# Using Multi-Attribute Predicates for Mining Classification Rules

Ming-Syan Chen

Electrical Engineering Department
National Taiwan University
Taipei, Taiwan, ROC
email: mschen@cc.ee.ntu.edu.tw

## Abstract

*In order to improve the efficiency of deriving classification rules from a large training dataset, we develop in this paper a two-phase method for multi-attribute extraction. A feature that is useful in inferring the group identity of a data tuple is said to have a good inference power to that group identity. Given a large training set of data tuples, the first phase, referred to as feature extraction phase, is applied to a subset of the training database with the purpose of identifying useful features which have good inference powers to group identities. In the second phase, referred to as feature combination phase, these extracted features are evaluated together and multi-attribute predicates with strong inference powers are identified. A technique on using match index of attributes is devised to reduce the processing cost.*

## 1 Introduction

Various data mining capabilities have been explored in the literature. Mining association rules has attracted a significant amount of research attention [3, 9, 11, 15]. For example, given a database of sales transactions, it is desirable to discover all associations among items such that the presence of some items in a transaction will imply the presence of other items in the same transaction. Another type of data mining is on ordered data, such as stock market and point of sales data. Interesting aspects to explore from these ordered data include searching for similar sequences [1, 16], e.g., stocks with similar movement in stock prices, and sequential patterns [4], e.g., grocery items bought over a set of visits in sequence. Mining on Web path traversal patterns was studied in [6]. In addition, one important application of data mining is the ability to perform classification in a huge amount of data. This is referred to as mining classification rules. Explicitly, mining classification rules is an approach of trying to develop rules to group data tuples together based on certain common features. For an example of commercial applications, it is desirable for a car dealer to know what are the common features of its most customers so that its sales persons will know whom to approach, and its catalogs of new models can be mailed directly to those customers with identified features. The business opportunity can thus be maximized. It is noted that due to the increasing use of computing for various applications, the importance of mining classification rules is growing at a rapid pace. The fast growth in the amount of data in those applications has furthermore made the efficient mining for classification rules a very challenging issue.

Classification rule mining has been explored both in the AI domain [12, 14] and in the context of databases [2, 5, 7, 8]. In machine learning, a decision-tree classification method, developed by Quinlan [13, 14], is one of the most important results, and has been very influential to later studies. It is a supervised learning method that constructs decision trees from a set of examples. The quality of a tree depends on both the classification accuracy and the size of the tree. Other approaches on data classification include statistical approaches [12], rough sets approach [17], etc. In the context of databases, an interval classifier has been proposed in [2] to reduce the cost of decision tree generation. An attribute-oriented induction method has been developed for mining classification rules in relational databases [8]. The work in [10] explores rule extraction in a database based on neural networks.

It is noted that in mining classification rules for a given database, one would naturally like to have a training dataset large enough so as to have a sufficient confidence on the rules derived. However, with a large training set, the execution time required for rule derivation could be prohibitive, in particular, when forming multi-attribute predicates is needed. When a sophisticated predicate is constructed from a combination of features, the execution time required grows exponentially with the size of a training database, which is highly undesirable in many applications. Consequently, we present in this paper a two-phase method for multi-attribute extraction and improve the efficiency of deriving classification rules in a large training dataset. A feature that is useful in inferring the group identity of a data tuple is said to have a good *inference power* to that group identity. Given a large training set of data tuples, the first phase, referred to as *feature extraction phase*, is applied to a subset of the training database with the purpose of identifying useful features which have good inference powers to group identities. Note, however, that in some cases the group identity is not so dependent on the value of a single attribute. Rather, the group identity depends on the combined values of a set of attributes. This is particularly true in a database where attributes have strong dependencies among themselves. Combining several individual features is thus required for constructing multi-attribute predicates with better inference powers. In the second phase, referred to as *feature combination phase*, those features extracted from the first phase are evaluated together and multi-attribute predicates with strong inference powers are identified. A technique on using *match index* of attributes is devised to reduce the processing cost. In essence, a match index is a heuristic indication on the combined inference power of multiple attributes, and can be used to identify uninteresting combined attributes and remove them from later processing. Note that being performed only on a subset of the training set, the feature extraction phase can be executed efficiently. On the other hand, since the features extracted are used to the whole training set in the feature combination phase, the confidence of the final classification rules derived can hence be ensured.

This paper is organized as follows. A problem description is given in Section 2. The two-phase method for mining classification rules is described in Section 3. Section 4 contains the summary.

## 2 Problem Description

In general, the problem on mining classification rules can be stated as follows. We are given a large database W, in which each tuple consists of a set of $n$ attributes (features), $\{A_1, A_2, ..., A_n\}$. The terms “attribute” and “feature” are used interchangeably in this paper. For example, attributes could be age, salary range, gender, zip code, etc. Our purpose is to classify all data tuples in this database into different groups according to their attributes. In order to learn proper knowledge on such classification, we are given a small training database E, in which each tuple consists of the same attributes as tuples in W, and additionally has a known group identity associated with it. An example group identity is the type of car owned, say “plain”, “good”, or “luxury”. We want to (1) learn the relationship between “attributes” and group identity from the training database E, and then (2) apply the learned knowledge to classify data in the large database W into different groups. Note that once the relationship between attributes and group identity is learned in (1), the process in (2) can be performed in a straightforward manner. Hence, we shall focus our discussion on methods for (1) in this paper, i.e., to identify attributes from $\{A_1, A_2, ..., A_n\}$ that have strong inference to the group identity.

| Label | Gender | Age | Beverage | State | Group id |
|---|---|---|---|---|---|
| 1 | M | 3 | water | CA | I |
| 2 | F | 4 | juice | NY | I |
| 3 | M | 4 | water | TX | I |
| 4 | F | 4 | milk | TX | I |
| 5 | M | 5 | water | NY | I |
| 6 | M | 3 | juice | CA | I |
| 7 | M | 3 | water | CA | II |
| 8 | F | 5 | juice | TX | II |
| 9 | F | 5 | juice | NY | II |
| 10 | F | 6 | milk | TX | III |
| 11 | M | 4 | milk | NY | III |
| 12 | M | 5 | milk | CA | III |
| 13 | F | 4 | milk | TX | III |
| 14 | F | 6 | water | NY | III |
| 15 | F | 6 | water | CA | III |

Table 1. A sample profile for classifying 15 children.

Consider a sample profile for 15 children in Table 1 as an example. In Table 1, each tuple, corresponding to each child, contains attributes: his/her gender, age, beverage preferred and state lived, and additionally his/her group identity. (For ease of exposition, each tuple is given a label in it first column, which is, however, not part of the attributes.) We now would like to explore the relationship between the attributes (i.e., gender, age, beverage and state in this case) and the group identity. As stated before, an attribute that is useful in inferring the group identity of a data tuple is said to have a good inference power to that group identity. A *predicate* in this study means a resulting classification rule from step (1) mentioned above, and will be used in step (2) to classify data tuples in the database W. In our discussion, the quality of a predicate refers to the combined inference power of the attributes this predicate is composed of. The proposed method consists of two phases: *feature extraction phase* and *feature combination phase*. Given a large training set of data tuples, the first phase, feature extraction phase, is to learn useful features, which have good inference powers to group identities, from a subset D of the training database E.

As mentioned earlier, in some cases the group identity is not so dependent on the value of a single attribute, but instead, depends upon the combined values of a set of attributes. This is particularly true in the presence of those attributes that have strong inference among themselves. Consider the profile in Table 2 as an example. In Table 2, it is found that a male with low income and a female with high income usually drive cars, whereas a male with high income and a female with low income ride bikes. In this case, exploring the relationship between "vehicle" (corresponding to the group id in Table 1) and "either gender or income attribute" will lead to little results, since neither gender nor income has a good inference power to the vehicle. However, a combination of gender and income (e.g., a male and low income) indeed has a good inference power to the vehicle. It can be seen that the type of vehicle in each tuple can in fact be determined from the combined value of gender and income. In view of this, to exploit the benefit of multi-attribute predicates, we devise the second phase, feature combination phase, which evaluates individual features extracted in the first phase and produces multi-attribute predicates with strong inference powers.

The two-phase method for mining classification rules can be summarized as follows.

**Algorithm M**: Mining classification rules

**Feature extraction phase:** To learn useful features, which have good inference powers to group identities, from a subset of the training database.

**Feature combination phase:** To evaluate extracted features based on the entire training database and form multi-attribute predicates with good inference powers.

| Label | Gender | Income | Vehicle |
|---|---|---|---|
| 1 | male | low | car |
| 2 | male | low | car |
| 3 | female | high | car |
| 4 | female | high | car |
| 5 | male | high | bike |
| 6 | male | high | bike |
| 7 | female | low | bike |
| 8 | female | low | bike |

Table 2. A sample profile for preferred vehicles.

## 3 Mining Classification Rules

We describe in this section a two phase method for mining classification rules. The first phase, feature extraction phase, is presented in Section 3.1, and the second phase, feature combination phase, is presented in Section 3.2.

As illustrated in Figure 1, the feature extraction phase is applied to a subset of the training database. In this phase, attributes that have good inference powers to group identities are identified. The operations of this phase are explained below. First, tuples in the database D are divided into several groups according to their group id's. The inference power of each attribute is then investigated one by one. Suppose $A$ is an attribute and $\{a_1,\ a_2, ..., a_m\}$ are $m$ possible values of attribute $A$. Also, the domain of the group identity $g$ is represented by domain($g$)= $\{v_1,\ v_2, ..., v_{|domain(g)|}\}$. The *primary group* for a value $a_i$ of attribute $A$, denoted by $v^{a_i}$, is the group that has the most tuples with their attribute $A$= $a_i$. Explicitly, use $n_A(a_i, v_k)$ to denote the number of tuples which are in group $v_k$ and have a value of $a_i$ in their attribute $A$. Then, we have

$$n_A(a_i, v^{a_i}) = \max_{v_k \in domain(g)} \{n_A(a_i, v_k)\}\,. \qquad (1)$$

The primary group for each value of attribute $A$ can hence be obtained. For the example profile in Table 1, if $A$ is "gender," then domain($A$)= {Male, Female}, and $n_A$(Male,I)=4, $n_A$(Male,II)=1, and $n_A$(Male,III)=2. Group I is therefore the primary group for the value "Male" of the attribute "gender".

The *hit ratio* of attribute $A$, denoted by $h(A)$, is defined as the percentage of tuples which, according to their corresponding attribute values, fall into their primary groups. Let $N$ denote the total number of tuples. Then,

$$h(A) = \frac{\sum_{1 \leq i \leq m} n_A(a_i, v^{a_i})}{N}. \qquad (2)$$

It can be seen that the stronger the relationship between an attribute and the group identity, the larger the hit ratio of this attribute will be. A hit ratio of an attribute would become one if that attribute could uniquely determine the group identity. The hit ratio is a quantitative measurement for the inference power of an attribute. According to the primary groups of various values of an attribute, the hit ratio of that attribute can be determined. Note that in essence we want to investigate the inference power of some combined features. However, to reduce the processing cost, we would like to restrict our attention to those attributes whose individual hit ratios meet a predetermined threshold. Specifically, we include attribute $A$ into a set $S_A$ for future processing only if the hit ratio of $A$ is larger than or equal to a predetermined threshold. Note that this is a greedy approach and does not guarantee providing the optimal solutions. As a consequence, those features with poor inference powers will be removed from later processing, and the processing cost can thus be reduced. The *most distinguishing attribute* refers to the attribute with the largest hit ratio. Formally, the flow of the feature extraction phase is summarized as follows.

**Feature extraction phase:**

**Step 1:** Divide tuples in the database D into several groups according to their group id's.

**Step 2:** Let $A$ denote the next attribute to process.

**Step 3:** Determine the primary group for each value of attribute $A$.

**Step 4:** According to the primary groups of various values of attribute $A$, obtain the hit ratio of $A$.

**Step 5:** Include attribute $A$ into set $S_A$ for future processing if the hit ratio of $A$ is larger than or equal to a predetermined threshold.

| Gender | I | II | III | (max, group) |
|---|---|---|---|---|
| Male | 4 | 1 | 2 | (4, I) |
| Female | 2 | 2 | 4 | (4, III) |
| hit ratio: | | | | $\frac{8}{15}$ |

Table 3. Distribution when the profile is classified by gender.

**Step 6:** If there is any more attribute to process then go to Step 2.
Otherwise stop.

For illustrative purposes, consider the example profile in Table 1, which can be viewed as a subset of the training database to be used in the feature extraction phase. First, we classify this profile according to gender, and obtain the results in Table 3. As explained earlier, Group I is the primary group for the value "Male" of attribute "gender". Also, it can be obtained that Group III is the primary group for the value "Female" of attribute "gender". As a result, there are 8 tuples, out of 15 tuples, fall into their primary groups. The hit ratio of attribute gender is thus $\frac{8}{15}$.

Similarly, it can be verified that the hit ratios of age, beverage and state are, respectively, $\frac{10}{15}$, $\frac{9}{15}$ and $\frac{6}{15}$. Finally, having the largest hit ratio among the four attributes, age is the most distinguishing attribute in this example. Suppose the predetermined threshold for the inclusion into $S_A$ is $\frac{8}{15}$. Then, attributes gender, age and beverage are included into $S_A$ whereas attribute state is not. The attributes collected in $S_A$ will be evaluated together in the feature combination phase to form multi-attribute predicates. Moreover, we have the following lemma which specifies the lower bound of the hit ratio of an attribute.

**Lemma 1:** Let $A$ be an attribute and $g$ be a group identity. Then,

$$h(A) \geq \frac{1}{|\text{domain}(g)|},$$

which is a tight lower bound.

Note that the feature extraction phase explores the relationship between the group identity and individual attributes. However, as explained by using the profile in Table 2 earlier, in some cases the group identity is not so dependent on the value of a single attribute, but instead, depends upon the combined values of a set of attributes. This is the very reason that the feature combination phase is called for.

The feature combination phase is applied to the entire training database with the purpose of evaluating the inference power resulting from combining attributes. As stated before, we shall only examine those attributes in $S_A$ so as to decrease the processing cost. In addition, a technique on using match index of attributes is devised to further reduce the processing cost. Specifically, instead of evaluating every pair of attributes, we shall only deal with those attribute pairs whose *match indexes* meet another threshold, where a match index is a heuristic indication on how likely a pair of attributes will lead to a strong inference power as a whole. Suppose that $A$ and $B$ are two attributes to be used to form a 2-attribute predicate. Let domain($A$)= $\{a_1, a_2, ..., a_m\}$ and domain($B$)= $\{b_1, b_2, ..., b_p\}$. Recall that $n_A(a_i, v_k)$ is the number of tuples which are in group $v_k$ and have their attribute $A = a_i$. $n_B(b_j, v_k)$ is defined similarly. Then, the *match index* of two attributes $A$ and $B$, denoted by $M_I(A,\ B)$, is defined as:

$$M_I(A, B) = \sum_{1 \le i \le m} \sum_{1 \le j \le p} \max_{v_k \in domain(g)} \{\min(n_A(a_i, v_k), n_B(b_j, v_k))\}\ .$$

Essentially, $M_I(A,B)$ is a heuristic indication on the combined inference power of $A$ and $B$. It can be observed that $M_I(A,B)$ is in fact an upper bound for the number of tuples falling into the primary groups of various values of attribute pair ($A$,$B$). Explicitly, $\frac{M_I(A,B)}{N}$ is an upper bound for the hit ratio of attribute pair ($A$,$B$). The use of match index will significantly reduce the processing cost while causing little compromise on the quality of the resulting predicates. The overall flow of the feature combination phase can be summarized as follows.

**Feature combination phase:**

**Step 1:** Let attribute pair ($A$,$B$) be the next attribute pair from $S_A$ to process.

**Step 2:** If the match index of ($A$,$B$) does not reach the predetermined threshold, go to Step 6.

**Step 3:** Determine the primary group for each value of attribute pair ($A$,$B$).

**Step 4:** According to the primary groups of various values of attribute pair ($A$,$B$), obtain the hit ratio of pair ($A$,$B$).

**Step 5:** Include attribute pair ($A$,$B$) into a set $S_C$ for future processing if the hit ratio of ($A$,$B$) is larger than or equal to a predetermined threshold.

**Step 6:** If there is any more attribute pair to process then go to Step 1.
Otherwise stop.

## 4 Conclusion

We have developed in this paper a two-phase method for multi-attribute extraction and improved the efficiency of classification rule derivation in a large training dataset. Given a large training set of data tuples, the feature extraction phase was applied to a subset of the training database with the purpose of identifying useful features which have good inference powers to group identities. In the feature combination phase, these extracted features were evaluated systematically and multi-attribute predicates with strong inference powers were identified. A technique on using match index of attributes was utilized to reduce the processing cost.

## Acknowledgements

M.-S. Chen is in part supported by National Science Council, Project No. NSC 87-2213-E-002-009 and Project No. 87-2213-E-002-101, Taiwan, ROC.

## References

[1] R. Agrawal, C. Faloutsos, and A. Swami. Efficient Similarity Search in Sequence Databases. *Proceedings of the 4th Intl. conf. on Foundations of Data Organization and Algorithms*, October, 1993.

[2] R. Agrawal, S. Ghosh, T. Imielinski, B. Iyer, and A. Swami. An Interval Classifier for Database Mining Applications. *Proceedings of the 18th International Conference on Very Large Data Bases*, pages 560–573, August 1992.

[3] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules in Large Databases. *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 478–499, September 1994.

[4] R. Agrawal and R. Srikant. Mining Sequential Patterns. *Proceedings of the 11th International Conference on Data Engineering*, pages 3–14, March 1995.

[5] T.M. Anwar, H.W. Beck, and S.B. Navathe. Knowledge Mining by Imprecise Querying: A Classification-Based Approach. *Proceedings of the 8th International Conference on Data Engineering*, pages 622–630, February 1992.

[6] M.-S. Chen, J.-S. Park, and P. S. Yu. Efficient Data Mining for Path Traversal Patterns. *IEEE Transactions on Knowledge and Data Engineering*, 10(2), April 1998.

[7] J. Han, Y. Cai, , and N. Cercone. Knowledge Discovery in Databases: An Attribute-Oriented Approach. *Proceedings of the 18th International Conference on Very Large Data Bases*, pages 547–559, August 1992.

[8] J. Han, Y. Cai, and N. Cercone. Data Driven Discovery of Quantitative Rules in Relational Databases. *IEEE Transactions on Knowledge and Data Engineering*, pages 29–40, February 1993.

[9] J. Han and Y. Fu. Discovery of Multiple-Level Association Rules from Large Databases. *Proceedings of the 21th International Conference on Very Large Data Bases*, pages 420–431, September 1995.

[10] H. Lu, R. Setiono, and H. Liu. NeuroRule: A Connectionist Approach to Data Mining. *Proceedings of the 21th International Conference on Very Large Data Bases*, pages 478–489, September 1995.

[11] J.-S. Park, M.-S. Chen, and P. S. Yu. Using a Hash-Based Method with Transaction Trimming for Mining Association Rules. *IEEE Transactions on Knowledge and Data Engineering*, 9(5):813–825, October 1997.

[12] G. Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. In G. Piatetsky-Shapiro and W. J. Frawley, editors, *Knowledge Discovery in Databases*, pages 229–238. AAAI/MIT Press, 1991.

[13] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[14] J.R. Quinlan. Induction of Decision Trees. *Machine Learning*, 1:81–106, 1986.

[15] R. Srikant and R. Agrawal. Mining Generalized Association Rules. *Proceedings of the 21th International Conference on Very Large Data Bases*, pages 407–419, September 1995.

[16] J. T.-L. Wang, G.-W. Chirn, T.G. Marr, B. Shapiro, D. Shasha, and K. Zhang. Combinatorial Pattern Discovery for Scientific Data: Some Preliminary Results. *Proceedings of ACM SIGMOD, Minneapolis, MN*, pages 115–125, May, 1994.

[17] W. Ziarko. The discovery, analysis, and representation of data dependancies in databases. In G. Piatetsky-Shapiro and W. J. Frawley, editors, *Knowledge Discovery in Databases*, pages 195–209. AAAI/MIT Press, 1991.