NEUTRAL TRANSFORMATION AND DYNAMIC CLUSTERING BASED ON NEAREST NEIGHBOR ESTIMATION

[1]M.Srinivas Reddy, [2]R.Shiva Prasad, [3]K.Ravi Prakash, [4]D Jyothi
[1,2,3]Assistant Professor, [4]UG Student, [1,2,3,4]Department of Computer Science Engineering, Brilliant Grammar School Educational Society Group of Institutions Integrated Campus, Hyderabad, India

ABSTRACT

Many rule-based classifiers have a basic flaw in that they locate the rules by using different heuristic techniques to reduce the search space and then choose the rules based on the sequential database coverage paradigm. For categorizing categorical and sparse high dimensional datasets, rule-based classifiers work well. Even worse, in order to expand to huge databases, these algorithms fall short of fully using certain more efficient search space trimming techniques. Because of this, the final set of rules they employ might not be the ones that are generally the best for all instances in the training database. The technique that is being suggested creates a classifier by directly mining the final set of classification rules. It uses an instance-centric rule-generation approach and it can assure for each training instance, one of the highest-confidence rules covering this instance is included in the final rule set, which helps in improving the overall accuracy of the classifier. The proposed system, introduce novel search strategies and pruning methods into the rule discovery process, which has high efficiency and good scalability.

Key words: Data mining rule mining classifier , instance-centric , prune.

INTRODUCTION

Classification is one of the most fundamental data mining jobs, and many different classification methods have been presented. One of these subcategories is rule-based classifiers. With the training database, they create a model that consists of a collection of excellent rules that can be used to forecast the class labels of unlabeled cases. Many studies have demonstrated that categorical databases and sparse high dimensional databases, such those that arise in the context of document classification, can both be classified quite effectively using rule-based classification algorithms.

Certain conventional rule-based algorithms, such as FOIL, RIPPER, and CPAR, find a set of classification rules one at a time and use a sequential covering approach to omit the positive examples that are covered by each newly found rule from the training set. This rule induction process is done in a greedy fashion as it employs various heuristics (e.g.,information gain) to determine how each rule would be extended.

Due to this heuristic rule induction process and the sequential covering framework, the final set of discovered rules are not guaranteed to be the best possible. For example, due to the removal of some training instances, the information gain is computed based on the incomplete information; thus, the variable (or literal) chosen by these algorithms to extend the current rule will be no longer the globally optimal one. Moreover, for multi-class problems, these algorithms need to be applied multiple times, each time mining the rules for one class. If the training database is large and contains many classes, the algorithms will be inefficient.

Since the introduction of association rule mining, many association-based classifiers havebeen proposed. Some typical examples like CBA and CMAR adopt efficient association rule mining algorithms (e.g., Apriori and FP-growth) to first mine a large number of high-confidence rules satisfying a user-specified minimum support and confidence thresholds and then use various sequential-covering-based schemes to select from them a set of high-quality rules to be used for classification.

Since these schemes defer the selection step only after a large intermediate set of high- confidence rules have been identified, they tend to achieve somewhat better accuracy than the heuristic rule induction schemes.

However, the drawback of these approaches is that the number of initial rules is usually extremely large, significantly increasing the rule discovery and selectiontime.

The proposed algorithm of this thesis, overcome the problems of both the rule-induction-based and the association-rule-based algorithms. It directly mines for each training instance one of the highest confidence classification rules that it supports and satisfies a user-specified minimum support constraint, and builds the classification model from the union of these rules over the entire set of instances.

It employs an instance-centric rule generation framework and is guaranteed to find and includethe best possible rule for each training instance. Moreover, since each training instance usually supports many of the discovered rules, the overall classifier can better generalize to new instances and thus achieve better classification performance.

To achieve high computational efficiency, the proposed system mines the classification rules for all the classes simultaneously and directly mines the final set of classification rules bypushing deeply some effective pruning methods into the projection based frequent item set mining framework. All these pruning methods preserve the completeness of the resulting rule-setin the sense that they only remove from consideration rules that are guaranteed not to be of high quality..

LITERATURE REVIEW

Two classes of algorithms that are directly related to this work are the traditional rule- induction-based methods and the other is the recently proposed association rule based methods.Both of these classes share the same idea of trying to find a set of classification rules to build the model. The rule-induction-based classifiers like C4.5 [13], FOIL [14], RIPPER [4], and CPAR [15] use various heuristics such as information gain (including Foil gain) and gini index to identify the best variable (or literal) by which to grow the current rule, and many of them followa sequential database covering paradigm to speed up rule induction.

The association based classifiers adopt another approach to find the set of classification rules. They first use some efficient association rule mining algorithms to discover the complete (or a large intermediate) set of association rules, from which the final set of classification rules can be chosen based on various types of sequential database covering techniques. Some typical examples of association-based (or Emerging Pattern-based) methods include CBA [12], CAEP [7], CMAR [11], ARCBC [2], and DeEPs [10].

In contrast to the rule-induction-based algorithms, HARMONY does not apply any heuristic pruning methods and the sequential database covering approach. Instead, it follows an instance- centric framework and mines the covering rules with the highest confidence for each instance, which can achieve better accuracy. At the same time, by maintaining the currently best rules for each training instance and pushing deeply several effective pruning methods into the projection- based frequent itemset mining framework [9], [1], [8], HARMONY directly mines the final set of classification rules, which avoids the time consuming rule generation and selection process used in several association-based classifiers [12], [11], [5].

The idea of directly mining a set of high confidence rules is similar to those in [3], [6]. The author of [3] investigated a brute-force technique for mining the set of high-confidence classification rules, and proposed several effective pruning strategies to control the combinatorialexplosion in the number of rule candidates. The FARMER algorithm [6] finds the interestingrule groups for micro array databases. It mines the rules in a row enumeration space, and fully exploits some pruning methods to prune the search space based on the user-specified constraints like minimum support, confidence, and chi-square.

Unlike [3], [6], HARMONY does not need the user to specify the minimum confidence and/or chisquare. Instead, it mines for each training instance one of the highest confidence frequent rules that it covers. In addition, by maintaining the currently best classification rules for each instance, HARMONY is able to incorporate some new pruning methods under the unpromising item (or conditional database) pruning framework.

It has been proven very effective in pushing deeply the length decreasing support constraintor tough block constraints into closed itemset mining [15], [8]. In addition, recently we noticed that a similar approach [5] to this research was independently proposed at the same time frame, and showed its high accuracy in classifying gene expression data.

SYSTEM MODEL

The system model adapt the traditional projection-based frequent item set mining frameworkto efficiently enumerate the classification rules, then focus on how to push deeply some effectivepruning methods into the rule enumeration framework and give the whole algorithm

A. Classification Rule Enumeration

The projection-based item set enumeration framework has been widely used in many frequent item set mining algorithms and will be used as the basis in enumerating the classification rules. Given a training database TrDB and a minimum support min sup, it first computes the frequent items by scanning TrDB once, and sorts them to get a list of frequent items (denoted by f list) according to a certain ordering scheme.

Assume the min sup is 3 and the lexicographical ordering is the default ordering scheme, the f list computed is {a, b, c, d, e}. The algorithm applies the divide-and-conquer method plus the depth-first search strategy. It first mines the rules whose body contains item 'a', then mines the rules whose body contains 'b' but no 'a', ..., and finally mines the rules whose body containsonly 'e'.

In mining the rules with item 'a', item 'a' is treated as the current prefix, and its conditional database (denoted by TrDB|a) is built and the divide-and-conquer method is applied recursivelywith the depth-first search strategy. To build conditional database TrDB|a, HARMONY firstidentifies the instances in TrDB containing 'a' and removes the infrequent items, then sorts theleft items in each instance according to the f list order, finally TrDB|a is built. Following thedivide-and-conquer method, the proposed algorithm first mines the rules with prefix 'ab', thenmines rules with prefix 'ac' but no 'b', and finally mines rules with prefix 'ae' but no 'b' nor 'c'. During the mining process, when it gets a new prefix, it will generate a set of classification rules with respect to. the training instances covered by the prefix. For each training instance, it always maintains one of its currently highest confidence rules mined so far. It computes the covering rules according to the class distribution with respect to the prefix P.

B. Ordering of the Local Items

In the above rule enumeration process, we used the lexicographical ordering as an illustration to sort the set of local frequent items in order to get the f list. Many frequent item set mining algorithms either adopt item support descending order or support ascending order as the orderingscheme. However, because we are interested in the highest confidence rules w.r.t. the training instances, both the support descending order and ascending order may not be the most efficient and effective ways. As a result, we propose the following three new ordering schemes as the alternatives.

To mine the highest confidence covering rules as quickly as possible, a good heuristic is to sort the local frequent items in their maximum confidence descending order.

The widely used entropy to some extent measures the purity of a cluster of instances. If the entropy of the set of instances containing P is small, it is highly possible to generate some high confidence rules with body P{xj}. Thus another good ordering heuristic is to rank the set of localfrequent items in their entropy ascending order.

C. Search Space Pruning

Unlike the association-based algorithms, it directly mines the final set of classification rules. By maintaining the current highest confidence among the covering rules for each traininginstance during the mining process, some effective pruning methods can be proposed to improve the algorithm efficiency.

METHODOLOGY

A. Classification Rule Generation

Enumerate the classification rules under the divide-and-conquer and depth first search paradigm, and proposed several pruning methods to speed up the enumeration of the highest confidence covering rules. By integrating the pruning methods with the rule enumeration, we getthe classification rule generation algorithm.

The algorithm first initializes the highest confidence classification rules w.r.t. each training instance to empty, then enumerates the classification rules by calling subroutine ruleminer(Ø, TrDB). Subroutine rule miner takes as input a prefix itemset pi and its corresponding conditional database cdb. For each conditional instance, it checks if a classification rule with higher confidence can be computed from the current prefix pi, if so, it replaces the corresponding instance's current highest confidence rule with the new rule. It then finds the frequent local items by scanning cdb, prunes invalid items based on the support equivalence item pruning method and the unpromising item pruning method. If the set of valid local items is empty or the whole conditional database cdb can be pruned based on the unpromising conditional database pruning method, it returns directly.

Otherwise, it sorts the left frequent local items according to the correlation coefficient ascending order, and grows the current prefix, builds the conditional database for the new prefix, and recursively calls itself to mine the highest confidence rules from the new prefix.

B. Building the Classification Model

After the set of highest confidence covering rules have been mined, it will be straightforward to build the classification model. HARMONY first groups the set of highest confidence covering rules into k groups according to their rule heads (i.e., class labels), where k is the total number ofdistinct class labels in the training database. Within the same group of rules, the algorithm sorts the rules in their confidence descending order, and for the rules with the same confidence, sorts them in support descending order. In this way, it prefers the rules with higher confidence and the rules with higher support if the confidence is the same.

C. New Instance Classification

After the classification model, CM, has been built, it can be used to classify a new test instance, ti, using the New Instance Classification algorithm. It first computes a score w.r.t. ti for each group of rules in CM, and predicts for ti a class label or a set of class labels if the underlying classification is a multi-class multi-label problem (i.e., each instance can be associated with several class labels).

D. Multi-class multi-label classification

The dominant factor-based method was proposed to predict the class labels for a multi-class multi-label classification problem. However, in many imbalanced classification problems, the average confidence of each group of classification rules may be quite different from each other, this uniform dominant factor based method will not work well. A large dominant factor may leadto low recalls (i.e., the percentage of the total test instances for the given class label that are correctly classified) for the classes with low average rule confidences, while a small dominant factor can lead to low precisions (i.e., the percentage of predicted instances for the given class label that are correctly classified) for the classes with high average rule confidences. To overcome this problem, the proposed algorithm adopts a weighted dominant factor-based method.

EXPERIMENTAL EVALUATION

The experimentation is done to obtain an accurate and efficient rule-based classifier with goodscalability, which should be able to overcome the problems of both the traditional rule based and the recently proposed association-based classifiers. Instead of using the sequential databasecovering to select the rules, our solution mines a set of high quality rules in an instance-centric manner and can assure that at least one of the highest confidence frequent covering rules (if thereis any) w.r.t. any training instance is included in the final result set of classification rules.

Specifically, given a training database TrDB and a minimum support threshold min sup, the problem of this study is to find one of the highest confidence frequent covering rules for each of the training instances in TrDB, and build a classifier from these classification rules.

The input training database must be in the form that is consistent; otherwise, the training database should be first converted to that form. For example, a numerical database should be firstdiscretized into a categorical one in order to use to build the model. In addition, although this study mainly focuses on mining any one of the highest confidence frequent covering rules for each training instance, it is straightforward to revise to mine the complete set of the highest confidence frequent covering rules or K highest confidence frequent covering rules for each training instance.

Implemented the proposed algorithm in Java and performed a thorough experimental study(Fig: 1, Fig: 2, Fig: 3). We first evaluated the system as a frequent item set mining algorithm to show the effectiveness of the pruning methods, the algorithm efficiency and scalability. Then we compared it with some well-known classifiers on both categorical and text databases. All the experiments except the scalability test were performed on a 2.6GHz Windows machine with1GB memory.

Many previous studies used some small databases to evaluate both the accuracy and efficiencyof a classifier. For example, most of the databases used only contain several hundred instances, which means the test databases contain too few test instances (i.e., only a few tens) if the 10-fold cross validation is adopted to evaluate the classification accuracy. In this thesis, main focus is made on relatively large databases (by large, we mean the database should contain no fewer than 1000 instances), although it report the comparison results for some small databases.

CONCLUSION

Designing accurate, efficient, and scalable classifiers is an important research topic in data mining, and the rule based classifiers have been proven very effective in classifying the categorical or high-dimensional sparse data. However, to achieve high accuracy, a good rule- based classifier needs to find a sufficient number of high quality classification rules and use themto build the model.

The proposed algorithm is an instance-centric classification rule mining paradigm and designed an accurate classifier. Several effective pruning methods and search strategies have alsobeen proposed, which can be pushed deeply into the projection-based frequent item set enumeration framework. The performance study shows that the proposed algorithm has high accuracy and efficiency in comparison with many well known classifiers for both the categorical data and the high dimensional text data. It also has good scalability in terms of the base size.

REFERENCES

1. R. Agarwal, C. Aggarwal, V. Prasad. A Tree Projection Algorithm for Generation ofFrequent Item Sets, Journal of Parallel and Distributed Computing. 61(3), 2001.
2. M. Antonie, O. Zaiane. Text Document Categorization by Term Association, ICDM’02.
3. R.J. Bayardo. Brute-force Mining of High-confidence Classification rules, KDD’97.
4. W. Cohen. Fast effective rule induction, ICML’95.
5. G. Cong, K. Tan, A. Tung, X. Xin. Mining Top-k Covering Rule Groups for GeneExpression Data, SIGMOD’05.
6. G. Cong, X. Xu, F. Pan, A. Tung, J. Yang. FARMER: Finding Interesting Rule Groups in Microarray Datasets, SIGMOD’04.
7. G. Dong, X. Zhang, L. Wong, J. Li. CAEP: Classification by aggregating emerging patterns,DS’99.

8. K. Gade, J. Wang, G. Karypis. Efficient Closed Pattern Mining in the Presence of Tough Block Constraints, KDD'04.
9. Han, et al. Mining Frequent Patterns without Candidate Generation, SIGMOD'00.
10. [10]J. Li, G. Dong, K. Ramamohanarao, L. Wong. DeEPs: A New Instancebased Discovery and Classification System, Machine Learning, 54(2), 2004.
11. [11]W. Li, J. Han, J. Pei. CMAR: Accurate and Efficient Classification based on multiple class-association rules, ICDM'01
12. [12]S B. Liu, W. Hsu, Y. Ma. Integrating Classification and Association Rule Mining, KDD'98.. [13]J. Quinlan. C4.5: Programs for Machine Learning, Morgan Kaufman, 1993.
13. [14]Y. Yang. An Evaluation of Statistical Approaches to Text Categorization, InformationRetrieval, Vol. 1, No. 1-2, 1999.
14. [15] X. Yin, J. Han. CPAR: Classification based on Predictive Association Rules, SDM'03