A Conflict-Based Confidence Measure for Associative Classification

Peerapon Vateekul and Mei-Ling Shyu
Department of Electrical and Computer Engineering
University of Miami
Coral Gables, FL 33124, USA
E-mail: p.vateekul@umiami.edu, shyu@miami.edu

Abstract

Associative classification has aroused significant attention recently and achieved promising results. In the rule ranking process, the confidence measure is usually used to sort the class association rules (CARs). However, it may be not good enough for a classification task due to a low discrimination power to instances in the other classes. In this paper, we propose a novel conflict-based confidence measure with an interleaving ranking strategy for re-ranking CARs in an associative classification framework, which better captures the conflict between a rule and a training data instance. In the experiments, the traditional confidence measure and our proposed conflict-based confidence measure with the interleaving ranking strategy are applied as the primary sorting criterion for CARs. The experimental results show that the proposed associative classification framework achieves promising classification accuracy with the use of the conflict-based confidence measure, particularly for an imbalanced data set.

1. Introduction

In recent years, extensive research has been carried out to integrate classification and association rule discovery [1, 2], which are two of the most important research areas in data mining [4, 8, 9, 11]. Such an integrated approach is called *Associative Classification (AC)* that can produce more efficient and accurate classifiers than some of the traditional classification techniques. Moreover, the generated classifiers in the form of class association rules (CARs), whose consequent part is a class label, are more comprehensive than some statistical classifiers such as Naïve BaYes. Some classifiers based on AC have been proposed such as CAEP [3], CMAR [5], CBA [6], and ADT [12]. These algorithms rank the generated CARs using a confidence measure.

According to the majority of AC algorithms, the rule ranking process plays an important role in the classification process since the accuracy is affected directly to the order of CARs. However, the confidence measure has some limitations [10]. Thus, in the past few years, various measures have been proposed such as Interest (*I*, also known as *lift*), Conviction (*V*), Correlation (ϕ), Cosine (*IS*), and etc. [10]. However, all of these measures still have some drawbacks and are mainly designed for association rule discovery task, not for classification tasks. Furthermore, for a classification task, the confidence measure may be not good enough due to a low discrimination power to the instances in the other classes. The reason is that a confidence measure is defined using a frequency count of the *exact matched instances* on a training data set. On a space of distribution, it is possible that a high confidence rule can be close to many instances in different classes. Thus, this rule has a possibility to misclassify instances in different classes on a testing data set.

In this paper, we propose a novel conflict-based confidence measure for ranking CARs in an associative classification framework, which better captures the conflict between a rule and a training instance. It quantifies the amount of conflict that a rule possesses with respect to all instances belonging to different classes in the data set. Moreover, we also propose an *interleaving* ranking strategy that can improve performance of CARs on both balance and imbalance data sets.

This paper is organized as follows. In Section 2, the associative classification approach is discussed. The overview of our proposed framework is described in Section 3. Then, we show the details of our proposed conflict-based confidence measure in Section 4. In Section 5, we present experimental results. Finally, we conclude our study in Section 6.

2. Associative Classification

Associative classification (AC) is a data mining technique that integrates classification with association rule mining (ARM) to find the rules from classification benchmarks. A generated rule for classification, called "class association rule" (CAR), is an implication of the form of $X \rightarrow c$, where itemset X is non-empty subset of all possible items in the database, $X \subseteq I$, $I=\{i_1, i_2,..., i_n\}$ where n is the number of itemsets, and c is a class identifier, $c \in \{c_1, c_2,..., c_m\}$ where m is the number of classes. Let a rule itemset be a pair $<X, c>$, containing the itemset X and a class label c. The rules are discovered in a training data set of transactions D_t. The strength of a rule R can be measured in terms of its support (*sup(R)*) and confidence (*conf(R)*). The support of R is the percentage of the instances in D_t satisfying the rule antecedent and having class label c as shown in Equation 1. The confidence of R is the percentage of instances in D_t satisfying the rule antecedent that also have the class label c as shown in Equation 2.

$$\sup(R) = P(X, c)\ ; \qquad (1)$$

$$conf(R) = \frac{\sup(R)}{\sup(X)} = \frac{P(X,c)}{P(X)} = P(c \mid X)\ . \qquad (2)$$

Referring to the support and confidence constraints, the rule $X \rightarrow c$ is a class association rule if the following two conditions are satisfied: $sup(X \rightarrow c) \geq minsup$ (minimum support threshold) and $conf(X \rightarrow c) \geq minconf$ (minimum confidence threshold). To find all CARs, many algorithms are commonly decomposed into three major processes: rule generation, rule selection, and classification. First, the rule generation process extracts all CARs that satisfy both minimum support and minimum confidence thresholds from the training data set. Second, the rule selection process applies the pruning techniques to select a small subset of high-quality CARs and builds an accurate model of the training data set. Finally, the classification process is used to classify an unseen data instance.

3. The Proposed Interleaving Conflict-Based Associative Classification Framework

Figure 1 shows the system architecture of our proposed framework. The main contribution of our proposed framework is that the proposed conflict-based confidence measure is used to rank the generated rules instead of the traditional confidence measure. We will describe the concept and calculation of our proposed measure in Section 4. As can be seen from Figure 1, our proposed framework consists of four modules, namely Rule Generator, Rule Re-ranking, Model Evaluation, and Classification.

The rule generator module generates the class association rules (CARs) from the training data. Based on the original Apriori algorithm [1], the generated CARs depend on the minimum support and minimum confidence thresholds. One of our main objectives is to demonstrate that the CARs ranked by the conflict-based confidence measure improve the classification accuracy over those ranked by the traditional confidence measure. We, therefore, aim to generate all possible rules by setting a low minimum support (5% in this study). In order to compare the ranking method, the minimum confidence threshold is set to 0%

In the rule re-ranking module, the whole set of generated CARs are sorted based on our proposed conflict-based confidence measure instead of the traditional confidence measure. In addition, the interleaving strategy is adopted, which gives a better result in both balanced and imbalanced data sets. This is motivated by the observations that the number of rules for the majority class (e.g., negative class) is usually much larger than the number of rules for the other class (e.g., positive class) for an imbalanced data set, and the rules for the negative class are most likely on the top of the sorted rule set. Consequently, classification accuracy of the positive class drops due to misclassifications. Our empirical studies showed that the interleaving ordered rule set achieves the better performance than the traditional ordered rule set. The interleaving strategy is as follows. First, we categorize the generated CARs by each class. Second, for each class, we sort the generated CARs based on our proposed conflict-based confidence measure. Finally, we interleave the CARs for each class that have the same rank to create the final rule set.

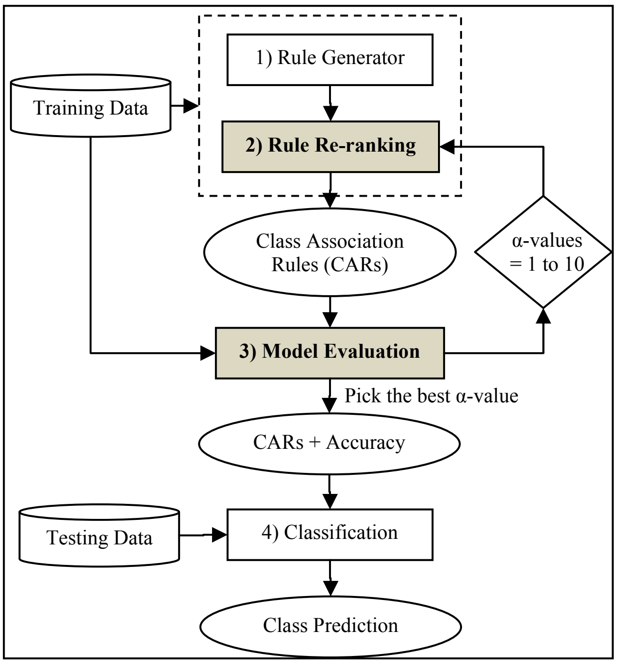

Figure 1. System architecture of the proposed interleaving conflict-based associative classification framework

As will be explained in Section 4.1, the conflict-based confidence measure directly depends on the α-value (*called the closeness threshold*). The model evaluation module aims to adaptively find the most suitable α-value based on the characteristics of each data set. In order to find the best α-value, an initial α-value, the maximum α-value, and an increment parameter are pre-defined. We start with the initial α-value for model evaluation based on the accuracy (i.e., F1 measure) of the generated CARs. For each iteration, we increase the α-value using the increment parameter until reaching the maximum α-value.

Finally, we choose the generated CARs with the α-value which gives the best accuracy.

The last module is classification that is used to classify a testing data instance. There are many constraints to classify a testing data instance based on the generated CARs, such as the first matched rule and the majority class of the k-top matched rules. In this paper, the classification module is based on the first matched rule constraint since it is simple and fast. Moreover, it is suitable for comparing the accuracy between CARs ranked by the tradition confidence measure and our conflict-based confidence measure.

4. Our Proposed Conflict-Based Confidence Measure

The proposed *conflict-based confidence measure (cfb-conf)* is based on a novel way of defining the conflict between a rule and a training data instance. It quantifies the amount of conflict that a rule possesses with respect to all instances belonging to different classes in the data set.

4.1. A conflict measure for a rule

A conflict measure between a given rule and a training data instance is defined under the following three assumptions.

Assumption 1: The conflict between a given rule and the instances belonging to different classes should be high. In the contrast, there is no conflict between the given rule and any instance belonging to the same class.

Assumption 2: If the classes of a given rule and an instance are different, and all the feature-value pairs in the antecedent of the rule are identical to those features in that instance, the conflict is the highest.

Assumption 3: If the classes of a given rule and the instance are different, and all the feature-value pairs in the antecedent of the rule are different from those features in that instance, there is no conflict (the lowest).

Table 1 shows an example of three data instances in class $C^{(2)}$, showing only the values of those features (i.e., f_1, f_3, f_4, f_6) appeared in the antecedent of rule (i.e., {(f_1=Y), (f_3=Y), (f_4=N), (f_6=N)}→ $C^{(1)}$). Here, $r_1^{(1)}$ denotes the first rule of class $C^{(1)}$ and (f_4=N) means that the value of f_4 in the rule is *N*. The conflict between $r_1^{(1)}$ and all data instances in $C^{(2)}$ should be high. Data instance $T_1^{(2)}$ gives the highest conflict to $r_1^{(1)}$ because it misclassifies this data instance. While there is no conflict to data instance $T_3^{(2)}$, $r_1^{(1)}$ can classify this data instance correctly. Though data instance $T_2^{(2)}$ will not be misclassified by $r_1^{(1)}$, this data instance is considered to be close to the rule because the values of features f_1, f_4, and f_6 are the same.

From the aforementioned assumptions, we propose a function to calculate the conflict between the given rule and the training data instance with the following characteristics:

(a) The conflict should be a decreasing function of the distance between the rule and the training data instance in an unmatched class.

(b) The conflict should be an increasing function of the amount of conflict between the class labels.

Table 1. The example data instances in class $C^{(2)}$

	f_1	f_3	f_4	f_6	Class
$T_1^{(2)}$	Y	Y	N	N	$C^{(2)}$
$T_2^{(2)}$	Y	N	N	N	$C^{(2)}$
$T_3^{(2)}$	N	N	Y	Y	$C^{(2)}$

We apply the Manhattan distance as shown in Equation 3 to analyze the distance between the j^{th} feature of the given rule $r_l^{(k)}$ of class k and the training data instance $T_i^{(k')}$ in class k' ($k' \neq k$). Let $F_l^{(k)}$ and F_i be the set of features appeared in the antecedent of rule $r_l^{(k)}$ and the corresponding features in training data instance $T_i^{(k')}$, respectively. We can calculate the normalized distance $Dis(F_l^{(k)}, F_i)$ by dividing the total number of features in the antecedent of the rule (as shown in Equation 4).

$$d_j = \begin{cases} 0, & \text{whenever the feature value are identical} \\ 1, & \text{otherwise} \end{cases}; \quad (3)$$

$$Dis(F_l^{(k)}, F_i) = \left(\sum_{j=1}^{N_{\bar{l}}} d_j \right) \div N_{\bar{l}}. \quad (4)$$

The proposed conflict measure between the given rule $r_l^{(k)}$ and the training data instance $T_i^{(k')}$ is calculated by referring to Characteristics (a) and (b) and their distance is calculated using Equation 5. Since we consider only the conflict of the rule to all data instances of the unmatched classes, the conflict is 0 if their classes are identical.

$$conflict_{(l,i)} = \begin{cases} \left(1 - \alpha\left(Dis(F_l^{(k)}, F_i)\right)\right)^2, & \text{if their classes are different} \\ 0, & \text{if((their classes are identical) or } (Dis(F_l^{(k)}, F_i) < \frac{1}{\alpha})) \end{cases}. \quad (5)$$

Moreover, we only consider the conflict of the rule and *close data instances* of unmatched classes, due to alleviating the effect of noisy data in the training data. We define the α-value as *a closeness threshold* to prune all *non-close data instances*. The constraint is that the conflict of the data instance whose distance to the given rule is larger than $\frac{1}{\alpha}$ is not considered. Thus, the conflict variation is primarily dependent on the α-value. The larger α-value is assigned, the more number of instances are pruned. From the example in Table 1, when the α-value is 3, the conflict between rule $r_1^{(1)}$ and data instance $T_1^{(2)}$ is 1 since its distance to the rule is 0. For data instance $T_2^{(2)}$, its distance to the rule is 0.25 and its conflict to $r_1^{(1)}$ is 0.0625. The conflict to data instance $T_3^{(2)}$ is 0 because its distance to $r_1^{(1)}$ is 1, which is greater than 0.33 (1/3).

Since the most suitable α-value varies from the characteristics of a data set, the model evaluation module in our proposed framework in Section 3 aims to find this value adaptively. There are three pre-defined values in this module which are the initial and maximum α-values, and the increment parameter. For the maximum α-value, we always set it to the number of features in the data set. The reason is best demonstrated by the following example. If the number of features is 4 and the assigned maximum α-value is 4, we will consider only the data instances whose distance to the rule is less than 0.25 (1/4). In the case of the data instance whose values of three features are matched and the value of one feature is not matched to the rule, we won't calculate the conflict of this data instance since their distance is 0.25. It means that we consider only the exact matched data instance whose distance is 0. In our experiments, we set the maximum α-value to 10 since there are 10 features in the experimental data set. For the other pre-defined values, we set the initial α-value to 1 and the increment parameter is also set to 1.

4.2. The proposed conflict-based confidence measure

Now we consider a conflict-based confidence measure of a rule defined in a way that it would take the close data instances into account. If a rule is not good, there will be a significant number of "*close data instances which belong to different classes"* and there will probably be a larger number of conflicting data instances. Consequently, the assigned confidence will be reduced significantly. Based on this fact, we propose a function that calculates a conflict-based confidence of each rule $r_l^{(k)}$ as shown in Equation 6, where N_{TC} is the total number of classes. We average the conflict measures for each data instance in the unmatched classes by dividing the number of all data instances in the different classes. Then, the average conflict measure is used as a penalty score minus from the maximum confidence value which is 1. From the example in Table 1, if there are only three data instances of the unmatched classes in the database, the conflict-based confidence of the rule $r_1^{(1)}$ is $(1-\frac{1+0+0.0625}{3})$ = 0.6458.

$$Confidence_l^{(k)} = 1 - \frac{\sum_{\left\{\forall T_i^{(k')}:C_i \neq C_k\right\}} Conflict_{(l,i)}}{\left|\left\{\forall T_i^{(k')} : C_i \neq C_k\right\}\right|}, \forall k' = \overline{1, N_{TC}}. \quad (6)$$

5. Experimental Results

In this section, we aim to evaluate the proposed conflict-based confidence measure with respect to the classification accuracy. Three performance evaluation metrics used are precision, recall, and F1 values as shown in Equations 7, 8, and 9. Let *TP, FP,* and *FN* be the numbers of true positive, false positive, and false negative, respectively. The F1 measure is considered as a more suitable performance metrics than precision and recall values individually, since it is the harmonic mean of precision and recall values.

$$\text{Precision} = \frac{TP}{TP + FP}; \quad (7)$$

$$\text{Recall} = \frac{TP}{TP + FN}; \quad (8)$$

$$\text{F1} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}. \quad (9)$$

Experiments were run on a 3.00 GHz Pentium 4 CPU with 1GB of RAM running under Windows Command Processor. The data set used is the *bank.arff* data set available at [7]. This data set has only nominal data and belongs to two classes (YES and NO). Moreover, we used this data set to generate balanced and imbalanced data sets as shown in Table 2. We used the whole *bank.arff* as a balanced data set and randomly removed some of data instances belonging to Class *"Yes"* to generate an imbalanced data set. For each data set, we also randomly separated data into two sets used for training and testing.

Table 2. The details of experimental data sets

Data Set	Class	Total	Training	Testing
Balanced	YES	274	183	91
	NO	326	217	109
	TOTAL	600	400	200
Imbalanced	YES	40	27	13
	NO	326	217	109
	TOTAL	366	244	122

5.1. The effect of various α-values

Since the conflict variation is primarily dependent on the α-value, this experiment aims to decide the most suitable α-value for each data set adaptively based on the characteristics of the data set. In our experiments, for each data set, ten α-values ranging from 1 to 10 with an increment by 1 are used. Figure 2 shows that the average F1 values on various α-values of the first-fold experiments on both balanced and imbalanced training data sets. Both graphs increase to the highest peak, decrease, and then converge to steady classification accuracy. As can be seen from this figure, we can conclude that the difference in accuracy can be large with respect to the different α-values. In Figure 2, the best α-values of the first-fold classification models on the balanced and imbalanced training data sets are 4 and 3, respectively.

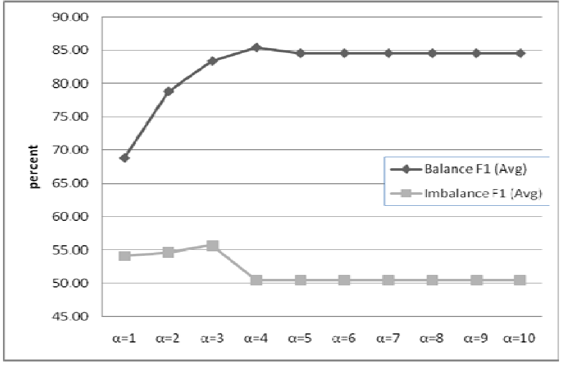

Figure 2. Average F1 values on various α-values of the first-fold experiments on both balanced and imbalanced training data sets

5.2. Performance comparison on a balanced testing data set

In this experiment, we intend to compare the classification accuracy on precision, recall, and F1 among the CARs ranked by the traditional confidence measure *(conf)* and the proposed conflict-based confidence measure based on the interleaving ordering strategy *(interleaving cfb-conf)*. To make the experimental results more convincing, 10-fold cross validation is used. For each run, referring to the previous experiment, we have to find the most suitable α-value for the generated CARs in the model evaluation module.

Table 3 shows the 10-fold cross validation comparison performance between the aforementioned methods on a balanced data set for Classes *"Yes"* and *"No"*, and their average values as well as the standard deviation values (in parentheses). As expected, the performance of the interleaving conflict-based associative classification framework can achieve better results of all performance evaluation measures on all classes with lower standard deviation values, comparing to the performance of the traditional confidence measure. Figure 3 demonstrates the bar chart that illustrates the average performance evaluation. According to this figure, we can conclude that the overall performance of CARs ranked by our proposed conflict-based confidence measure is better than that of the rules ranked by the traditional confidence measure.

5.3. Performance comparison on an imbalanced testing data set

We also conducted an experiment on an imbalanced training data set. Table 4 shows the comparison performance between the aforementioned methods on an imbalanced testing data set for Classes *"Yes"* and *"No"*, and their average values as well as the standard deviation values (in parentheses). Figure 4 shows the overall performance comparison between the aforementioned two rule ranking strategies. As can be seen from this figure, the performance of the precision, recall, and F1 values of the CARs ranked by the traditional confidence measure are very bad, since all rules totally misclassify the instances of Class *"Yes"* (the non-majority class). Since it cannot classify the instances of Class *"Yes"* for all 10 runs, the evaluation measures (Precision, Recall and F1) for each run are equal. Thus, the standard deviations (SD) of these measures are 0. In contrast, the experimental results of those ranked by our proposed interleaving conflict-based confidence measure show better performance.

Table 3. Performance comparison of 10-fold cross validation between conf & interleaving cfb-conf methods on a balanced testing data set

Class	Measures / Eval.	Conf (±SD)	Interleaving Cfb-Conf (±SD)
Yes	Precision(YES)	83.28 (±3.96)	82.24 (±2.86)
	Recall (YES)	75.71 (±4.29)	82.64 (±3.14)
	F1(YES)	79.18 (±2.21)	82.36 (±1.58)
No	Precision (NO)	81.21 (±2.44)	85.50 (±1.97)
	Recall (NO)	87.06 (±4.16)	84.95 (±3.33)
	F1(NO)	83.96 (±1.92)	85.17 (±1.57)
AVG	**Precision (Avg)**	**82.25** (±2.09)	**83.87** (±1.47)
	Recall (Avg)	**81.39** (±1.92)	**83.30** (±1.46)
	F1 (Avg)	**81.81** (±1.97)	**83.83** (±1.46)

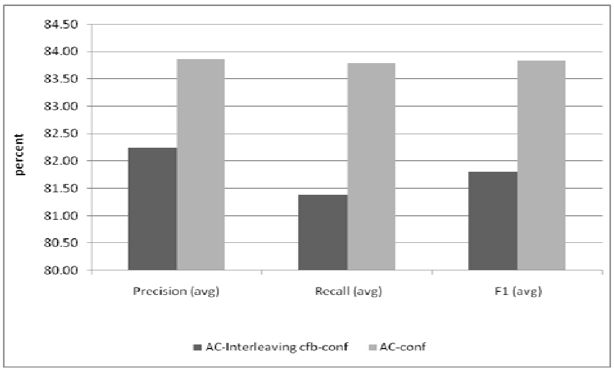

Figure 3. Average precision, recall, and F1 values of 10-fold cross validation between conf & interleaving cfb-conf on a balanced testing data set

6. Conclusion

In this paper, we propose an interleaving conflict-based associative classification (AC) framework. However, the traditional confidence measure which is used by most of the AC algorithms in the ranking process has a low discrimination power. To address this issue, a new confidence measure called *"conflict-based confidence measure"* is proposed which applies a distance function to find a conflict between a rule and all training data instances belonging to different classes in the training data set. Moreover, our proposed framework incorporates *an interleaving ordering strategy* in ranking the rules. The experimental results show that the CARs ranked by our proposed conflict-based confidence measure achieve a better performance than those ranked by the traditional confidence measure in both balanced and imbalanced data sets.

Table 4. Performance comparison of 10-fold cross validation between conf & interleaving cfb-conf methods on an imbalanced testing data set

Class	Measures Eval.	Conf (±SD)	Interleaving Cfb-Conf (±SD)
Yes	Precision(YES)	0.00 (±0.00)	11.09(±0.47)
	Recall (YES)	0.00 (±0.00)	89.23 (±10.38)
	F1(YES)	0.00 (±0.00)	19.70 (±0.94)
No	Precision (NO)	89.34 (±0.00)	93.81 (±5.54)
	Recall (NO)	100.00 (±0.00)	14.77 (±7.80)
	F1(NO)	94.37 (±0.00)	24.70 (±10.90)
AVG	**Precision (Avg)**	**44.67** (±0.00)	**52.45** (±2.98)
	Recall (Avg)	**50.00** (±0.00)	**52.00** (±2.25)
	F1 (Avg)	**47.19** (±0.00)	**52.22** (±2.57)

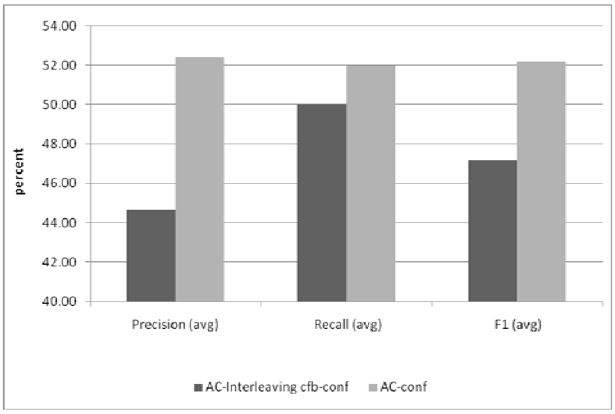

Figure 4. Average precision, recall, and F1 values of 10-fold cross validation between conf & interleaving cfb-conf on an imbalanced testing data set

7. References

[1] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," *Proceedings of the 20th International Conference on Very Large Databases*, 1994, pp. 487-499.

[2] R. Agrawal, T. Imielinski., and A. Swami, "Mining association rules between sets of items in large databases," *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, Washington, DC, 1993, pp. 207-216.

[3] G. Dong, X. Zhang, L. Wong, and J. Li, "CAEP: Classification by Aggregating Emerging Patterns," *Proc. Second Int'l Conf. Discovery Science (DS'99)*, Dec. 1999, pp. 30-42.

[4] C. Haruechaiyasak, M.-L. Shyu, and S.-C. Chen, "Identifying Topics for Web Documents through Fuzzy Association Learning," *International Journal of Computational Intelligence and Applications (IJCIA)*, Special Issue on Internet-Based Intelligent Systems, , September 2002, vol. 2, no. 3, pp. 277-285.

[5] W. Li, J. Han, and J. Pei, "CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules," *Proc IEEE Int'l Conf. Data Mining (ICDM'01)*, Washington, DC, Nov. 2001, pp. 369-376.

[6] B. Liu, W. Hsu, and Y. Ma, "Integrating Classification and Association Rule Mining," *Proc. Fourth Int'l Conf. Knowledge Discovery and Data Mining (KDD'98)*, New York, Aug. 1998, pp. 80-86.

[7] B. Mobasher, "http://maya.cs.depaul.edu/~Classes/Ect584/Weka/classify.html," School of CTI, DePaul University, 2005.

[8] M.-L. Shyu, S.-C. Chen, and S.H. Rubin, "Stochastic Clustering for Organizing Distributed Information Source," *IEEE Transactions on Systems, Man and Cybernetics*, Part B, vol. 34, no. 5, October 2004, pp. 2035-2047.

[9] S.P. Subasingha, J. Zhang, K. Premaratne, M.-L. Shyu, M. Kubat, and K.K.R.G.K. Hewawasam, "*Using Association Rules for Classification from Databases Having Class Label Ambiguities: A Belief Theoretic Method,"* Edited by T.Y. Lin, Y. Xie, A. Wasilewska, and C.-J. Liau, Data Mining: Foundations and Practice, pp. 523-546, Springer-Verlag, 2008, ISBN 978-3-540-78487-6.

[10] P.-N. Tan, M. Steinbach, and V. Kumar, "Introduction to Data Mining," *Addison Wesley; US Ed Edition*, May 2005, Chapter6, Section 6.7 pp. 370-786.

[11] F. Thabtah, "Challenges and Interesting Research Directions in Associative Classification," *Sixth IEEE International Conference on Data Mining-Workshops (ICDMW'06)*, Dec 2006, pp. 785-792.

[12] K. Wang, S. Zhou, and Y. He, "Growing Decision Trees on Support-Less Association Rules," *Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD'00),* Boston, Massachusetts, Aug. 2000, pp.265-269.