
2	Gspacing	0.13	7	Habitat	0.16
7	Habitat	0.09	2	Gspacing	0.11
3	Ringnumber	0.05	6	Cshape	0.05
4	Sroot	0.04	3	Ringnumber	0.04

(continued)

# of Values	Variables	ST Values	# of Values	Variables	MI Values
6	Cshape	0.02	4	Sroot	0.04
4	Csurface	0.02	10	Ccolor	0.03
10	Ccolor	0.02	4	Csurface	0.02
4	Veilcolor	0.02	4	Veilcolor	0.02
4	Ssabovering	0.01	4	Ssbelowring	0.01
4	Ssbelowring	0.01	4	Ssabovering	0.01
2	Sshape	0.01	2	Gattachment	0.01
2	Gattachment	0.01	2	Sshape	0.01
1	Veiltype	0.00	1	Veiltype	0.00

Table 3

Comparison between ST and MI for Wine Dataset

# of Values	Variables	ST Values	# of Values	Variables	MI Values
5	Flavanoids	0.48	5	Flavanoids	0.88
5	Color	0.42	5	Diluted	0.85
5	Diluted	0.36	5	Color	0.79
5	Proline	0.35	5	Proline	0.74
5	Hue	0.30	5	Hue	0.62
5	Alcohol	0.24	5	Phenols	0.56
5	Phenols	0.23	5	Alcohol	0.53
5	Magnesium	0.18	5	Magnesium	0.37
5	Alcalinity	0.17	5	Proanthocyanins	0.33
5	Proanthocyanins	0.15	5	Alcalinity	0.31
5	Malidacid	0.14	5	Malidacid	0.28
5	Nonflavanoids	0.13	5	Nonflavanoids	0.27
5	Ash	0.05	5	Ash	0.09

Table 4

Comparison between ST and MI for Iris Dataset

# of Values	Variables	ST Values	# of Values	Variables	MI Values
5	Petal Width	0.67	5	Petal Width	1.31
5	Petal Length	0.64	5	Petal Length	1.22
5	Sepal Length	0.27	5	Sepal Length	0.61
5	Sepal Width	0.23	5	Sepal Width	0.50

This observation of MI preference for multi-valued attributes is in accord with Blanchard et al. (2005). In contrast, the procedure based on ST produces a more stable selection of variables which does not favor the multi-valued nature of attributes. This is in agreement with the claim by Zhou and Dillon (1991) that ST is fair in handling multi-valued variables. However, the question still remains of how the ST and MI methods compare to each other when used for the purpose of feature subset selection. When using an attribute relevance measure for the feature subset selection problem, commonly a relevance cut-off point is chosen below which all attributes are removed. Hence, in the ranking of attributes according to their decreasing ST and MI values in Tables 1- 4, a relevance cut-off needs to be set. Here, the cut-off point was selected based on the significant difference between the ST and MI values in decreasing order. The significant difference was considered to occur in the ranking at the position where attribute's ST and MI value is less than half of the previous attribute's ST and MI value in the ranking, respectively. At this point and below in the ranking, all attributes are considered as irrelevant. In Tables 1 - 4, all the attributes that are considered as irrelevant based on this way of determining the cut-off value are shaded gray. As one can see, the way in which feature subsets would be selected based on ST and MI measures differs for the Adult dataset only. Hence, the performance of these two subsets when used for generating association rules for classification purposes will be evaluated next. Additionally, in the Iris dataset (Table 4), all input variables were considered in the experiments, as Iris dataset consists of only 4 attributes, and complexity problems would not occur.

The comparison result for the Adult dataset are shown in Table 1, where the capabilities of attributes in predicting the values of attribute Income (<=50K and >50K) are measured. For the Adult dataset results presented in Table 1, the relevance cut-off value is 0.01. This is due to the ST value of attribute Hours Per Week being more than double the ST value for attribute Work class. Thus, the subset of data now consists of 10 attributes: Marital Status, Relationship, Capital Gain, Education Number, Education, Sex, Occupation, Age, Capital Loss and Hours Per Week. Similarly for the Mushroom dataset in Table 2, the subset of data after the feature subsets selection process consists of 11 attributes: Odor, SporePrintColor, Gsize, Ringtype, Bruises, Gcolor, Pop, Scabovering, Scbelowring, Gspacing and Habitat. For the Wine dataset (Table 3), only the Ash input variable has been discarded from further analysis.

For the Adult dataset, by ranking the attributes based on ST values, 10 input attributes are selected based on the aforementioned way of determining the cut-off value, while 13 input attributes are favored based on MI ranking. The

cut-off point at and below of which all attributes are considered as irrelevant is shown in Table 1, where cells of attributes removed are shaded gray. Rules are then generated based on these 10 and 13 input variables and evaluated for their accuracy and coverage rate. Accuracy rate (AR) is typically defined as the number of correctly classified instances. Additionally, coverage rate (CR) refers to the percentage of captured/covered instances from the database. Thus, our aim is to evaluate these extracted rules in terms of correctly predicting the class value from the training datasets and correctly predicting the class value from the testing/unseen dataset. They are also evaluated for their coverage rate on both training and testing datasets. As depicted in Table 5, for this dataset, the selection of 10 input attributes that were ranked based on ST resulted in 303 rules in comparison to 1726 rules when they were ranked by MI. This was not at the cost of a reduction in coverage rate; moreover, accuracy was slightly better for both the training and testing datasets.

Table 5

Rules Evaluation between Attributes Selected Based on ST and MI for Adult Dataset

	Data Partition	Symmetrical Tau (ST)			Mutual Information (MI)		
		# Of Rules	AR %	CR%	# Of Rules	AR %	CR %
Initial # of Rules	Training	2192	68.98	100.00	2192	68.98	100.00
	Testing		69.05	100.00		69.05	100.00
Rule # from feature subset	Training	303	67.46	100.00	1726	67.36	100.00
	Testing		67.45	100.00		67.38	100.00

As shown in the experiments, the ST has more advantageous properties in comparison to MI, as the feature subset selected according to the ST measure resulted in fewer rules with slightly higher accuracy at the same coverage rate of 100%. In addition, from the ranking of the different attributes relevance measures, it was shown that MI tends to favor multi-valued attributes in comparison to ST. Given these observation as well as others' claims (Shaharanee & Hadzic, 2013; Zhou & Dillon, 1991) in regards to the advantageous properties of ST over other existing measures, the ST feature selection criterion was used within the framework as the first step to remove any irrelevant attributes. This would prevent the generation of rules that include any irrelevant attributes. Hence, in the experiments, it is not necessary to use ST to further verify the rules as the rules were created from the attribute subset considered as relevant according to the measure.

CONCLUSION

This paper has presented an empirical analysis of the usefulness and implication behind using feature subset selection prior to association rule generation with respect to their classification accuracy and coverage rate. Datasets of varying complexity were used and the characteristics of different feature subset selection techniques were investigated. The experimental results show that, pre-processing tasks are essential in guaranteeing a good data mining model especially for the association rule mining. The advantage of selecting certain features for inclusion in the association rule mining process is that only those attributes relevant to the classification task at hand are used to generate the association rules. Furthermore, the result reveals that there is a strong case for applying a feature subset selection process prior to the association rule generation, which in itself significantly reduces the rule quantity and complexity of the task. This will reduce a large number of irrelevant rules while at the same time preserving relatively high accuracy rate and capable of retaining the rule coverage rate. As part of our ongoing work, we intend to extend the works by utilizing the chosen feature subset selection techniques towards more complex datasets discovered from semi-structured data.

REFERENCES

Asuncion, A., & Newman, D. J. (2007). UCI Machine Learning Repository. *University of California Irvine School of Information*. University of California, Irvine, School of Information and Computer Sciences. Retrieved from http://www.ics.uci.edu/~mlearn/MLRepository.html

Bayardo Jr., R. J., & Agrawal, R. (1999). Mining the Most Interesting Rules. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 145–154). New York, NY, USA: ACM.

Blanchard, J., Guillet, F., Gras, R., & Briand, H. (2005). *Using Information-Theoretic Measures to Assess Association Rule Interestingness*. In Proceedings of the Fifth IEEE International Conference on Data Mining. IEEE Computer Society.

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. *Knowledge and Information Systems*, *34*(3), 483–519. doi:10.1007/s10115-012-0487-8

Brin, S., Motwani, R., & Silverstein, C. (1997). *Beyond market baskets: generalizing association rules to correlations*. In Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data. Tucson, Arizona, United States: ACM.

Cheng, H., Yan, X., Han, J., & Hsu, C.-W. (2007). Discriminative frequent pattern analysis for effective classification. In *Proceeding of the 23rd International IEEE Conference on Data Engineering* (pp. 716-725).

Cheng, H., Yan, X., Han, J., & S. Yu, P. (2008). Direct discriminative pattern mining for effective classification, In *Proceeding of the 24th International IEEE Conference on Data Engineering* (pp. 67-178).

Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis*, *1*(1-4), 131–156. doi:10.1016/S1088-467X(97)00008-5

Geng, L., & Hamilton, H. J. (2006). Interestingness measures for data mining. *ACM Computing Surveys*, *38*(3), 9. doi:10.1145/1132960.1132963

Hadzic, F., & Dillon, T. S. (2006). *Using the Symmetrical Tau (t) criterion for feature selection in decision tree and neural network learning*. In Proceedings of the SIAM 2nd Workshop on Feature Selection for Data Mining: Interfacing Machine Learning and Statistics.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann Publishers.

Hilderman, R., & Hamilton, H. (2001). *Evaluation of interestingness measures for ranking discovered knowledge*. In Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining: Springer-Verlag.

Jaroszewicz, S., & Simovici, D. (2001). A general measure of rule interestingness. In L. Raedt & A. Siebes (Eds.), *Principles of data mining and knowledge discovery SE-21* (Vol. 2168, pp. 253–265). Springer Berlin Heidelberg.

Ke, Y., Cheng, J., & Ng, W. (2008). An information-theoretic approach to quantitative association rule mining. *Knowledge and Information Systems*, *16*(2), 213–244. doi:10.1007/s10115-007-0104-4

Kudo, M., & Sklansky, J. (2000). Comparison of algorithms that select features for pattern classifiers. *Pattern Recognition*. doi:10.1016/S0031-3203(99)00041-2

Li, J., Shen, H., & Topor, R. (2002). Mining the optimal class association rule set. *Knowledge-Based Systems*, *15*, 399–405. doi:10.1016/S0950-7051(02)00024-2

Molina, L. C., Belanche, L., & Nebot, A. (2002). Feature selection algorithms: A survey and experimental evaluation. In *Proceding of IEEE International Conference on Data Mining (ICDM '02)* (pp. 306–313).

Olanweraju, R. F., Aburas, A. A., Omran, O. K., & Abdalla, A.-H. H. (2010). Damageless Digital Watermarking using Complex Valued Artificial Neural Network. *Journal of Information and Communication Technology*, 9, 111–137.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*, 1226–1238. doi:10.1109/TPAMI.2005.159

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106.

Quinlan, J. R. (1993). *C4. 5: Programs for machine learning* (Vol. 1). Morgan kaufmann.

Roiger, R. J., & Geatz, M. W. (2003). *Data Mining: A Tutorial-Based Primer*. Addison Wesley.

Shaharanee, I. N. M., & Hadzic, F. (2013). Evaluation and optimization of frequent, closed and maximal association rule based classification. *Statistics and Computing*. doi:10.1007/s11222-013-9404-6

Shaharanee, I. N. M., Hadzic, F., & Dillon, T. S. (2011). Interestingness measures for association rules based on statistical validity. *Knowledge-Based Systems*, *24*(3), 386–392. doi:10.1016/j.knosys. 2010.11.005

Shaharanee, I. N. M., & Jamil, J. M. (2013). Features selection and rule Removal for frequent Association rule based classification. In *Proceedings of the 4th International Conference on Computing and Informatic* (pp. 377–382).

Smyth, P., & Goodman, R. M. (1992). An information theoretic approach to rule induction from databases. *IEEE Transactions on Knowledge and Data Engineering*, *4*(4), 301–316. doi:10.1109/69.149926

Tan, P. N., Kumar, V., & Srivastava, J. (2002). Selecting the right interestingness measure for association patterns. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Edmonton, Alberta, Canada: ACM.

Tan, P.-N., Steinbach, M., & Kumar, V. (2014). *Introduction to data mining*. Essex: Pearson Education Limited.

Yusof, S. A. M., Paulraj, M., & Yaacob, S. (2008). Classification of Malaysian vowels using formant based features. *Journal of Information and Communication Technology*, *7*, 27–40.

Zhou, X. J., & Dillon, T. S. *A statistical-heuristic feature selection criterion for decision tree induction*. 13 IEEE Transactions on Pattern Analysis and Machine Intelligence 834–841 (1991). IEEE Computer Society. doi:10.1109/34.85676