An Overview of Associative Classifiers

Yanmin Sun, Andrew K. C. Wong, Fellow, IEEE Yang Wang, Member, IEEE

Abstract

Associative classification is a new classification approach integrating association mining and classification. It becomes a significant tool for knowledge discovery and data mining. This paper presents a critical overview of certain aspects on modelling associative classifiers including event association mining, computational complexity in mining associations, methods of constructing associative classifiers based on the association patterns discovered, some similarities and differences among the reported associative classification systems, and experimental comparisons of these associative classifiers with the conventional classification system such as C4.5.

Index Terms-Data mining, classification, classifier design and evaluation, pattern discovery

1 Introduction

A new classification approach known as *associative classification* integrates association mining and classification into a single system [4, 8, 9, 10, 13, 15, 18]. Association mining, or pattern discovery, aims to discover descriptive knowledge from database, while classification focuses on building a classification model for categorizing new data. By and large, both association pattern discovery and classification rule mining are essential to practical data mining applications. Considerable efforts have been made to integrate these two techniques into one system. A typical associative classification system is constructed in two stages: 1) discovering all the event associations (in which the frequency of occurrences is significant according to some tests); 2) generating classification rules from the association patterns to build a classifier. In the first stage, the learning target is to discover the association patterns inherent in a database (also referred to as knowledge discovery). In the second stage, the task is to select a small set of relevant association patterns discovered to construct a classifier given the predicting attribute.

Experiments reported in [4, 8, 9, 10, 15, 18] showed that associative classification systems achieve competitive classification results with traditional classification approaches such as C4.5. The reasons are that the associative classifier is composed of high quality rules, which are generated from highly confident event associations that reflect the close dependencies among events; and, when classifying a new object, only significant rules are used. Meanwhile, classification rules induced from significant event associations are more easily understood by humans. Another advantage of this approach is its greater flexibility in handling unstructured data.

This study presents an overview of some of the recent works on constructing associative classifiers, focusing especially on the algorithms for association mining and associative classification system modelling. The paper is organized as follows. Following the introduction (Section 1), Section 2 is the study on association mining algorithms together with the computational complexity in mining associations. Section 3 is the study on modelling associative classification systems and the similarities and differences among three representative association classifiers reported. Section 4 is an experimental comparison of these three associative classifiers and their comparisons with the conventional classification system such as C4.5. Section 5 finally draws the conclusion and highlights the findings of this study.

2 Association Mining

Association mining was first proposed to analyze basket data. The basket problem assumes that in a grocery shop there are a large number of *items*, such as bread, milk, butter, beer, diapers and so on. Marketers would like to know what items people often buy together. For example, it may be found through analyzing transaction data that customers usually buy milk, butter and bread together. Marketers can then use this information to place these items in proper locations and adjust their selling strategies. In the like manner, the same sort of question is encountered in recommender systems, diagnosis decision support, intrusion detection, etc. The challenge is how to discover those events that are frequently associated together from large databases especially when no domain knowledge is available. Ever since its introduction, association mining has become an important technique for knowledge discovery from databases (KDD). Apriori algorithm [1] is originally reported for the analysis of transactional data. This algorithm regards an *itemset*, e.g. {milk, butter, bread}, as a *frequent itemset* if its frequency, which indicates how often the component items occur together, is greater than a pre-defined threshold. Each frequent itemset denotes one association pattern in a transactional data set. Another well developed method, which we refer to as *high-order pattern discovery* [16], detects association patterns using residual analysis which provides a rigorous statistical base to justify the significance of the discovered patterns. In their presentation, more general and formal terminologies are used: an item is defined as a *primary event*, an itemset as a *compound event* and a frequent itemset as an *event association* or *association pattern*. To present these two association mining algorithms consistently, it is better to use the same terminologies.

2.1 Data Representative

Consider a data set D containing M Samples. Every sample $\underline{x}$ is described in terms of N attributes, each of which can assume values in a corresponding discrete finite alphabet. Let $X = X_1, X_2, \cdots, X_N$ represent this attribute set. Each attribute, $X_i, 1 \leq i \leq N$, can be seen as a random variable taking on values from its alphabet $\alpha_i = \alpha_i^1, \cdots, \alpha_i^{m_i}$, where m_i is the cardinality of the alphabet of the i^{th} attribute. An additional attribute Y is considered as the target class with a set of k memberships $\{y_1, y_2, \cdots, y_k\}$ denoting the set of k class members. Then a sample can be represented as $\{\underline{x}, y\}$, where, $\underline{x} = \{x_1, \cdots, x_N\}$ is a realization of X, x_i can assume any value in α_i and y can assume any value in Y.

2.2 Terminology, Notations and Definitions

Definition 2.1 A <u>primary event</u> *of a random variable* $X_i (1 \leq i \leq N)$*, is a realization of* X_i *which takes on a value from* α_i*.*

We denote the $p^{th}(1 \leq p \leq m_i)$ primary event of X_i as

$$[X_i = \alpha_i^p]$$

or simply x_{ip}. We use x_i denoting a realization of X_i.

Let s be a subset of integers $\{1, \cdots, N\}$ containing k elements ($k \leq N$), and X^s be a subset of X such that

$$X^s = \{X_i | i \in s\}$$

Then x_p^s denotes the pth realization of X^s. We use x^s denoting a realization of X^s.

Definition 2.2 A compound event *associated with the variable set* $X^s = \{X_i | i \in s\}$ *is a set of primary event instantiated by a realization* x^s*. The* order *of the compound event is* $|s|$.

Definition 2.3 *Let T be a statistical test. If a compound event x^s passes the test T, we say that the primary events of x^s compose* an event association *or x^s is* an association pattern *of order $|s|$.*

2.3 Association Mining Algorithms

2.3.1 Apriori Algorithm

Let us denote the frequency of observed occurrences of a compound event x^s as o_{x^s}. The *support* of the compound event x^s is defined as its probability in the data set. That is

$$support(x^s) = \frac{o_{x^s}}{M} \tag{2.1}$$

where, M is the data size. A compound event x^s can be considered as an association pattern only if its support is greater than a pre-defined *minimum support*.

A general rule has the form of $A \Rightarrow B$ denoting that the observation of A infers that B is probably true. An association rule denotes the causal relationship between two compound events. Let l and k be two subsets of integers $\{1, \cdots, N\}$, where $l \cap k = \phi$. Then X^l and X^k are two subsets of X

$$X^l = \{X_i | i \in l\}$$

and

$$X^k = \{X_j | j \in k\}$$

such that

$$X^l \bigcap X^k = \phi$$

Let x^l denote a realization of X^l and x^k a realization of X^k. To test if $x^l \Rightarrow x^k$ is an association rule, the measure *confidence* is defined as

$$confidence(x^l \Rightarrow x^k) = \frac{support(x^k, x^l)}{support(x^l)} \tag{2.2}$$

If $confidence(x^l \Rightarrow x^k)$ is greater than a pre-defined *minimum confidence*, $x^l \Rightarrow x^k$ can be considered as an association rule.

To detect all association patterns, it makes multiple passes over the database. In the first pass, the algorithm simply counts primary event occurrences to determine the 1-order compound event. In each subsequent pass, say pass k, the algorithm starts with the (k-1)-order event associations found in the (k-1)th pass and generates new possible compound events. Next, the database is scanned and the supports of candidates are counted to determine which of the candidates are association patterns. See [1] for more details.

2.3.2 High-Order Pattern Discovery Using Residual Analysis

This method tests the statistical significance of the frequency of occurrence of a pattern candidate against that of its *expected occurrences* [16]. The frequency of expected occurrences of a compound event x^s is its expected total number under the assumption that the variables in X^s are mutually independent. The frequency of expected occurrences of x^s is denoted as e_{x^s}, such that

$$e_{x^s} = M \cdot \prod_{i \in s, x_i \in x^s} P(x_i) \tag{2.3}$$

where $P(x_i)$ is estimated by the ratio of observed frequency of x_i to the sample size M.

To test whether or not x^s is a significant association pattern, *standardized residual* defined in[5] is used to scale this deviation between o_{x^s} and e_{x^s}:

$$z_{x^s} = \frac{o_{x^s} - e_{x^s}}{\sqrt{e_{x^s}}} \tag{2.4}$$

Standardized residual z_{x^s} is the square root of χ^2. It has an asymptotic normal distribution with a mean of approximately zero and a variance of approximately one. Hence, if the absolute value of z_{x^s} exceeds 1.96, then, by a conventional criteria, x^s is considered as a significant association pattern with a confidence level of 95%. Standardized residual is considered to be of normal distribution only when the asymptotic variance of z_{x^s} is close to 1, otherwise, it has to be adjusted by its variance for a more precise analysis. The *adjusted residual* is expressed as:

$$d_{x^s} = \frac{z_{x^s}}{\sqrt{v_{x^s}}} \tag{2.5}$$

where v_{x^s} is the maximum likelihood estimate of the variance of z_{x^s}. More details can be found in [16].

2.4 Computational Complexity

Association mining is time consuming when data arrays contain a large number of rows and/or columns. Many studies [1, 6, 16] indicated its inherent nature of a combinatorially explosive number of event associaitons. Consider a data set with N *M-ary* attributes. The total number of combinations of kth order association patterns is given by

$$p_k = (M)^k \cdot \binom{N}{k}, \qquad 2 \leq k \leq N \tag{2.6}$$

where p_k denotes the total number of primary event combinations at order k. There are $\binom{N}{k}$ sets of variables of size k and M^k possible events for each variable set. Let the ratio of the number of pattern candidates of order k to the number of candidates of order $k-1$ be

$$\zeta_k = \frac{p_k}{p_{k-1}} = \frac{M \cdot (N - k + 1)}{k} \tag{2.7}$$

From the equation, we observe that as the order of the patterns increases from k=2 and upwards, especially when k is small compared to N, ζ_k increases rather fast. That is the size of the search space for kth order patterns increases fast with respect to that of the $(k-1)$th order patterns. For real world data, the pattern associations are sparsely scattered rather than uniformly distributed in the hypothesis space. If a compound event is not an association pattern of $(k-1)$-order, it cannot be expanded as a higher order event association. All k-order association pattern candidates are generated from $(k-1)$-order association patterns. Thus, the search complexity cannot be determined exactly since it is highly dependent upon the characteristics of the input data. Some research efforts are reported to deduce the computational complexity of association mining. Among them, the algorithm *FP-growth* [6] mines the association patterns without repeatedly scanning the database and check a large set of candidates by pattern matching when using the Apriori algorithm .

3 Associative Classifiers

3.1 Associative Classifiers Based on Apriori Algorithm

Apriori algorithm finds all association rules in the database that satisfy the pre-defined minimum support and minimum confidence constraints. For these association rules detected, there is not a fixed target at the right-hand-side. For classification purpose, rules for prediction should have an identical pre-determined target attribute. Works trying to induce classifiers from these discovered association rules are reported in[9, 10, 13, 18]. Classification rules are extracted from association rules by restricting the right-hand-side to the classification attribute. CBA [10] ranks these classification rules in sequence of their confidence, support and the order of generation. A minimum set of classification rules are then chosen according to the training error rate. In classifying an unknown case, the first rule that satisfies the case will be used. If there is no rule that applies, the default class (the majority class) will be taken. CMAR reported in [9] suggests a *weighted* χ^2 analysis to perform a classification based on multiple association rules. Given a new data object, CMAR collects the subset of rules matching the new object from the set of rules for classification. These rules may not be consistent with the class labels. CMAR first groups those according to class labels. Then, a *"combined effect"* is accounted for each group by adopting a *weighted*

χ^2 as the measure to determine the final class membership of the object. Generally speaking, with Apriori algorithm, the setting of minimum support and confidence is rather ad hoc. The user typically changes parameters and runs the mining algorithm many times in search of the "optimal" results. Such a process is very time consuming and little has been done to alleviate such problem[2]. Meanwhile, how to measure the rule qualities when a large number of rules are generated is another challenging issue.

3.2 Classification by Emerging Patterns

Emerging patterns (EPs) are defined as event associations whose supports change significantly from one dataset to another [3]. A data set is partitioned into several subgroups according to their class labels. The difference of the supports of an event association in one subgroup from those of the opposing group is measured. It is referred to as the *growth rate*. Those patterns whose growth rates satisfy a predefined threshold are detected. They are regarded as capturing the class discriminant information. Hence, such a pattern discovery process is directly classification-oriented. Both CAEP [4] and DeEPs [8] employ EPs as classification rules. CAEP first finds out all the EPs from the training data for each class. When classifying a new object by aggregating the differentiating power of a set of EPs that apply, a score is obtained for each class, and that with the highest score wins. Arguing that to discover all EPs from the training data is time consuming, DeEPs proposes a "lazy" learning approach. Whenever a new instance is being considered, DeEPs uses it as a filter to remove irrelevant training values in order to reduce the training space. Boundary EPs are detected for each class. To classify this instance, a collective score for each class is calculated by summarizing the frequencies of the selected EPs pertaining to each class. As an EP discovery process is instance-based, all the training data should be stored for re-learning during the entire classification process.

3.3 High-Order Pattern and Weight-of-Evidence Rule Based Classifier

High-order pattern and weight of evidence rule based classifier (HPWR) [14, 15] is a well developed classification system. As introduced in Section 2.3.2, the algorithm of high-order pattern discovery detects significant association patterns by using residual analysis in statistics. At the next stage, classification rules are generated using *weight of evidence* to quantify the evidence of significant association patterns in support of, or against a certain class membership[15].

Based on the mutual information in information theory, the difference in the gain of information when predicting attribute Y takes on the value y_i and when it takes on some other values, given $\underline{x}$, is a measure of evidence provided by $\underline{x}$ in favor of y_i being a plausible value of Y as opposed to Y taking other values. This difference, denoted by $W(Y = y_i/Y \neq y_i|\underline{x})$, is defined as the weight of evidence, which has the following forms:

$$
\begin{aligned}
& W(Y = y_i/Y \neq y_i|\underline{x}) && (3.1)\\
= \ & I(Y = y_i : \underline{x}) - I(Y \neq y_i : \underline{x}) && (3.2)\\
= \ & log\frac{P(Y = y_i|\underline{x})}{P(Y = y_i)} - \log\frac{P(Y \neq y_i|\underline{x})}{P(Y \neq y_i)} && (3.3)\\
= \ & \log\frac{P(\underline{x}|Y = y_i)}{P(\underline{x}|Y \neq y_i)} && (3.4)
\end{aligned}
$$

where $I(\cdot)$ is the mutual information. The weight of evidence is positive if $\underline{x}$ provides positive evidence supporting Y taking on y_i, otherwise, it is negative, or zero.

Suppose that there are n sub-compound events $\underline{x}_1, \cdots, \underline{x}_n$ are detected, where $(\underline{x}_k, Y = y_i)$ is a significant association pattern and $\cup_{k=1}^{n}\underline{x}_k = \underline{x}$; $\underline{x}_p \bigcap \underline{x}_q = \Phi$ when $p \neq q$, $1 \leq k, p, q \leq n$. According to [14], the weight of evidence $W(Y = y_i/Y \neq y_i|\underline{x})$ can be obtained from the sum of the weight of evidence for all association patterns in $\underline{x}$ supporting $Y = y_i$. That is

$$
\begin{aligned}
& W(Y = y_i/Y \neq y_i|\underline{x}) \\
= \ & \log\frac{P(\underline{x}_1|Y = y_i)}{P(\underline{x}_1|Y \neq y_i)} + ... + \log\frac{P(\underline{x}_n|Y = y_i)}{P(\underline{x}_n|Y \neq y_i)} && (3.5)
\end{aligned}
$$

$$= \sum_{k=1}^{n} W(Y = y_i / Y \neq y_i | \underline{x}_k) \quad (3.6)$$

Thus, the calculation of weight of evidence is to find a proper set of disjoint significant event associations from $\underline{x}$ and to sum individual weight of evidence provided by each of them. The task is to maximize the term in Equation (3.6). The most plausible value y_i of Y is the one that renders the highest weight.

3.4 Analysis

Among these three types of associative classification systems, certain similarities and differences exist.

1. Associative classifiers based on Apriori algorithm like CBA, and HPWR are typical associative classification systems. Traditionally, association and classification are two independently important tasks for practical applications. Association is mainly used in data mining for discovering descriptive knowledge from databases, while classification is addressed in the field of machine learning for exploring boundaries among classes. As association pattern mining and classification rule mining are both indispensable in a data mining system, there are needs to integrate both into an association classification system as reflected by many such research efforts. In general, the pattern discovery phase detects all event associations without necessarily relating the class it might be associated with. When the predicting attribute for classification is given at the second stage, a subset of association patterns or rules relevant to the predicting attribute is selected to construct a classifier. Theoretically, when a different predicting attribute is assigned, no re-learning is necessary for pattern discovery. Both CBA and HPWR share such a view. On the other side, the learning processes of CAEP and DeEPs are more like those of a traditional classifier. Their classification rules are generated from emerging patterns (EPs). The discovering of EPs is class-based (one class against other classes) assuming the predicting attribute is known at the pattern discovery phase. This learning process serves the purpose of classification instead of exploring descriptive knowledge across the entire database.

2. HPWR, CAEP and DeEPs use multiple rules to classify a new object: HPWR employs weight of evidence, which accounts for the strength of the association between the class membership and all the admissible statistically significant conditions. The total weights of evidence provided by several applicable patterns are "addable" if they are conditionally independent [17]; CAEP obtains a score for each class by aggregating the differentiating power of EPs which apply to the test object; and DeEPs determines collective scores for all classes by compactly summarizing the frequencies of the discovered boundary EPs. Thus, all relevant EPs of a class contribute to the final decision. On the other side, CBA uses only one rule for prediction. One problem with this is that it cannot handle partial information from the test object [15]. For example, if an unknown instance to be classified is $O = [A, B, C]$ and according to *rule1:* $A \Rightarrow$ *class1*, O belongs to *class1*, but according to *rule2*: $B \Rightarrow$ *class2* and *rule3*: $C \Rightarrow$ *class2*, O belongs to *class2*. Then CBA will classify O as *class1* since *rule1* precedes *rule2* and *rule3* even though the combination of *rule2* and *rule3* might be more determining.

3. HPWR discovers all the association patterns using *residual analysis*. The statistical significance of an association pattern is guaranteed, which eliminates the need of using unstandardized (widely varying) and arbitrary thresholds. Meanwhile, the residual is easily interpreted in terms of the degree of satisfaction in the discovery when compared with others; CBA employs Aproiri algorithm to detect association rules by testing their supports and confidences; EPs used by CAEP and DeEPs are discovered by calculating their supports in one class against the others and obtaining a growth rate reflecting the significance of the support changes. One learning issue with the latter two pattern discovery methods is the setting of the threshold: in Apriori algorithm, which is the minimum support and minimum confidence; and in EP mining, which is the growth rate (set as infinite in DeEPs).

4 Experimental Comparisons

In this section, we conduct experiments to evaluate the performances of the associative classification systems, as well as to compare them with the conventional classification system

C4.5. In [12], performances of C4.5 and boosting to C4.5 are examined and compared. 27 datasets taken from UCI Machine Learning Repository [11] with considerable diversity in sample size, number of classes, and number of attributes are used for the evaluation. In [8], 40 datasets are used to test both DeEPs and CBA. Among them 25 datasets are tested in [12] except "Audiology" and "Phoneme". We will conduct experiments on HPWR using these 25 datasets. The descriptions of these datadets are summarized in Table 1.

Table 1: Description of Datasets

	Data set	# Attributes	#Class	# Instances
1	Anneal	38	6	798
2	Auto	25	7	205
3	Breast-w	9	2	699
4	Chess	36	2	3196
5	Cleve	13	2	303
6	Crx	15	2	690
7	Diabetes	8	2	768
8	German	20	2	1000
9	Glass	9	7	214
10	Hepatitis	19	2	155
11	Horse	22	2	368
12	Hungarian	13	2	294
13	Hypo	25	2	3163
14	Iris	4	3	150
15	Labor	16	2	57
16	Letter	16	26	20000
17	Lymp	18	4	148
18	Segment	19	7	2310
19	Sick	25	2	3163
20	Sonar	60	2	208
21	Soybean	35	19	683
22	Splice	60	3	3190
23	Vehicle	18	4	846
24	vote	16	2	435
25	Waveform	21	3	5000

In experiments reported in [12], continuous data in each data set is pre-discretized using the commonly used discretization utility in MLC++ [7] with the default setting. The classification performances are evaluated by their classification accuracies which are based on the percentage of correct predictions on the test sets. 10-fold cross validation is used, in which a data set is divided into 10 subsets; each subset is in turn used as testing data while the remaining data is used as the training data set; then the average accuracy across all 10 trials is reported. In our experiments, in order to compare the performances fairly, we adopt the same routines as described above. Table 2 shows the results.

Table 2: Comparison of Associative Classifiers with C4.5

DataSet	C4.5	DeEP	CBA	HPWR
Anneal	92.33	94.41	98.10	96.90
Auto	82.34	67.65	79.00	80.74
Breast-w	94.72	96.42	95.28	96.82
Chess	91.45	97.81	98.12	94.83
Cleve	77.06	81.17	77.24	93.75
Crx	85.30	84.18	85.90	83.89
Diabetes	74.61	76.82	72.90	73.30
German	71.56	74.40	73.20	76.30
Glass	67.52	58.49	72.60	69.76
Hepatitis	79.61	81.18	80.20	89.07
Horse	85.08	84.21	82.10	83.22
Hungarian	78.47	81.11	81.87	84.73
Hypo	99.52	97.19	98.40	97.58
Iris	95.20	96.00	92.90	96.91
Labor	80.88	87.67	83.00	94.97
Letter	88.01	93.60	51.76	72.30
Lymp	78.31	75.42	77.33	85.24
Segment	96.79	94.98	93.51	86.66
Sick	98.66	94.03	97.30	95.45
Sonar	74.38	84.16	78.30	83.21
Soybean	92.42	90.00	92.23	87.34
Splice	94.09	69.71	70.03	92.73
Vehicle	72.91	70.95	68.78	64.87
Vote	94.94	95.17	93.54	91.91
Waveform	72.67	84.36	75.34	77.83
Average	84.75	84.43	82.76	86.01

When each individual associative classifier is compared with C4.5, over these 25 datasets, DeEPs performs better on 12 datasets, similarly (difference around or below 1%) on 4 datasets, worse on 8 datasets and the average accuracies between them are similar; CBA performs better on 8 datasets, similarly on 7 datasets, worse on 10 datasets, and the average accuracy is lower than C4.5 by 1.99%; HPWR performs better on 13 datasets, similarly on 1 dataset, worse on 11 datasets, and the average accuracy is slightly higher by 1.26%. These comparison results manifest that associative classifiers are competitive with the conventional classifiers such as C4.5. When these three associative classifiers are compared, DeEPs performs the best on 8 datasets; CBA on 7 datasets and HPWR on the remaining 10 datasets; the average accuracy of HPWR is slightly better.

5 Conclusion

In this overview, we show that the new classification approach, associative classification, is developed by integrating two important techniques of data mining, association pattern mining and classification system modelling, into one system. Two well known association mining algorithms are Apriori algorithm and high-order pattern discovery. The main concern in detecting event associations is the computational complexity due to the inherent nature of a combinatorially explosive number of event associations especially when data arrays contain a large number of rows and/or columns. Ever since a large set of association patterns are detected, many research works have been done to apply these discoveries for classification tasks. In this overview, three types of associative classification systems are studied. Analysis on the similarities and differences among them are provided. Further experimental results indicate: 1) associative classifiers are competitive with the conventional classifiers such as C4.5; and 2) among these three associative classifiers compared, HPWR performs better than the other two.

References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proceedings of International Conference on Very Large Data Bases(VLDB'94)*, pages 487–499, 1999.

[2] G. Cong and B. Liu. Speed-up interative frequent intemset mining with constraint changes. In *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM-2002)*, pages 107–114, Maebashi, Japan, December 2002.

[3] G. Dong and J. Li. Efficient mining of emerging paterns: Discovering trends and differences. In S. Chaudhui and D. Madigan, editors, *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 43–52. ACM Press, San Diego, CA, 1999.

[4] G. Dong, X. Zhang, L. Wong, and J. Li. CAEP:classification by aggregating emerging patterns. In *Proceedings of The Second International Conference on Discovery Science (DS'99)*, pages 43–55, Japan, December 1999.

[5] S. J. Haberman. *The Analysis of Residuals in Cross-Classified Tables*, volume 29. Biometrics, 1973.

[6] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. pages 1–12, May 2000.

[7] R. Kohavi, D. Sommerfield, and J. Dougherty. *Data Mining Using MLC++: A machine learning library in C++. Tools with Artificial Intelligence.* IEEE CS Press, 1996.

[8] J. Li, G. Dong, K. Ramamohanarao, and L. Wong. DeEPs: a new instance-based lazy discovery and classification system. *Machine Learning*, 54(2):99–124, 2004.

[9] W. Li, J. Han, and J. Pei. CMAR: Accurate and efficient classification based on multiple class-association rules. In *Proceedings of The 2001 IEEE International Conference on Data Mining (ICDM'01)*, pages 369–376, San Jose, CA, November 2001.

[10] B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. In *Proceedings of the Fourth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 80–86, New York, NY, August 1998.

[11] P. M. Murph and D. W. Aha. *UCI Repository Of Machine Learning Databases*. Department Of Information and Computer Science, University Of California: Irvine, 1991.

[12] J. R. Quinlan. Bagging, Boosting, and C4.5. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence and the Eighth Innovative Applications of Artificial Intelligence Conference*, pages 715–730, Menlo, Park, August 1996.

[13] K. Wang, S. Zhou, and Y. He. Growing decision tree on support-less association rules. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining(KDD'00)*, pages 265–269, Boston, MA, August 2000.

[14] Y. Wang. *High-Order Pattern Discovery and Analysis of Discrete-Valued Data Sets*. PhD thesis, University of Waterloo, Waterloo, Ontario, Canada, 1997.

[15] Y. Wang and A. K. C. Wong. From association to classification: Inference using weight of evidence. *IEEE Trans. On Knowledge and Data Engineering*, 15(3):764–767, 2003.

[16] A. K. C. Wong and Y. Wang. High order pattern discovery from discrete-valued data. *IEEE Trans. On Knowledge and Data Engineering*, 9(6):877–893, 1997.

[17] A. K. C. Wong and Y. Wang. Pattern discovery: A data driven approach to decision support. *IEEE Transactions on Systems, Man, and Cybernetics*, 33(1):114–124, 2003.

[18] X. Yin and J. Han. CPAR: Classification based on predictive association rules. In *Proceedings 2003 SIAM International Conference on Data Mining(SDM'03)*, pages 331–335, San Francisco, CA, May 2003.