CLASSIFICATION ON MULTI-LABEL DATASET USING RULE MINING TECHNIQUE

Ravi Patel[1], Jay Vala[2], Kanu Patel[3]

[1]*Information Technology, GCET*
[2]*Information Technology, GCET*
[3]*Information Technology, BVM*

Abstract

Most recent work has been focused on associative classification technique. Most research work of classification has been done on single label data. But it is not appropriate for some real world application like scene classification, bioinformatics, and text categorization. So that here we proposed multi label classification to solve the issues arise in single label classification. That is very useful in decision making process. Multi-label classification is an extension of single-label classification, and its generality makes it more difficult to solve compare to single label. Also we proposed classification based on association rule mining so that we can accumulate the advantages of both techniques. We can get the benefit of discovering interesting rules from data using rule mining technique and using rule ranking and rule pruning technique we can classified that rules so that redundant rules can be reduced. So that Here proposed work is done on multi label dataset that is classified using rule mining algorithm. Here proposed approach is an accurate and effective multi label classification technique, highly competitive and scalable than other traditional and associative classification approaches.

----------***----------

1. INTRODUCTION

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks.[1,2] 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 are: Classification, Clustering, Association Rule Discovery, Regression etc.

2. 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 [3]. 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 provisional 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.

Two main algorithms: apriori algorithm and FP-Growth algorithm
Here we are using FP-Growth algorithm instead of apriori for associative classification. some advantages of FP-Growth :Uses compact data structure ,Eliminates repeated database scan[12]

3. CLASSIFICATION

In classification [4], 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 [9] 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 manage rs could analyze the customers’

behaviors 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 [8], rule based classifier[5], Bayesian network (BN)[6], Decision tree[7].

3.1 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 [14], CAEP [15], and ADT [16]. 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 [14, 15, 16] show that association based classification may have better accuracy in general.

3.2 Multi-Label Classification [13]

For the growing quantity of data in fields such as bioinformatics, scene classification and text categorization, multi-label classification is more appropriate In multi-label classification is a variant of the classification problem where multiple target labels must be assigned to each instance. Multi-label classification should not be confused with multiclass classification, which is the problem is categorizing instances into more than two classes.

There are two main methods for tackling the multi-label classification problem. Problem transformation methods and algorithm adaptation methods. Problem transformation methods transform the multi-label problem into a set of binary classification problems. Algorithm adaptation methods adapt the algorithms to directly perform multi-label classification.

3. PROPOSED WORK

Here proposed work consists of following steps:

1. Discover the set of frequent items.
2. Rule generation using threshold value of confidence and support.
3. Re-rank the execution order of rule using rule ranking technique.
4. Prune the redundant rule using some pruning technique.
5. Generate prune rule.
6. Predict test case using the prune rule.

In Step 3 Two Techniques is used to improve the existing algorithm. There are two techniques is used in implementation of the proposed algorithm which are described below.

1. Rule Pruning Technique
2. Rule Ranking Technique

Before pruning starts the rules must be sorted in according to descending order in confidence, support, and number of items in the rule antecedent.

Rule Ranking Technique of CaR is conducted according to confidence in descending order. Then, CaR with the same confidence are ranked according to the support value in descending order. if the CaR with the same confidence level as well as same support level they rank according to rule length with long rule to short rule.

Rule Pruning Technique Here find the rule which satisfy the if rule on right hand side (consequent) is the class label, and the rule on left hand side (antecedent) is attribute values[10,11].

Data Set format: Suppose One Dataset Weather has following nominal attributes with values.

Data Set Weather:

No	Outlook	temperature	Humidity	Windy	Play
1	Sunny	HOT	HIGH	False	no
2	Sunny	HOT	HIGH	True	no

(a). Data set with nominal attributes

No	Outlook	temperature	Humidity	Windy	Play
1	1	6	7	10	12
2	1	6	7	9	12

(b). Converted into numeric data set

Data Set Format

Here we have to take unique numeric value for every value of attributes.

- Outlook=Sunny=1, Outlook=Overcast=2, Outlook=Rainy=3
- Temperature=hot=4, temperature=mild=5, temperature=cool=6
- Humidity=high=7, humidity=normal=8
- Windy=TRUE=9, windy=FALSE=10
- Play=yes=11, play=no=12

3.1 Implementation

For that implementation we are not using weka tool because there is no provision in weka to set more than one attribute as a class-label. Means weka tool is not supporting to multilabel classification. That's why here we are using net beans for generating association rule from FP-Growth algorithm.

3.2 Function of Proposed Algorithm

In Existing Algorithm there are the functions which are used to generate association rules those are given below.

- Outputsrules() is used for calling Outputsrule().
- Outputsrule() is used for displaying the rules.
- OutputsItemSet() is used for rule format how will be display from itemset.
- InserrulesintoRulelist() is used for inserting rules as per Descending order of confidence.

Changes made in the functions of proposed algorithm:
The Function Outputsrules() which is used to generate Association rules based on threshold of Confidence value and Support value Parameter of the algorithm.

So, in this proposed pruning method we have implemented one new function Outputsrules1() for generating association rules which satisfy our condition :

Condition: Right side of the rule:-Class Attribute
Left Side of the rule:-Simple Attribute

In this Outputsrule1() function First of all we are checking all the attribute at the right side of the rule if right side of the rule is class label then this rule will be consider as a classified rule. If the rule satisfy above condition then we will check left hand side of the rule if left hands side of rule have Simple attribute then this rule will be generate after this pruning step Using InserrulesintoRulelist(), Rules are sorted according to Descending order of Confidence value. if Confidence is same then it will check Support of the rule, if support of the rules are also same then it will check for the no of items in antecedent of the rule. Means according to length rules are sorted and pruned.

4. RESULTS AND ANALYSIS

Here we have taken two data sets for associative classification. We have shown graphs for no. of rules reduction and accuracy before and after applying our proposed Classification approach.

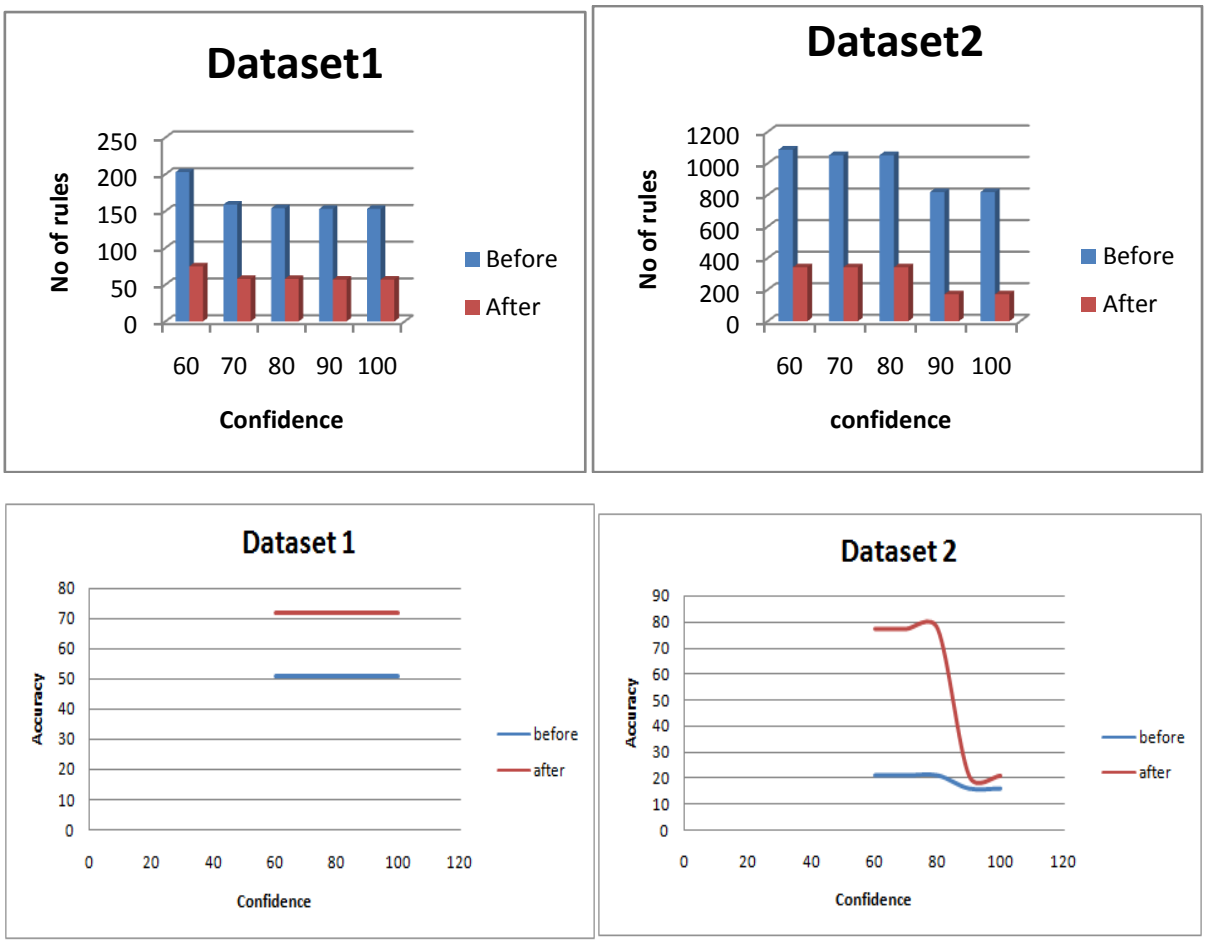
Number of Rules with Different Confidence Level

Table 1 Accuracy with different confidence level

	Confidence									
	60		70		80		90		100	
	Before	After	Before	After	Before	After	Before	After	Before	After
DataSet1	203	75	159	58	154	58	153	57	153	57
DataSet2	1087	342	1052	342	1052	342	818	171	818	171

Table 2 Here graphs shows no. Of rules reduction and accuracy

	Confidence									
	60		70		80		90		100	
	Before	After	Before	After	Before	After	Before	After	Before	After
DataSet1	51.72	72.42	51.72	72.42	51.72	72.42	51.72	72.42	51.72	72.42
DataSet2	20.83	77.08	20.83	77.08	20.83	77.08	16	20.83	16	20.83



These results show that if value of confidence is increase then the numbers of rules are decreased but also the accuracy of the algorithm is decrease. After 80% confidence in dataset2 the accuracy of the algorithm is decreased drastically and in dataset 1 no more reduction in the number of rules after confidence level 80%. So the better value of confidence for the datasets is 80%.

5. CONCUSIONS AND FUTURE WORK

Our Proposed work of Multi-label Associative Classification algorithm uses two methods called rule ranking and rule Pruning, for ranking and removing unnecessary rules which are generated by the FP-Growth association rule mining algorithm. So, we can able to increase the accuracy of FP-Growth algorithm. Using this technique we are able to overcome the problems of single label classification and redundant rules generation.

Our proposed multi label classification algorithm could be improve for executing dataset which have character or Nominal attribute. And also improve the accuracy of this algorithm by applying heuristic search methods of artificial intelligence to discover more informative rules Or generated rule could be passed in the genetic algorithm as an initial population to achieve more accuracy.

REFERENCES

[1] Shelly Gupta , Dharminder kumar ,Anand Sharma "data mining classification techniques applied for breast cancer diagnosis and prognosis", ijcse vol. 2 no. 2 ,2011

[2] Tipawan Silwattananusarn and Dr. KulthidaTuamsuk "Data Mining and Its Applications for Knowledge Management : A Literature Review from 2007 to 2012" IJDKP Vol.2, No.5, September 2012

[3] Agrawal, R., Srikant, R.: Fast algorithms for mining association rule. In: Proceedings of the 20th International Conference on Very Large Data Bases, pp. 487–499 (1995)

[4] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. JohnWiley & Sons, 1973

[5] G.L. Pappa and A.A. Freitas, Automating the Design of Data Mining Algorithms. An Evolutionary Computation Approach, Natural Computing Series, Springer, 2010

[6] G.F. Cooper, P. Hennings-Yeomans, S.Visweswaran and M. Barmada, "An Efficient Bayesian Method for Predicting Clinical Outcomes from Genome-Wide Data", AMIA 2010 Symposium Proceedings, 2010, pp. 127-131

[7] M. Garofalakis, D. Hyun, R. Rastogi and K. Shim, "Building Decision Trees with Constraints", Data

Mining and Knowledge Discovery, vol. 7, no. 2, 2003, pp. 187 – 214

[8] Y. Singh Y, A.S. Chauhan, "Neural Networks in Data Mining", Journal of Theoretical and Applied Information echnology, 2005, pp. 37-42

[9] B.Santhosh Kumar, K.V.Rukmani .,"Implementation of web usage mining using APRIORI and FP growth algorithms" IJANA'2010

[10] Fadi Thabtah,Wa'el Hadi,Hussein Abu-Mansour,L. McCluskey "A New Rule Pruning Text Categorisation Method ,7th International Multi-Conference on Systems, Signals and Devices,2010.

[11] Yuhanis Yusof, Mohammed Hayel Refai "MMCAR: Modified Multi-class Classification based on Association Rule"978-1-4673-1090-1/12/2012 IEEE

[12] B.Santhosh Kumar, K.V.Rukmani .,"Implementation of web usage mining using APRIORI and FP growth algorithms" IJANA'2010

[13] Grigorios Tsoumakas, Ioannis Katakis:"Multi-Label Classification: An Overview "International Journal of Data Warehousing and Mining, 3(3), 1-13, July-September 2007

[14] B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. In *KDD'98*, New York, NY, Aug.1998.

[15] G. Dong, X. Zhang, L. Wong, and J. Li. Caep: Classification by aggregating emerging patterns. In *DS'99 (LNCS1721)*, Japan, Dec. 1999.

[16] K. Wang, S. Zhou, and Y. He. Growing decision tree on support-less association rules. In *KDD'00*, Boston, MA,Aug. 2000.