Analysis of Rule Ranking and Rule Pruning in Associative Classification Technique
Ravi Patel¹ Jay Vala² Kanu Patel³
¹²Information Technology, GCET
³Information Technology, BVM

Abstract--- In Data Mining, Association rule mining and classification are most important tasks for decision making process based on data. Single-Label classification only predict the single class label but nowadays most of the Application require more than one class label for Prediction eg. Music and text categorization, Medical diagnosis. Above issue can be solved by Multi-label classification But the disadvantage of that is in this technique due to more number of label number of rules generated by the algorithm is too large and redundant. So discovery of strong rules is more harder this problem can be solved by the ranking and pruning method using Parameter like Entropy and Information Gain. Still not giving good result so future work is we can apply pruning using the factors like Confidence and database Coverage. In this paper some of the techniques for rule ranking and pruning are explained for associative classification.

I. INTRODUCTION
Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks.[2,3] In general, data mining tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions.
- Various types of task of data mining
  1. Classification
  2. Clustering
  3. Association Rule Discovery
  4. Sequential Pattern Discovery
  5. Regression
  6. Deviation Detection

A. Classification
In classification[1], by the help of the analysis of training data we develop a model which used to predict the class of objects whose class label is unknown. The model is trained so that it can distinguish different data classes. The training data is having data objects whose class label is known in advance.
Classification analysis is the Also known as supervised classification, uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. Whilst single-label classification, which assigns each rule in the classifier the most obvious class label, has been widely studied [8] little work has been conducted on multi-label classification. The classification algorithm learns from the training set and builds a model. This model is used to classify new unclassified data objects. For example, after starting a credit policy, the OurVideoStore manager rs could analyze the customers’ behaviours of their credit, and label accordingly the customers who received credits with three possible labels “safe”, “risky” and “very risky”. The classification analysis would generate a model that could be used for acceptance or rejection of credit requests in the future. Many techniques for classification are neural network[7], rule based classifier[4], Bayesian network (BN)[5], Decision tree[6].

B. Association Rule Mining
Association rule mining is the discovery of association rules. Association rule mining, one of the most important and well researched techniques of data mining, was first introduced in [1]. It studies the frequency of items occurring together in transactional databases, and according to support count, detects the frequent item sets. Another threshold, confidence, which is the conditional probability than an item appears in a transaction when another item appears, is used to pinpoint association rules. Association rule mining is used for market basket analysis. For example, it could be useful for the OurVideoStore manager to know what movies are often rented together or if there is a relationship between renting a movies and buying chips or popcorn. The discovered association rules are of the form: A →B [s,c], where A and B are conjunctions of attribute value-pairs, and s (for support) is the probability that A and B appear together in a transaction and c (for confidence) is the conditional probability that B appears in a transaction when A is in transaction. For example, the association rules:
- Rent Type(X, “game”) ∧ Age(X, “13-19”) →Buys(X, “chips”) \( [s=2\%, c=55\%] \)

would indicate that 2% of the transactions considered are of customers aged between 13 and 19 who are renting a game and buying a chips, and that there is a certainty of 55% that teenage customers who rent a game also buy chips.

The two basic parameters of Association Rule Mining (ARM) are: support and confidence:
Support(s) of an association rule is defined as the percentage/fraction of records that contain \( X \cup Y \) to the total number of records in the database. The count for each item is increased by one every time the item is encountered in different transaction \( T \) in database \( D \) during the scanning process. Support(s) is calculated by the following
\\
Support (XY) = \frac{Support count of XY}{Total number of transaction in D}
\\
Suppose the support of an item is 0.1%, it means only 0.1 percent of the transaction contain purchasing of this item.
Confidence of an association rule is defined as the percentage/fraction of the number of transactions that contain \( X/Y \) to the total number of records that contain \( X \), where if the percentage exceeds the threshold of confidence an interesting association rule \( X \rightarrow Y \) can be generated.
\\
Confidence(X/Y) = \frac{Support (XY)}{Support(X)}
Confidence is a measure of strength of the association rules, suppose the confidence of the association rule X/Y is 80%, it means that 80% of the transactions that contain X also contain Y together, similarly to ensure the interestingness of the rules specified minimum confidence is also pre-defined by users.

Association rule mining is to find out association rules that satisfy the pre-defined minimum support count and confidence threshold from a given database. Apriori and FP-Growth are most popular algorithm for association rule mining[8].

C. Associative Classification
Recent studies propose the extraction of a set of high quality association rules from the training data set which satisfies certain user-specified frequency and confidence thresholds. Effective and efficient classifiers have been built by careful selection of rules. e.g., CBA [9], CAEP [10], and ADT [11]. Such a method takes the most effective rule(s) from among all the rules mined for classification. Since association rules explore highly confident associations among multiple variables, it may overcome some constraints introduced by a decision-tree induction method which examines one variable at a time. Extensive performance studies [9, 10, 11] show that association based classification may have better accuracy in general. Some rule ranking and pruning are used by these types of techniques.

II. METHODS FOR RULE PRUNING AND RANKING
Associative classification is combination of association rule and classification. this is highly accurate classifier than traditional classification techniques like greedy and decision tree. However, the size of the classifiers produced by associative classification algorithms is usually large and contains insignificant rules. This may degrade the classification accuracy and increases the classification time, thus, pruning becomes an important task. This is used for finding relationship between attribute in large database,[12]

Problem:- X \rightarrow Y Here Y must be class attribute[12,14]

III. APPROACHES OF RULE PRUNING AND RANKING
A. Text Pruning Method.[12]
- Database Coverage:- the method in which post pruning technique is used . Pruning after rules are created.
- Discard specific rules with the less confidence value then general rule is called as redundant rule pruning.
- Discard the rules that incorrectly classify training data. Before pruning starts, the rules must be sorted in descending manner according to confidence, support, and number of items in the rule antecedent. According to the above method rule will be ranked.
- Important Parameter:- Confidence level, support level and number of items
  - in the rule antecedent. We can also Improve the performance of association classifiers by rule Prioritization method.[15]
- Re-rank the rule execution order of CAR using rule priority this can also improve the accuracy.
  - In general, the ranking of CARs is conducted according to confidence in descending order. Then, CARs with the same confidence are ranked according to the support value in descending order. if the CARs with the the class assignment (prediction) method makes a group of rule prediction instead of utilizing only a single rule it increase the accuracy.
- Require condition: rule on right hand side (consequent) is the class label,
- and the rule on left hand side (antecedent) is attribute values [12,14].
- same confidence level as well as same support level they rank according to
- rule length with short rule to long rule.

IV. ALGORITHM
Let r1 and r2 be two rules. Then, r1 precedes r2, denoted as r1 > r2, if
- conf(r1) > conf(r2), or
- conf(r1) = conf(r2) and sup(r1) > sup(r2), or
- conf(r1) = conf(r2) and sup(r1) = sup(r2) and len(r1) > len(r2), or
- conf(r1) = conf(r2) and sup(r1) = sup(r2) and len(r1) = len(r2) and
  lex(r1) > lex(r2)
- here conf(r) and sup(r) are the confidence level and support level of r, respectively, len(r) used for showing the number of items in the rule r, and lex(r) used for position of rule’s item according to lexicographical order of items.
- Some time spare rules may be useful in testing time .
- If there is no spare rule then it classifies that into the default class.
  - This Approach is called as L3 approach.
  - Level I is for classifying documents based on rules which can classify at least one training document correctly.
  - Level II is for classifying documents based on spare rules which classify no documents during the training process.
  - Level III is for classifying documents into a default class.[15]

1. We can also improve the accuracy of the association classifier assigning weight
There are many domains such as medical, where the maximum accuracy of the model is desired and hence the accuracy of the associative classifiers. We use weighted association rule mining (WARM). All predicting class label not have same importance in any prediction techniques. So different weights can be assigned to different attributes according to their predicting capability. That takes advantage of weighted association rule mining. This is used for improving the accuracy of class prediction. Here each and every parameter as assigned some weight. So we have to calculate weighted confidence and support of the derived rule,[13]

2. Rule ranking method using mutually associated Pattern
- rule generation:-Rule generation using CUA
- rule Pruning:-we need pruning to delete redundant and noisy information.
- Delete all negative rule concept we can used. And support and confidence also will be used.
- rule Ranking:-above parameter like confidence ,support and fever condition
- in the left hand side of the rule will be selected first.[14,15]
3. We can also calculate the rank of the rule using below equations [16]
   - \( \text{rank}(r) = \sup(r) \times \text{conf}(r) \),
   - \( \text{rank}(r) = \frac{\text{conf}(r)}{\sup(r)} \),
   - \( \text{rank}(r) = \frac{\sup(r)}{\text{conf}(r)} \),
   - \( \text{rank}(r) = \sup(r) + \text{conf}(r) \),
   - \( \text{rank}(r) = \sup(r) - \text{conf}(r) \),
   - \( \text{rank}(r) = \text{conf}(r) - \sup(r) \).

so using these type of technique we can improve classification technique. So these methods are useful for ranking and pruning of redundant rules generated by association rule mining

V. CONCLUSION

Associative classification is most useful technique for world wide application, mostly for multilabel classification. But if database is large then it’s too much hard to handle large amount of rules. So that using rule ranking and rule pruning method some of redundant rules can be pruned so that accuracy and effectiveness are achieved. Also number of rules is reduced

REFERENCES

[14] Yuhani Yusof, Mohammed Hayel Refai "MMCAR: Modified Multi-class Classification based on Association Rule" 978-1-4673-1090-1/12/2012 IEEE
[16] Guangfei Yang, Shingo Mabu, Kaoru Shimada, and Kotaro Hirasawa1 "A New Associative Classification Method by Integrating CMAR and RuleRank Model based on Genetic Network Programming",Fukuoka International Congress Center, Japan,August 18-21, 2009,