FEATURES SELECTION AND RULE REMOVAL FOR FREQUENT ASSOCIATION RULE BASED CLASSIFICATION

Izwan Nizal Mohd Shaharanee[1] and Jastini Jamil[1]

[1]*School of Quantitative Sciences, Universiti Utara Malaysia, Malaysia, {nizal.jastini@uum.edu.my}*

ABSTRACT. The performance of association rule based classification is notably deteriorated with the existence of irrelevant and redundant features and complex attributes. Association rules naturally often suffer from a large volume of rules generated, many of which are not interesting and useful. Thus, selecting relevant feature and/or removing unrelated rules can significantly improve the association rule performance. In this paper, we explored and compared feature selection measures to filter out irrelevant and redundant features prior to association rules generation. Rules that encompassed with irrelevant/redundant features were removed. Based on the experimental results, removing rules that hold irrelevant features slightly improve the accuracy rate and capable to retain the rule coverage rate.

Keywords: features selection, rules removal, frequent item set mining

INTRODUCTION

The number of patterns/association rules generated through frequent item sets mining can be quite large, while usefulness of each rule for the classification/prediction task may be limited. 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, creating problems such as over fitting (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.

RELATED WORKS

The feature subset selection as describes in (Han & Kamber, 2001) is a ways 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 & Kamber, 2001) 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". Han and Kamber in (Han & Kamber, 2001) asserted that, domain expertise can be employed in order to pick up useful attributes. However, because the data mining task involves a large volume of data and unpredictable behavior of data during data mining, this task is often expensive and time consuming.

The test of statistical significance is one of the prominent approaches in evaluating attributes/features usefulness. Stepwise forward selection, stepwise backward selection and a combination of both are three commonly used heuristic techniques utilized in statistical significance tests such as linear regression and logistic regression (Han & Kamber, 2001). Moreover, the application of correlation analysis such as the chi-squared test is also valuable in identifying redundant variables for features subset selection. Another powerful technique for this purpose is the Symmetrical Tau (Zhou & Dillon, 1991), which is a statistical-heuristic feature selection criterion. It measures the capability of an attribute in predicting the class of another attribute. Additionally, information gain is another attributes' relevance analysis method employed in the popular ID3 (J Ross Quinlan, 1986) and C4.5 (John Ross Quinlan, 1993) as reported in (Han & Kamber, 2001), for selecting the most prominent class distinguishing attributes as split nodes in the decision tree.

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, $AT = \{at_1, at_2, ..., at_{|AT|}\}$ the set of input attributes in D, and $Y = \{y_1, y_2, ..., y_{|Y|}\}$ the class attribute with a set of class labels in D. Let an association rule mining algorithm be denoted as AR_{AL}, the set of association rules 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 $AT' \subseteq AT$ and $ac(AR(D') \geq ac(AR(D') - \varepsilon$, where ε is an arbitrary user defined small value to reflect noise present in real-world data. In other words, the task is to find the optimal set of attributes, $AT_{OPT} \subseteq AT$, such that the accuracy of the association rule set using AR_{AL} is maximized.

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, real-world datasets of varying complexity obtained from the UCI Machine Learning Repository. 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 & Kamber, 2001). Other discrete attributes in the Adult and Mushroom datasets were 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, 2012). 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. 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 PerWeek (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 2 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	0.1448		6	Relationships	0.1662
6	Relationship	0.1206		7	Marital Status	0.1575
6	Capital Gain	0.0706		16	Education	0.0934
8	Education Number	0.0688		14	Occupation	0.0932
16	Education	0.0528		8	Education Number	0.0900
2	Sex	0.0470		10	Age	0.0894
14	Occupation	0.0469		10	Hours Per Week	0.0545
10	Age	0.0432		6	Capital Gain	0.0475
5	Capital Loss	0.0361		2	Sex	0.0374
10	Hours Per Week	0.0354		5	Capital Loss	0.0238
7	Work Class	0.0166		7	Work Class	0.0171
5	Race	0.0085		41	Native Country	0.0093
41	Native Country	0.0077		5	Race	0.0083
10	FNLWGT	0.0002		10	FNLWGT	0.0002

Table 2. Comparison between ST and MI for Mushroom Dataset

Feature Subset Selection Based on ST				Feature Subset Selection Based o MI		
# of Values	Variables	ST Values		# of Values	Variables	MI Values
9	Odor	0.5872		9	Odor	0.9127
9	SporePrintColor	0.3246		9	SporePrintColor	0.4812
2	Gsize	0.2866		12	Gcolor	0.4078
5	Ringtype	0.2585		5	Ringtype	0.3172
2	Bruises	0.2487		9	Scabovering	0.251
12	Gcolor	0.2172		9	Scbelowring	0.2404
9	Scabovering	0.1462		2	Gsize	0.2271
6	Pop	0.1454		6	Pop	0.197
9	Scbelowring	0.1405		2	Bruises	0.1897
2	Gspacing	0.1298		7	Habitat	0.1578
7	Habitat	0.0980		2	Gspacing	0.1088
3	Ringnumber	0.0460		6	Cshape	0.0487
4	Sroot	0.0439		3	Ringnumber	0.0409
6	Cshape	0.0299		4	Sroot	0.0402
4	Csurface	0.0234		10	Ccolor	0.0356
10	Ccolor	0.0227		4	Csurface	0.0249

4	Veilcolor	0.0214		4	Veilcolor	0.0222
4	Ssabovering	0.0169		4	Ssbelowring	0.0166
4	Ssbelowring	0.0150		4	Ssabovering	0.0163
2	Sshape	0.0150		2	Gattachment	0.0122
2	Gattachment	0.0146		2	Sshape	0.0108
1	Veiltype	0.0000		1	Veiltype	0.0000

Table 3. Comparison between ST and MI for Wine Dataset

# of Values	Variables	ST Values		# of Values	Variables	MI Values
5	Flavanoids	0.4810		5	Flavanoids	0.8796
5	Color	0.4226		5	Diluted	0.8476
5	Diluted	0.3610		5	Color	0.7914
5	Proline	0.3543		5	Proline	0.7422
5	Hue	0.3019		5	Hue	0.6242
5	Alcohol	0.2367		5	Phenols	0.5591
5	Phenols	0.2312		5	Alcohol	0.5275
5	Magnesium	0.1840		5	Magnesium	0.3717
5	Alcalinity	0.1680		5	Proanthocyanins	0.3275
5	Proanthocyanins	0.1525		5	Alcalinity	0.3143
5	Malidacid	0.1403		5	Malidacid	0.2821
5	Nonflavanoids	0.1313		5	Nonflavanoids	0.2730
5	Ash	0.0499		5	Ash	0.0996

Table 4. Comparison between ST and MI for Iris Dataset

# of Values	Variables	ST Values		# of Values	Variables	MI Values
5	Petal Width	0.6738		5	Petal Width	1.311
5	Petal Length	0.6355		5	Petal Length	1.226
5	Sepal Length	0.2724		5	Sepal Length	0.618
5	Sepal Width	0.2301		5	Sepal Width	0.508

This observation of MI preference for multi-valued attributes is in accord with that of (Julien, Fabrice, Regis, & Henri, 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 & Dillon, 1991) that ST is fair in handling multi-valued variables. However, the question still remains of how the ST and MI methods compare 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/MI values in decreasing order. The significant difference was considered to occur in the ranking at the position where that attribute's ST/MI value is less than half of the previous attribute's ST/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.

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 which all attributes are considered as irrelevant, is shown in Table 1, where cells of attributes removed are shaded gray. For example, the comparison results 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 6.2, the relevance cut-off value is 0.0166. 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.

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

CONCLUSION

As shown in the experiment section for the Adult dataset, the ST has more advantageous properties in comparison with MI, as the feature subset selected according to the ST measure, resulted in many less rules which at the same time had a slightly higher accuracy and the same coverage rate of 100%. In addition, from the ranking of the different attributes relevance measures (i.e. Tables 1 - 4), it was shown that MI tends to favor multi-valued attributes in comparison to ST. Given these observation as well as others' claims (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 in order 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.

REFERENCES

Cheng, H., Yan, X., Han, J., & Hsu, C.-W. (2007). Discriminative frequent pattern analysis for effective classification (p. 10 pp.). Pisctaway, NJ, USA: IEEE.

Han, J., & Kamber, M. (2001). *Data mining : concepts and techniques*. *The Morgan Kaufmann series in data management systems* (p. xxiv, 550 p.). San Francisco: Morgan Kaufmann.

Julien, B., Fabrice, G., Regis, G., & Henri, B. (2005). Using Information-Theoretic Measures to Assess Association Rule Interestingness. *Proceedings of the Fifth IEEE International Conference on Data Mining*. IEEE Computer Society. doi:http://dx.doi.org/10.1109/ICDM.2005.149

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

Quinlan, John Ross. (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. (2012). *Quality and Interestingness of Association Rules Derived from Data Mining of Relational and Semi-structured Data*. Curtin University.

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