Rule Pruning with Correctly Classify in Associative Classification

Mr. Jogendraprasad Bagdi[1]
[1]Department of Computer Science
[1]Gujarat Technological University, India

Abstract— Recent studies in data mining revealed that Associative Classification (AC) data mining approach builds competitive classification classifiers with reference to accuracy when compared to classic classification approaches including decision tree and rule based. Nevertheless, AC algorithms suffer from a number of known defects as the generation of large number of rules which makes it hard for end-user to maintain and understand its outcome and the possible over-fitting issue caused by the confidence-based rule evaluation used by AC. This thesis attempts to deal with reduces the number of generated rules without having large impact on the prediction rate of the classifiers. In this paper proposed method FPCC (Full And Partial Rule Coverage Correctly Classifier) is used to discover the rules which are fully matched and partially match with correct classifier.

Key words: data mining, Rule Pruning, Associative Classification

I. INTRODUCTION

Knowledge Discovery and Data Mining (KDD) is playing an important role in extracting knowledge in this era of data overflow. KDD consists of many methods and techniques that can be applied to different data to extract knowledge. Some of the methods include association, classification, and clustering.

II. LITERATURE SURVEY

This chapter covers the studies and work related to the topic.

A. Classification based on Association Rules (CBA)

This algorithm first generates all the association rules and then selects a small set of rules to form the classifiers. When predicting the class label for a coming sample, the best rule is chosen. It consists of two parts, a *rule generator* (called CBA-RG), which is based on algorithm Apriori for finding association rules and a *classifier builder* (called CBA-CB).

B. CMAR (Classification based on Multiple Association Rules)

The associative classification suffers from the huge set of mined rules and sometimes biased classification or over fitting because the classification is done based on only single high-confidence rule. This associative classification method, CMAR (Classification based on Multiple Association Rules) is proposed in which the classification is performed based on a weighted analysis using multiple strong association rules. The classification is performed based on a weighted X^2 analysis using multiple strong association rules.

C. CPAR (Classification Based On Predictive Association Rules)

The CPAR combines the advantages of both associative classification and traditional rule-based classification. Instead of generating a large number of candidate rules CPAR adopts a greedy algorithm to generate rules directly from training data. To avoid over fitting it uses expected accuracy to evaluate each rule and uses the best k rules in prediction.

D. MCAR (Multi-Class Classification Based On Association Rule)

This algorithm consists of two main phases: rule generation and a classifier builder. In the first phase, the training data set is scanned once to discover frequent one rule items, and then MCAR recursively combines rule items generated to produce potential frequent rule items (candidate rule items) involving more attributes. The supports and confidences for candidate rule items are calculated simultaneously, where any rule item with support and confidence larger than minimum support and minimum confidence, respectively, is created as a potential rule. In the second phase, rules created are used to build a classifier by considering their effectiveness on the training data set. Only effective rules will be kept in the final classifier.

III. MOTIVATION

MCAR produces classifiers with slightly more rules than current AC techniques, resulting in reduced error rate. MCAR algorithm is highly competitive when compared with traditional classification algorithms such as RIPPER, C4.5 and scales well compared with popular AC like CBA with regards to prediction power, rules features and efficiency.

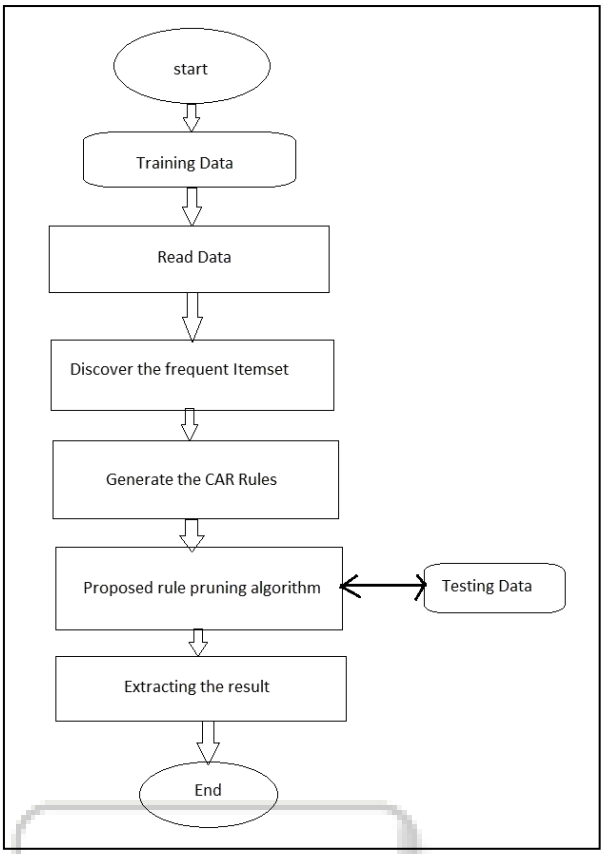
IV. PROPOSED METHODOLOGY

A. The AC Model Working:

The user selects the training dataset and then defines the required thresholds (minimum support and minimum confidence).

The AC system starts processing the training data by producing the complete frequent rule items set and then produce the set of Class Association Rules.

A subset of these rules is selected (interesting ones) through evaluation procedure. The selected rules are then used to form the classifier. Lastly, the classifier is applied on the testing dataset. Here, we developed Rule pruning algorithm which mines single label cases (each case is assigned to one class label only).



B. Proposed Algorithm

Input: Trainnig datased T, minsup, minconf
Output: classifier Cl

- Read for the set F1 frequent 1-ruleitem
- j←1
- Repeat while until Fj is null
- Fj+1←gen(Fj) //IT GOES TO GENERATE FUNCTION
- TempCl←TempCl U Fj+1
- j= j+1
- goto step 4
- Rank(TempCl) //IT GOES TO RANKING FUNCTION
- Cl←Evaluate (TempCl on T)
- Output Cl

C. Generate Function

- Input: set of ruleitem S
- Output: A set S' produced ruleitems
  - S'←0
  - Do
  - for each pair of disjoint i1, i2 in S Do
  - if(<i1 U i2>, c) passes the minsup
  - if(<i1 U i2>, c) passes the minconf
  - S'← S' U (<i1 U i2>, c)
  - endif
  - endif
  - end
  - end
  - Return S'

D. Rank Function:

Given two rules, *ra* and rb, *ra* precedes *rb*(*ra*>*rb*) if:

- The confidence of *ra* is greater than that of *rb*.
- The confidence values of *ra* and *rb* are the same, but the support of *ra* is greater than that of *rb*.
- Confidence and support values of *ra* and *rb* are the same, but *ra* has fewer conditions in its left hand side than of *rb*.
- Confidence, support and cardinality of *ra* and *rb* are the same, but *ra* is associated with a more representative class than that of *rb*.
- All above criteria are identical for *ra* and *rb*, but *ra* generated from an items and columns that have higher order than that of *rb*.

E. Evaluate:

- Input: Training dataset T and set of ranked rules R
- Output: Classifier cl
- For Every rule ri in R
- if ri fully match a training case and Correctly classify it then
- Insert the rule at the end of cl
- Remove all training cases in T covered by ri.
- else if r partially match at least a single case then and the class is matched
- insert the rule at the end of cl
- remove all cases in T covered by ri else
- Discard ri and remove it from R
- end
- Next r

1) Features of the proposed Algorithm

MCAR consider a rule significant during building the classifier if it's fully and correctly cover a training instance. Proposed algorithm employs a new rule evaluation which considers the rule significant if it's partially covers training cases.

V. RESULTS AND ANALYSIS

A. Class Distribution on Dataset

Dataset	No Of Attribute	No Of Class	Class Distribution	No Of Cases
Tic-Tac	10	2	65% 35%	958
Contact	5	3	65% 35%	24
Wheather	5	2	65% 35%	14
Zoo	18	7	65% 35%	101

B. Accuracy on Different Dataset

Dataset	C 4.5	Ripper	PRMF	FPCC(Proposed Method)
Tic-Tac	83.23	98.80	99	99.5
Contact	66.66	66.66	66.66	66.66
Wheather	60	60	78	80
Zoo	94.28	88.57	92	92.5

Above Dataset, C4.5 and Ripper Algorithm is Tested on Weka. And PRMF (Partial Rule Matching Filtering) and FPCC (Full and Partial Rule Coverage Correctly Classifier). In these Datasets PRMF and FPCC method is applied with rule Support and Confidence is 20% and 40% respectively.

C. No of Rules on Different Dataset

Data	C4.5	Ripper	MCAR	PRFM	FPCC
Tic-Tac	95	9	27	4	4
Contact	4	3	53	7	6
Zoo	9	6	8	7	7
weather	3	4	6	3	4

VI. CONCLUSION

Associative classification technique is most useful for multi label classification. 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.

VII. FUTURE WORK

An excessive CPU time is required during the process of discovering the frequent items, generated the rule and learns the classifier which impacts the efficiency. Accuracy may achieve more if there is Class Distribution is Exact.

REFERENCES

[1] A Survey on Algorithms for Market Basket Analysis , By Gajalakshmi & M. S. Murali Dhar IJCSM December 2013
[2] Improving rule sorting, predictive accuracy and training time in associative classification Fadi Thabtah a,*, Peter Cowling b, Suhel Hammoud c F. Thabtah et al. / Expert Systems with Applications xx (2005) 1–13
[3] A New Class Based Associative Classification Algorithm Author: Zhonghua Tang and Qin Liao(IAENG International Journal of Applied Mathematics, 36:2, IJAM_36_2_3)
[4] A Survey of Associative Classification Algorithms By Sohil Gambhir, Prof. Nikhil Gondliya(IJERT NOV 2012)
[5] PARTIAL RULE MATCH FOR FILTERING RULES IN ASSOCIATIVE CLASSIFICATION Journal of Computer Science 10 (4): 570-577, 2014 ISSN: 1549-3636 © 2014 Science Publications
[6] Comparative Analysis of Different Techniques in Classification Based on Association Rules By Nitendra Kumar Vishwakarma, Jitendra Agrawal, Shikha Agrawal, Sanjeev Sharma IEEE 2013
[7] A review of associative classification mining by Thabtah, Fadi Abdeljaber 2007
[8] Data Mining Concepts and Techniques by Jiawei Han & Micheline Kamber.Pulication Elseiver.
[9] Multi-class Classification based on Association Rule, Proceedings of the ACS/IEEE 2005 International Conference on Computer Systems and Applications Author: Fadi Thabtah , Peter Cowling , Yonghong Peng.
[10] A review of associative classification mining by Thabtah, Fadi Abdeljaber
The Knowledge Engineering Review, Vol. 22:1, 37–65. 2007, Cambridge University Press doi: 10.1017/S0269888907001026 Printed in the United Kingdom
[11] Pruning Techniques in Associative Classification: Survey and Comparison by Fadi Thabtah
[12] Analysis of Rule Ranking and Rule Pruning With Correctly Classify In Associative Classification by Ravi Patel, Jay Vala, Kanu Patel in (IJSRD/Vol. 2/Issue 03/2014/244

