Feature Selection by Mining Optimized Association Rules based on Apriori Algorithm

K. Rajeswari, PhD
Dept. of Computer Engineering
Pimpri Chinchwad college Of Engineering
Pune, India

ABSTRACT
This paper presents a novel feature selection based on association rule mining using reduced dataset. The key idea of the proposed work is to find closely related features using association rule mining method. Apriori algorithm is used to find closely related attributes using support and confidence measures. From closely related attributes a number of association rules are mined. Among these rules, only few related with the desirable class label are needed for classification. We have implemented a novel technique to reduce the number of rules generated using reduced data set thereby improving the performance of Association Rule Mining (ARM) algorithm. Experimental results of proposed algorithm on datasets from standard university of California, Irvine (UCI) demonstrate that our algorithm is able to classify accurately with minimal attribute set when compared with other feature selection algorithms**.**

Keywords
Feature selection, Association Rule Mining (ARM), Apriori, Classification**..**

1. INTRODUCTION
Many times, data sets for analysis contain hundreds of attributes, which may be irrelevant to the mining task or redundant. Attribute subset selection or Feature selection is a technique to extract closely related features and remove irrelevant or useless features according to an objective function. The aim of Feature selection is to minimize the number of features such that the probability distribution of the resulting data classes is near to the original distribution of all the features [1]. An exhaustive search for the optimal subset of attributes can be prohibitively expensive, especially as total number of records (n) and the number of data classes increase. Association rule mining, one of the most important and well researched techniques of data mining, was initially introduced in [2]. This technique is utilized in our work with reduced data set related to the desired class label and with reduced features. There are many reasons for subset selection of the features instead of all the features [3]. To measure a diminished set of features is cheaper, faster with increased accuracy by exclusion of irrelevant features. Differentiating relevant and irrelevant features, gives a proper insight about the nature of prediction problem and understanding of final classification model.

For feature selection various heuristic methods used are stepwise forward selection, stepwise backward elimination, combined forward selection and backward elimination, random generation and decision tree induction. Stepwise forward selection is a feature selection method which starts with an empty set of attributes, best of the original attributes is found and added entirely after a single consideration of its usefulness. The pitfalls of this method include a high susceptibility to getting trapped by local optima, and a one track process that easily discards a feature entirely after a single consideration of its usefulness. Variations of this method is found in [4][5][6]. Stepwise backward elimination procedure starts with full set of attributes and at each step, a worst attribute is removed. INTERACT is a backward elimination algorithm [7]. Recent references about implicit enumerative techniques of selection features adapted to regression models are found in [8 - 10]. Also the problems of feature selection is dealt in.[2][11-17]. Principal component analysis (PCA) is one of the famous method used [11]. But it is disadvantageous, as all the data need to be processed when new data is added. Decision trees are used [16-18] which uncovers relevant attributes one by one iteratively. Mutual Information is used as a feature selector in [19]. Stepwise regression [20] uses a statistical F-Test technique and best first search uses greedy hill climbing [21] for Feature selection. Taguchi method is used to find the Neural Network structure for feature section [22]. Measures like Information measures, distance measures, dependence measures, accuracy measures, consistency measures are used for evaluating the goodness of features [23-27]. Wrapper methods with Genetic algorithms are used for feature selection [28][29]. The drawback of Genetic algorithms is over fitting. Although classification algorithms like decision tree, neural networks, bayes classifier classify the given data set, it is found [30] that Feature selection or attribute selection play a major role in improving the efficiency of the classifier.

Our work minimizes an exhaustive search as the data set itself is reduced, as per the class labels desired during classification. The organization of paper is as follows – Chapter 2 discusses about Association rule mining process, Chapter 3 is about data set reduction. Chapter 4 gives an overview about Feature selection, Chapter 5 gives details of implementation and in chapter 6 result analyses is discussed.

2. ASSOCIATION RULE MINING
Let I = {i1, i2, i3,..., id} be the set of all items in a market basket data and T = {t1,t2,t3,...,tn} be the set of all transactions. Each transaction t_i contains a subset of items chosen from Item set I. A collection of zero or more items is termed an item set. Support count is an important property of an item set. Support count refers to the number of transactions that contain a particular item set. Mathematically, the support count, $\sigma(X)$, for an item set X can be given as follows:

$$\sigma(X)= |\{t_i|X \subseteq t_i, t_i \in T\}|$$

An association rule is an implication expression of the form A → B , where A and B are disjoint item sets, i.e., A ∩ B = ∅. There are two important basic measures for association rules, minimum support and confidence. Generally minimum support and confidence are predefined by user/analyst so that the rules which are not so interesting or not useful can be deleted. Support is the total count of number of transactions where all items in A and B are together. Confidence determines how frequently items in B appear in transactions that contain A. The formal definitions of these metrics are given below,

Support(A → B)= σ(A and B)
Confidence(A → B)= σ(A and B)/ σ(A)

Apriori algorithm is used to find the frequent item sets [31].

3. DATA SET REDUCTION FOR FEATURE SELECTION

Feature selection is an important preprocessing technique to improve performance of association rule mining process. It improves the accuracy of the classifier. For any given dataset, the features can be analyzed, to find their association with the class label using Apriori algorithm. Rules are generated for item sets with expected minimum support and confidence. If the customer is interested on a particular class label, then the tuples with this particular class label is taken for analyzing the association level existing between attributes and desirable class label. This method of selecting tuples based on desirable class label increases the efficiency of Apriori algorithm by reducing the number of iterations and time involved in finding association rules .The subset features found after reducing the tuples is fed to the classifier to check the performance of classifier.

For the heart disease data set, the class label will be 'Have risk' and 'No risk' cadres. As user will be more interested in the class label 'Have risk' only , the 'No risk' cadre tuples can be removed from the data set to find the association rules for consequences 'Have Risk'. This reduces the data set size at least 40%, thereby improving the performance of Apriori algorithm. We have obtained the similar accuracy, sensitivity and specificity values with a reduced data set as that of original data set. Accuracy is tested by using C 4.5 decision tree classifier in Weka [32].

4. FEATURE SELECTION

In this paper, a novel feature selection method is proposed based on Association analysis. It extracts the features by analyzing the correlation between features found by association rule mining. Based on the consequence, the data set is first reduced. This reduced data set with the desired consequence is used for association rule mining. This reduces the memory utilization, and time taken for each iteration of the association rule mining process. After selecting the features, complete set of data is given to classifier to test the accuracy using 10 fold cross validation. We found that, the results obtained are similar to that of the original data set.

4.1 Implementation details

Our Feature selection algorithm has four steps- dataset reduction, frequent item set generation using Apriori, association rule generation and feature selection. Algorithm Database_Reduction is used to reduce original dataset. Steps 4-12 are repeated for each tuple in original dataset, if tuple is not contributing to class label then it is deleted from dataset and the size of dataset is updated.

Apriori algorithm is used for mining frequent item sets for Boolean association rules. It uses prior knowledge of frequent item sets and explores k+1 item sets from k item sets. to generate all k- frequent item sets . It follows antimonotonic property ie if a set does not pass test, all of its supersets also will fail in the test. If P(I) < min_sup, then P(I∪A) < min_sup. A two step process is followed namely Join Step and Prune step. From these steps, frequent item sets are found and association rules are generated.

Let Ck denote the set of candidate k-item sets and Fk denote the set of frequent k-itemsets. The frequent itemsets generation algorithm has two important characteristics:

(1) It is a level-wise algorithm; i.e., mapped to the lattice structure, it traverses the item set lattice one level at a time, from frequent 1-itemsets to the maximum size of frequent item sets;

(2) It uses a generate-and-test strategy for finding frequent item sets. At each iteration, new candidate item sets are generated from the frequent item sets found in the previous iteration. The support for each candidate is then counted and tested against the minimum support threshold [16]. The second step is to construct association rules that satisfy the user-defined minimum confidence by using frequent itemsets. Suppose one of the frequent itemsets is Fk,

$$Fk = \{i1, i2, i3, \ldots, ik\},$$

association rules with this itemsets can be generated in the following way: the first rule is

$$\{i1, i2, \ldots, ik-1\}$$

by checking the confidence this rule can be determined as interesting or not.

Algorithm Feature selection based on association rule mining returns all closely related features. Dataset D is discretized in step 2. Output of step 2 is given to step 3 that will reduced the size of descretized dataset to improve the efficiency of association rule mining thereby improving efficiency of this feature selection algorithm. All possible rules of interested consequences are generated in step 4. Some rules among all generated rules may not be useful and are deleted based on lift value (lift <=1). If lift value is equal to one, it means the antecedent and consequent of the rule r are not related. If lift values is less than one, it means the antecedent and consequent of the rule r are related negatively and if it is greater than one then positively related. Feature selection based on association rule mining In step 12-13 antecedent attributes of selected rules are included in result set and are returned as a final selected feature set in step 16.

Algorithm 1: Database_Reduction

D : Training set

N: Number of tuples in D

C : Class Attribute

Method:

1. Set D'=D
2. Set N'=N
3. Set i=1;
4. Repeat
5. For each T_i Є D do
6. Set flag=false
7. For each c Є C do
8. flag=flag v T_i[c]
9. If(flag = false)
10. Set D'=D'- T_i
11. Set N'=N'-1;
12. Until i < > N

Algorithm 2: Apriori algorithm for frequent itemsets generation

minsup: Minimum support threshold

N :Number of tuples in original data set D

Method:

1. k=1
2. F_k={i|i∈I∧ Support({i})≥N×minsup}
3. Repeat
4. k=k+1
5. C_k=candidates generated from F_k-1
6. For each instance t ∈ T do
7. C_t=subset(C_k, t)
8. For each candidate itemset c ∈ C_t do
9. Support(c)= Support(c)+1
10. End for
11. End for
12. F_k={c|c∈C_k∧Support(c)≥N×minsup}
13. Until F_k=Null
14. Result=∪F_k

Algorithm 3: Rules generation from frequent itemset generated by Apriori algorithm

minconf: Minimum confidence threshold

Method:

1. For each frequent k-itemset f_k, k≥2 do
2. H_1={i|i∈f_k}
3. call apr-genrules(f_k, H_1)
4. End for

Function apr-genrules(f_k, H_m)

1. k=|f_k|
2. m=|H_m|
3. If k > m+1 then
4. H_{m+1}=m+1
5. For each h_{m+1} ∈ H_{m+1} do
6. Conf=Support(f_k)/ Support(f_k-h_{m+1})
7. If Conf ≥ minconf then
8. Return the rule (f_k-h_{m+1})→h_{m+1}
9. Else
10. delete h_{m+1} from H_{m+1}
11. End if
12. End for
13. apr-genrules(f_k, H_{m+1})
14. End if

Algorithm 4: Feature selection based on association rule mining

D': Reduced Training set

N': Number of tuples in D'

N: Number of tuples in D

C: Class Attribute

Method:

1. Result=Null
2. Discretize(D)
3. Database_Reduction(D,N)
4. Rules=apriori(D',N',N,minsup,minconf)
5. For rach rule r Є Rules do
6. If lift(r) < 1
7. Delete r from Rules
8. End if
9. End for
10. If Rules = Null then break
11. Else
12. r = getRule(Rules)
13. F= Select_Antecedent_Attributes(r)
14. End if
15. Return Result

5. EXPERIMENTAL RESULTS

5 .1 Efficiency

Performance of proposed algorithm is mainly depending on the run time of Apriori algorithm.

The computational complexity of the Apriori algorithm can be affected by support threshold, number of items, number of transactions and average transaction width [34]. The computational complexity of the Apriori algorithm on a dataset having N tuples has following parameters:

5.1.1 Time required for generation of frequent 1-itemsets - O(NI) where I is the average no. of Item set and N is the total number of transactions.

5.1.2 Time required for candidate generation C- It includes merging cost and pruning cost.

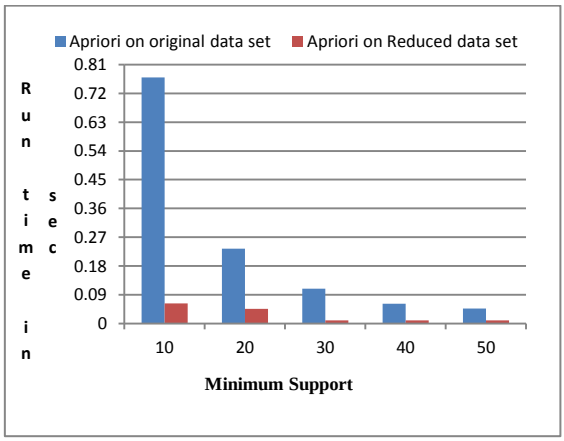

Figure 1. Execution time of Apriori Algorithm with different minimum support threshold.

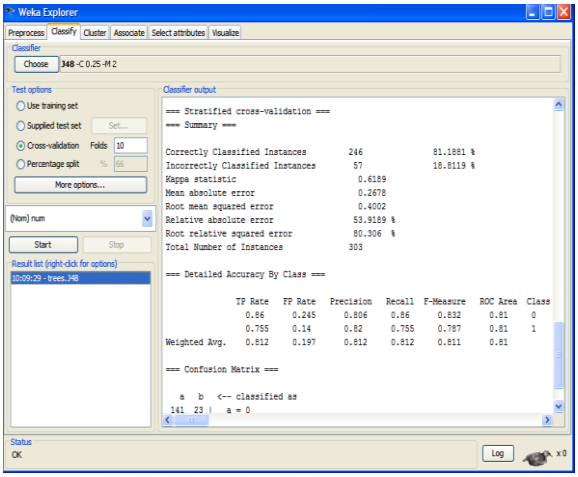

Figure 2. Result of Classifier C 4.5[28] on Heart dataset [29] with all attributes

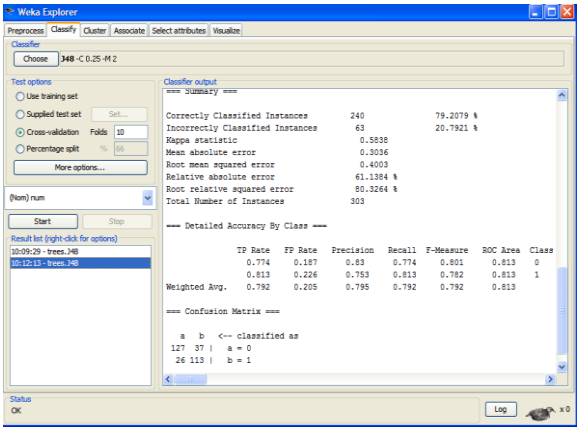

Figure 3. Result of Classifier C 4.5[28] on Heart dataset [29] with selected attributes using feature selection algorithm

5.1.3 Time required for support counting- The cost for support counting is $O(N\sum_k(\begin{smallmatrix} w \\ k \end{smallmatrix})\alpha_k)$ where w is the maximum transaction width and α_k is the cost for updating the support count of a candidate k-itemset.

The computational complexity for the Apriori algorithm on reduced dataset having N' tuples is very much less than the same on original dataset having N tuple (Note N' is atleast 40% less than N) and is shown in figure 1.

5.2 Effectiveness

Effectiveness of the proposed algorithm is checked.

6. CONCLUSION

In this paper, we have proposed a novel method of feature selection using association rule mining on reduced data set based on desired class label attribute. By reducing dataset the performance of Apriori algorithm is improved significantly, thereby improving association rule mining process. Our result shows that this method is effective and efficient for most of the real datasets from UCI repository with acceptable classifier accuracy.

7. REFERENCES

[1] Jaiwei Han and Micheline Kamber, "Data Mining Concepts and Techniques", *Second Edition, Elsevier, Morgan Kaufmann publishers*.

[2] R. Agrawal, T. Imielinski, and A. Swami, "Database mining: A perfor- mance perspective," *IEEE Trans. Knowledge Data Eng., vol. 5, Dec. 1993*.

[3] Reunanen, J. (2003). "Overfitting in making comparisons between variable selection methods". *Journal of Machine Learning Research, 3 (7/8), 1371—1382.*

[4] K.Z. Mao, "Fast Orthogonal Forward Selection Algorithm for Feature Subset Selection". *IEEE Transactions on Neural Networks, 2002. 13(5): 1218-1224*.

[5] J. Jelonek, Jerzy S., "Feature Subset Selection for Classification of Histological Images. *Artificial Intelligence in Medicine", 1997. 9:22-239.*

[6] B. Sahiner, H.P. Chan, N. Petrick, R.F. Wagner, and L. Hadjiiski, "Feature Selection and Classifier Performance in Computer-Aided Diagnosis: The Effect of Finite Sample Size" *Medical Physics, 2000. 27(7): 1509-1522*.

[7] Z. Zhao, H. Liu, Searching for Interacting Features, *IJCAI 2007*.

[8] Gatu C. And Kontoghiorghes E.J. (2003)."Parallel Algorithms for Computing all Possible Subset Regression Models Using the {QR} Decomposition". *Parallel Computing, 29, pp.505-521.*

[9] Gatu C. And Kontoghiorghes E.J. (2005). "Efficient Strategies for Deriving the Subset {VAR} Models". *Computational Management Science, 2 (4):253-278.*

[10] Gatu C. And Kontoghiorghes E.J. (2006)."Branch-and-bound Algorithms for Computing the Best-Subset Regression Models". *Journal of Computational and Graphical Statistics, 15 (1):139-156.*

[11] T. Joliffe, "Principal Component Analysis", *New York: Springer- Verlag, 1986*.

[12] K. L. Priddy et al., "Bayesian selection of important features for feed- forward neural networks", *Neurocomput., vol. 5, no. 2 and 3, 1993.*

[13] L. M. Belue and K. W. Bauer, "Methods of determining input features for multilayer perceptrons," *Neural Comput., vol. 7, no. 2, 1995*.

[14] J. M. Steppe, K.W. Bauer Jr., and S. K. Rogers, "Integrated feature and architecture selection*," IEEE Trans. Neural Networks, vol. 7, July 1996.*

[15] Q. Li and D. W. Tufts, "Principal feature classification," *IEEE Trans. Neural Networks, vol. 8, Jan. 1997.*

[16] R. Setiono and H. Liu, "Neural network feature selector," *IEEE Trans. Neural Networks, vol. 8, May 1997.*

[17] R. Quinlan, C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann.

[18] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees. Belmont, *CA:Wadsworth, 1984*.

[19] Hanchuan Peng, Fuhui Long, Chris Ding, Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy, *IEEE Trabsactions on Pattern Alalysis and machine Intelligence, vol. 27, No. 8, August 2005.*

[20] S Nojun Kwak and Chong-Ho Choi, Input Feature Selection by Mutual Information Based on Parzen Window, *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, No. 12, December 2002.*

[21] Thomas Drugman, Mihai Gurban and jean-Philippe Thiran,' Feature Selection and Bimodal Integration for Audio-Visual Speech Recognition', *School of Engineering-STI Signal Processing Institute*

[22] Georgia D. Tourassi, Erik D. Frederick, Mia K. Markey, Carey E., Floyd, Jr., "Application of the mutual information criterion for feature selection in computer-aided diagnosis", *North Carolina, Medical Physics, vol. 28, No. 12, December 2001*.

[23] Gang Wang, Frederick H. Lochovsky, Qiang Yang, "Feature Selection with Conditional Mutual Information MaxiMin in Text Categorization", *Department of Computer Science, Hong Kong University of Science and technology, Kowloon, Hong Kong, 2004*.

[24] J. J. Liu, G. Cutler, W. Li, Z. Pan, S. Peng, T. Hoey, L. Chen, and X. B. Ling, "Multiclass Cancer Classification and Biomarker Discovery Using GA-Based Algorithms*," Bioinformatics, vol. 21, pp.2691-2697, 2005.*

[25] L. Li, T.A.Darden, C.R.Weingberg, and Levine., "Gene Assessment and Sample Classification for Gene Expression Data Using a Genetic Algorithm / k-Nearest Neighbor Method," *Combinatorial Chemistry & High Throughput Screening, vol. 4, pp. 727-739, 2001*.

[26] Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research, vol. 3, pp. 1157-1182, 2003.*

[27] Huizhen Liu, Shangping Dai, Hong Jiang, 'Quantitative association rules mining algorithm based on matrix', 978-1-4244-4507-3/09©2009 IEEE.

[28] Weka Software http://www.cs.waikato.ac.nz /ml/ weka.

[29] Murphy P. M. and Aha. D. W. (1994). "UCI repository of Machine Learning, University of California", *Department of Information and Computer Science, http://www.ics. uci.edu/~ mlearn/ ML Repository.html*.

[30] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Networks, vol. 5, July 1994*.

[31] N. R. Draper and H. Smith, Applied Regression Analysis, *2nd ed. New York: Wiley, 1981*.

[32] P. H.Winston, Artificial Intelligence, *MA: Addison-Wesley, 1992.*

[33] G. E. P. Peterson et al., "Using Taguchi's method of experimental design to control errors in layered perceptrons*," IEEE Trans. Neural Networks, vol. 6, July 1995.*

[34] Pang-Ning Tan, Michael Steinbach, Vipin Kumar "Introduction to Data Mining", *Addison Wesley.*

Table 1. Features selected and Accuracy of different methods on various UCI datasets [33]

Datasets / Method			Association Mining		Genetic Search		Chi-square		Information Gain	
			No. of Attributes selected	Accuracy	No. of Attributes selected	Accuracy	No. of Attributes selected	Accuracy	No. of Attributes selected	Accuracy
Name	No. of Attributes	Accuracy								
Vote	16	96.32	4	95.843	7	96.32	16	96.32	16	96.32
Zoo	17	92.07	12	90.67	13	92.07	17	92.07	17	92.07
Breast Cancer	9	94.84	8	92.8	9	94.84	9	94.84	9	94.84