ISSN No. 0976-5697

Volume 4, No. 8, May-June 2013

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Implementation of the Associative Classification Algorithm and Format of Dataset in Context of Data Mining

Gajraj Singh
Research Scholar, NIMS University, Jai[pur, India
singhgajraj83@gmail.com

DR. P.K. YADAV
Associate Professor G.C.
Kanwali (Rewari) India

Abstract: Construction of classification models based on association rules. Although association rules have been predominantly used for data exploration and description, the interest in using them for prediction has rapidly increased in the data mining community. In order to mine only rules that can be used for classification, I had modified the well known association rule mining algorithm Apriori to handle user-defined input constraints. We considered constraints that require the presence/absence of particular items or that limit the number of items in the antecedents and/or the consequents of the rules. We developed a characterization of those item sets that will potentially form rules that satisfy the given constraints. This characterization allows us to prune during item set construction. This improves the time performance of item set construction. Using this characterization, we implemented a classification system based on association rules. Furthermore, I enhanced the algorithm by relaying on the typical support/confidence framework, and mining for the best possible rules above a user-defined minimum confidence and within a desired range for the number of rules[9]. This avoids long mining times that might produce large collections of rules with low predictive power.

I. INTRODUCTION

Algorithm is perfect method to mine interested rule set, is based on association rule mining with some variation of classification association mining algorithm. The algorithmic approach is to mine class association rules. The most significant concern is how the interestingness of an association rule is measured. In the case of classification, we are interested in a highly accurate rule set. The rule set should be able to generalize beyond the training instances. For this purpose, we developed a method which can be decomposed in three parts. Find frequent item sets and frequent class association rules. The provided support threshold value is used to remove the uninterested element.

Figure 1 Class Association Rule Mining

Find the strong class association and confidence threshold value helps to accomplish this task and prune the weak rules .Subset of selected class association rule is used to design a classifier and rest of the class association rules are removed[8].

In short, our interestingness measure should prefer more accurate rules. The Apriori algorithm [1] has become the standard approach to mine association rules. We have adopted it to mine class association rules with some modification in the way explained by Liu et al. [2]. It generates frequent items. An item set is called frequent when its support is above a predefined minimum support.
Figure 1 Class Association Rule Mining

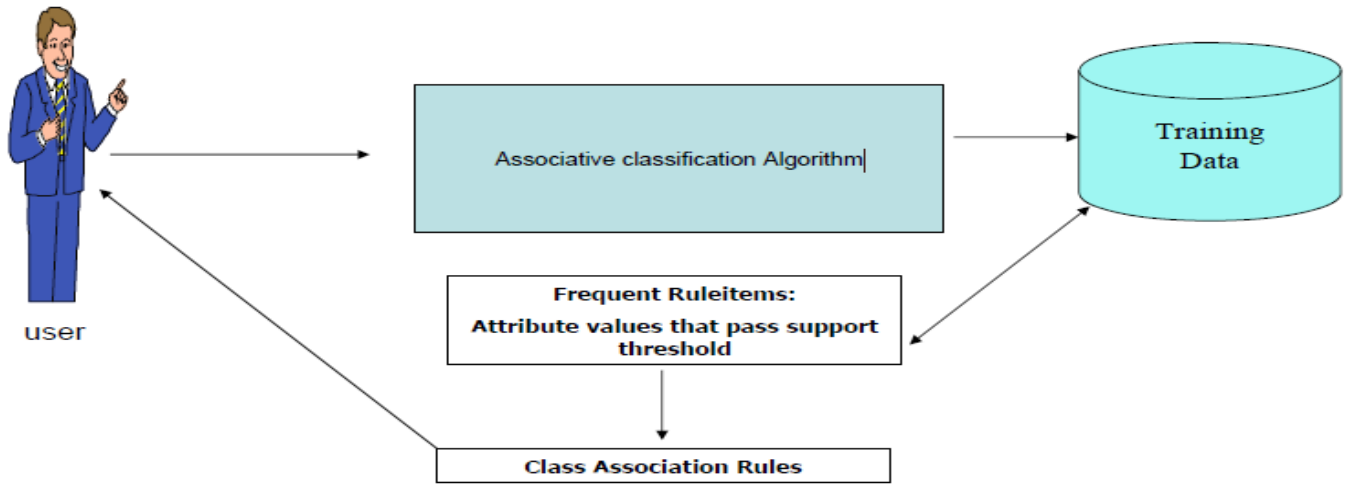

Figure 1 Class Association Rule Mining

II. MATERIALS AND METHODS

A. Tracing of proposed Algorithm using Example:

Consider example of figure-1, for proposed algorithm. Strong class association rule set shown in Table -2 which is getting by pruning the weak class association rules which satisfies confidence threshold are further classifies based on subset construction[3].

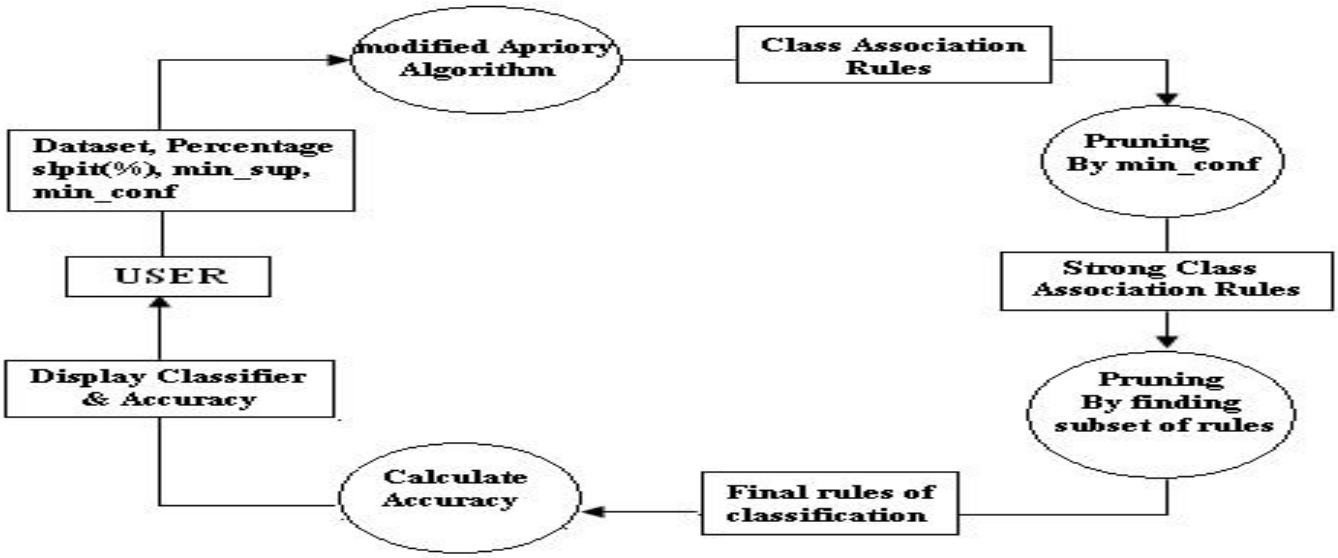

Figure-2 Data Flow Diagram

B. Algorithm of Proposed System:

Algorithm of Classification using association is divided into four phases[7].

Input: Database, minimum support, minimum confidence, percentage Split(%)

Output: Classification using association

Process

Phase - 1

/*Finding frequent item set and generate class association rules based on support measurement*/

Define class attribute C.

Divide dataset in training and testing by given percentage Split(%).

Find frequent item set in form X → C and Generate Class association rules where the item C is a distinct class label from training dataset.

Calculate support value of each rule.

Discard each rule with support < minimum support.

Phase - 2

/*Finding strong class association rules by pruning week rules based on confidence measurement*/

Calculate confidence value of each rule.

Find the weak rules which having confidence < minimum confidence.

Prune the weak rules and Generate strong class association rules.

Phase - 3

/* Find a subset of selected class association rules is used to design a classifier and rest of the class association rules are removed */

Find the rule R1 which is subset of R2 from selected class association rules.

Discard the rules R2 if confidence (R1) >= confidence (R2).

Generate classifier which is based on association rules.

Phase - 4

/* Calculate Accuracy of model.*/

Test each rule generated in classifier on testing dataset.

Find the accuracy of each rule.

Calculate accuracy of classifier model by Calculating average of all rule's accuracy[3].

C. Flow Of Proposed Algorithm:

a. Main Steps of proposed algorithm:

a) Set minimum support, minimum confidence, percentage split (PS in%).
b) Select Database.
c) Divide Dataset into Training and Test dataset according to PS.
d) Finding frequent item set and generate class association rules[10].
e) Discard those items having the support value below minimum support.
f) Calculate the confidence and perform Pruning by discarding all rules having less confidence then minimum confidence.
g) Find the rule X which is subset of selected class association rules of Y.
h) Discard rules Y if confidence(X) >= confidence(Y).
i) Display associative classifier's rules.
j) Calculate Accuracy of the model by testing the rules on test dataset.
k) Display Model Accuracy

III. RESULTS AND DISCUSSION

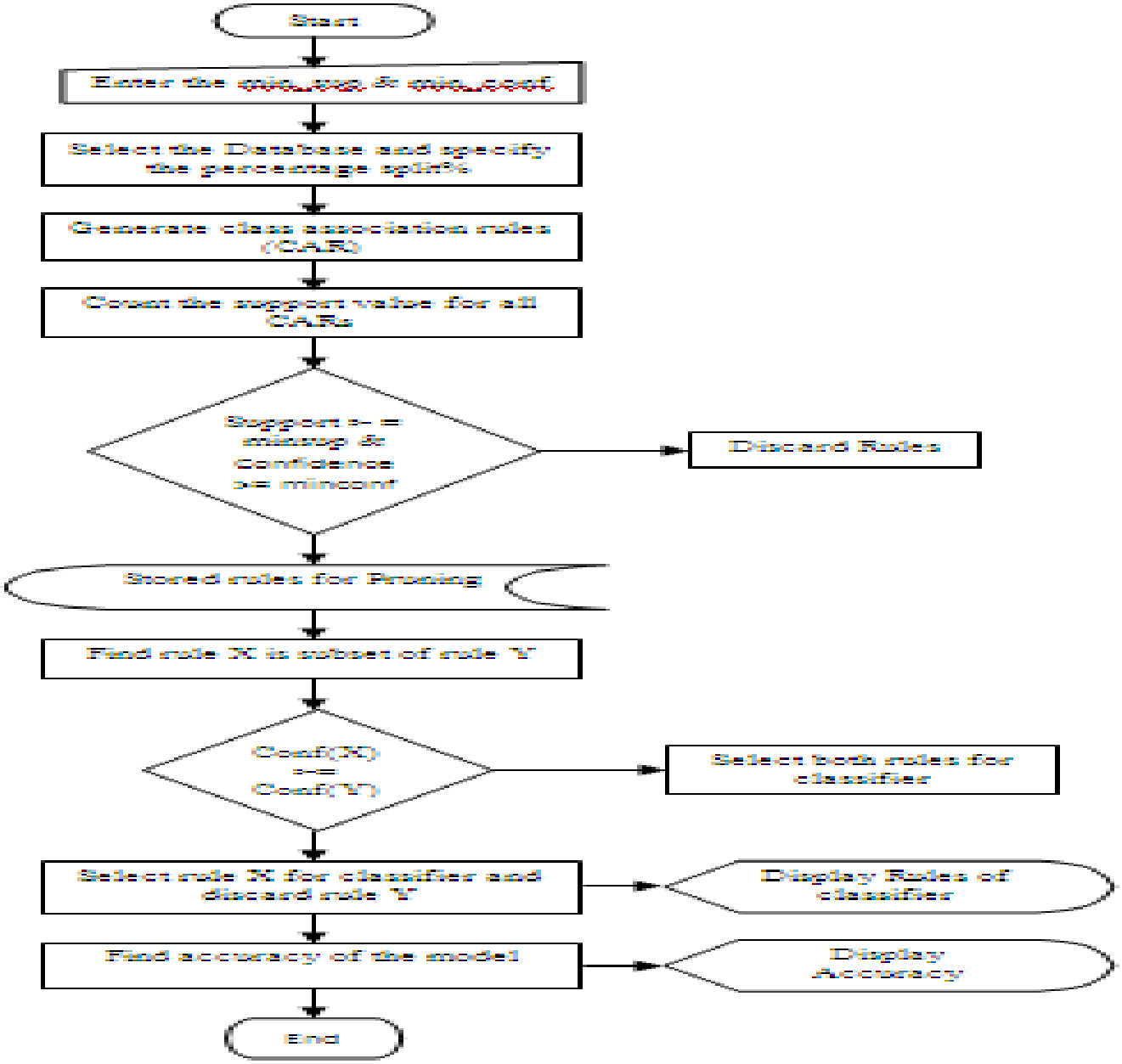

Figure: 3

A. Generate Class Association Rules (Modified Apriori algorithm):

In order to mine class association rules the basic algorithm has to be changed. Conceptually, class association rules differ from standard association rules in their consequence. They have the form X → Y, where the item Y is a distinct class label and {Y} $\notin$ X. We use the adapted version of apriori that is employed in the well-known CBA method [2] for classification based on association rules. CBA's modified apriori uses a virtual

B. Flow chart of proposed system:

Division of the training set into subsets each of which contains only instances labeled with the same class. Frequent item sets are found in each of these subsets separately. Once the frequent item sets are identified rule generation is simple. The rule body is the frequent item set itself and the rule head is the associated class label. The virtual division allows us to calculate the support s(X∩ Y) of the whole rule X →Y . However we cannot calculate the confidence of the rule. Confidence is defined as ˆc(X →Y) = s(X ∩Y)/ s(X). This problem is the rationale for a virtual division instead of an actual division[4]. Now the extended (or in a sense restricted) apriori can mine class association rules.

Input: Transaction dataset, minimum support
Output: Class association rule

C. Pruning:

As the algorithmic approach for classification using association rules consists of three major steps: mining, pruning, and classifying. The pruning step links the mining of descriptive patterns—association rules—with the building of a predictive model, by choosing the patterns which are used for the global.

Class Association rule mining normally results in a large set of association rules which does not only affect the computation time during classification, but also makes it more difficult for a human expert to understand and analyze the results. The overall goal should be to trim the mined set of class association rules to one which is as powerful and as small as possible. Therefore, pruning is an essential step in classification using association rules and a crucial difference between existing approaches. However, there are different approaches to build a classifier out of a set of rules, but the pruning methods decides which rules belong to the final classifier. The intended purpose of pruning is to reduce the size of the mined rule set without losing its discriminative power. We evaluate this in our experiments.

Input: Class association rule set sorted according to interestingness measure, minimum Confidence[5].
Output: Strong class association rules

D. Classification:

This section introduces the final step in classification using association rules: find minimal set of rules as the classification. The input to this step is a pre-processed set of strong class association rules. In a more abstract view of the entire process they correspond to a set of descriptive patterns. For a set of classification rules there are three fundamental ideas of how to use them for classification [11]. The basic decision involves whether to consider every rule that covers an instance to a certain extent or to consider only a single rule. Instead of weighting each rule equally we can consider more elaborate weighting schemes. The use of weighting schemes is easily accommodated, because association rule mining algorithms produce a sorted set of rules according to their interestingness measure. The simple majority vote algorithm does not take any information out of the sort order and therefore cannot profit from it. Hence, in order to reveal the differences between association rule mining algorithms, weighted schemes are preferable. We refer to this kind of algorithms as weighted vote algorithms. Apart from these voting algorithms which consider each rule (or subset of rules when the weight of some rules is set to 0), the other possibility is to look only at a single rule. Therefore the sorted set of class association rules is used as a sorted list and the first rule that covers the instance to be classified is used for prediction. These approaches are called decision list algorithm [6].

Input: Strong class association rule set
Output: classifier model

Table .1 Example of proposed algorithm before subset construction.

Strong Class Association Rule		**Confidence**
Antecedent	**Consequent**	
X1	C2	4/5
X2	C1	3/3
Z1	C1	4/5
X2Z1	C1	3/3

As in subset Construction we find the rule which is subset of other one. In the above table the rule X2 -> C1 is subset of X2,Z1 -> C1 with equal class attribute and confidence value.

Table.2 Example of Proposed classifier algorithm

Final Rules for Classifier Model		**Confidence**
Antecedent	**Consequent**	
X1	C2	4/5
X2	C1	3/3
Z1	C1	4/5

IV. REFERENCE

[1] B. Liu, W. Hsu, and Y. Ma, “Integrating Classification and Association Rule Mining”, Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD-98), AAAI Press, New York City, NY, United States, 1998, pp. 80-86.

[2] Syed Zahid Hassan and Brijesh Verma, “A Hybrid Data Mining Approach for Knowledge Extraction and Classification in Medical Databases”, Seventh International Conference on Intelligent Systems Design and Applications, 2007, pp 503-508.

[3] Quanzhong Liu, Yang Zhang and Zhengguo Hu, “Extracting Positive and Negative Association Classification Rules from RBF Kernel1”, International Conference on Convergence Information Technology, 2007, pp1285-1291.

[4] Kamber, Jiawei Han and Micheline, "Data Mining: Concepts and Techniques, 2nd ed," 2006.

[5] Jianchao Han, “Learning Fuzzy Association Rules and Associative Classification Rules”, IEEE International Conference on Fuzzy Systems, 2006, pp 1454-1459.

[6] M.H.Margahny and A.A.Mitwaly, “Fast algorithm for mining association rule”, AIML 05 Conference, 19-21 December 2005, pp 36-40.

[7] F. Coenen and P. Leng, “An Evaluation of Approaches to Classification Rule Selection”, Proceedings of the 4th IEEE International Conference on Data Mining (ICDM-04), IEEE Computer Society, Brighton, United Kingdom, November 2004, pp. 359-362.

[8] SU-LAN ZHANG and JI-FU ZHANG, “A NEW CLASSIFICATION MINING MODEL BASED ON THE DATA WAREHOUSE”, Proceedings of the Second International Conference on Machine Learning and Cybernetic, 2003, pp 168-171.

[9] X Yin and J. Han, “CPAR: Classification based on Predictive Association Rules”, Proceedings of the Third SIAM International Conference on Data Mining (SDM-03), SIAM, San Francisco, CA, United States, 2003, pp. 331-335.

[10] W. Li, J. Han, and J. Pei, “CMAR: Accurate and Efficient Classification based on Multiple Class association Rules”, In Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM-01), IEEE Computer Society, San Jose, CA, United States, 2001, pp. 369-376.

[11] Jochen Hipp, Ulrich Guntzer and Gholamreza Nakhaeizadeh, “Algorithm of association rule mining-a general survey and comparison”, ACM SIGKDD, Vol 2, Issue 1, July 2000, pp 58-64.