Research Journal of Applied Sciences, Engineering and Technology 6(24): 4681-4684, 2013
DOI:10.19026/rjaset.6.3485
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: April 09, 2013 Accepted: May 03, 2013 Published: December 25, 2013

Research Article

Improving Efficiency of Classification using PCA and Apriori based Attribute Selection Technique

[1]K. Rajeswari, [2]Rohit Garud and [3]V. Vaithiyanathan
[1]Pimpri Chinchwad College of Engineering, Pune and Ph. D Research Scholar, SASTRA University, Tanjore, India
[2]Pimpri Chinchwad College of Engineering, Pune, India
[3]Professor and Associate Dean Research, SASTRA University, Tanjore, India

Abstract: The aim of this study is to select significant features that contribute for accuracy in classification. Data mining is a field where we find lots of data which can be useful or useless in any form available in Data Warehouse. Implementing classification on these huge, uneven, useless data sets with large number of features is just a waste of time degrading the efficiency of classification algorithms and hence the results are not much accurate. Hence we propose a system in which we first use PCA (Principal Component Analysis) for selection of the attributes on which we perform Classification using Bayes theorem, Multi-Layer Perceptron, Decision tree J48 which indeed has given us better result than that of performing Classification on the huge complete data sets with all the attributes. Also association rule mining using traditional Apriori algorithm is experimented to find out sub set of features related to class label. The experiments are conducted using WEKA 3.6.0 Tool.

Keywords: Apriori, bayes, classification, data mining, decision tree classifier j48, features, mult layer perceptron, WEKA 3.6.0

INTRODUCTION

Data mining is a field were huge amount of data which is been mined form data warehouse. Classification is a technique which is used to label the attributes of the table and classify the data into different similar type of classes or category. Classification is used to predict the type of class techniques available in data mining, classification which it belongs. Classification is divided into two categories supervised and unsupervised, Supervised classification is the technique in which label is already known before Classification and in Unsupervised we need to find it based on the training sets and apply it on test data. To apply classification on this huge data set will take large amount of time to compute as well we are not sure about the accuracy of the results. This study proposes a method where classification technique is used only with the important attributes using feature selection techniques namely PCA (Principal Component Analysis) and Association rule mining technique which will select the subset attributes significant for classification.

Association rule mining: Association rule mining is business intelligence technique which defines how the attributes of the relations are closely related to each other and also how all subsets of the attribute are dependent on the class label. Apriori is the algorithm which gives us set of rule based on support and confidence of the attributes in the subsets.

Feature selection: Feature selection is technique of selecting a attribute form a relation which is more important to describe the relation and to make a decision and to be decision attribute. PCA is one of the feature selection techniques which select the attributes which will give more prominent result and increases the accuracy of the classification algorithm.

Classification techniques:

- **J48:** J48 as shown in Fig. 1 is the classification algorithm based on decision tree. It creates a tree of attributes which depicts the arrangement of attribute in the tree structure based on the highest value of the Information Gain and Entropy.
- **Multi-layer-perceptron:** It is the classification algorithm based on neural network which takes a lot of time to execute but the result accuracy is efficient.
- **Bayes:** It is the classification algorithm based on Bayesian technique which is based on conditional probability and bayes theorem.

Corresponding Author: K. Rajeswari, Pimpri Chinchwad College of Engineering, Pune and Ph. D Research Scholar, SASTRA University, Tanjore, India

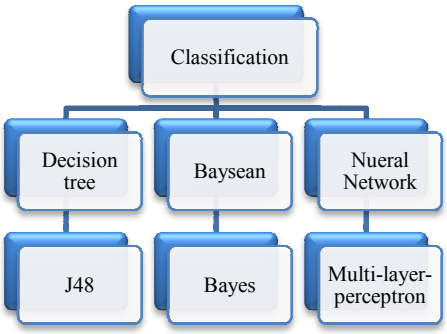

Fig. 1: Classification category

In this study we propose classification with selected subset of significant features which will provide good accuracy.

LITERATURE SURVEY

In our proposal we chose J48 as a decision tree algorithm, Bayes as Bayesian type and Multi-layer-Perceptron as Neural Network based classification algorithm because they are the best in their fields of classification techniques. The study (Sprinkhuizen-Kuyper, 2008) depicts how decision tree is useful in classification techniques with improved efficiency and pruning as well as less computations. The study (Phyu, 2009) concludes the comparisons of the classification algorithms based on accurate system results and also depicts how decision tree and bayes classification technique is well suited for good accuracy. The study (Nguyen and Grenville, 2008) states that J48 is better classification algorithm than Bayes classification algorithm. The study (Atlas *et al*., 1989) signifies the importance of Multi-layer-Perceptron algorithm in classification technique. Machine learning approaches have focused on models (e.g., neural nets, Bayesian nets, hyper planes) that are unfamiliar to most non-analyst users. Although data mining models in the form of if-then rules (Usama *et al*., 1996; Holte, 1993; Thames *et al*., 2003), decision trees (Quinlan, 1993) and association rules (Agrawal *et al*., 1996; Han and Yongjian, 1995) are considered to be easy to understand, problems arise when the size of trees or the number of rules become very large. Feature selection using hashing and application of Apriori with selected features is discussed in Rajeswari and Vaithiyanathan (2012a). Neural networks as feature selector is a novel method proposed in Rajeswari and Vaithiyanathan (2012b). But the time taken for training to model the independent variables to dependent variables is large (Rajeswari and Vaithiyanathan, 2012c).

PROPOSED METHODOLOGY

In our Methodology we insist to used a apply PCA or Apriori before classification can get applied onto the huge data set which will increase the efficiency of the classification algorithm and then we compare results for

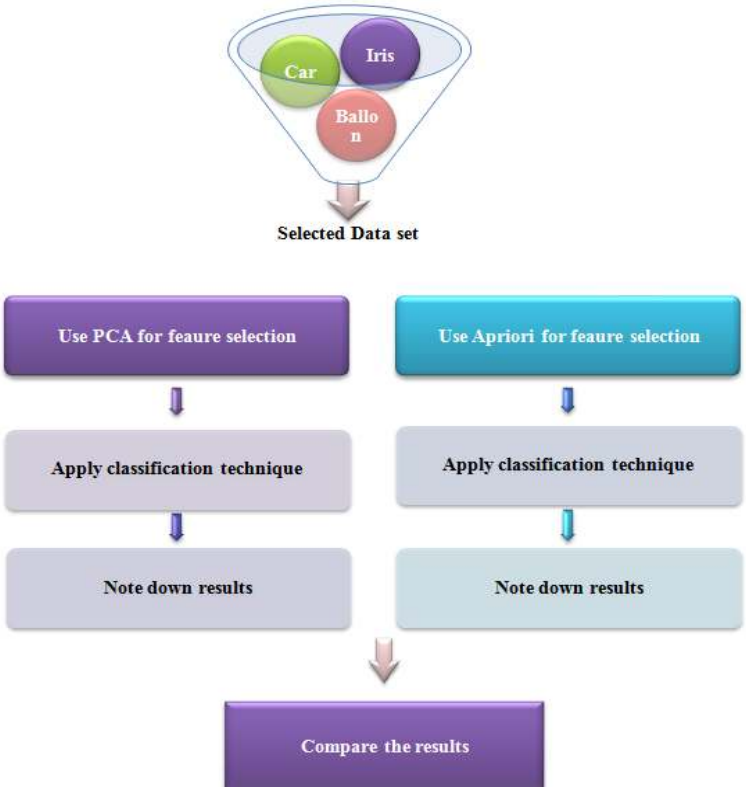

Fig. 2: Proposed methodology

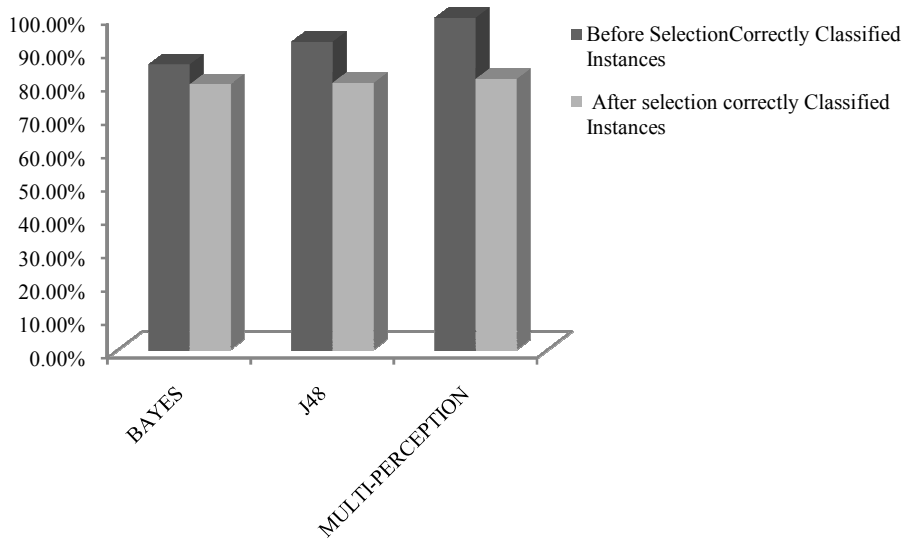

Fig. 3: Correctly classified instances for car data set

Table 1: Comparison of accuracy before and after selection of attributes with different classifiers

Classification technique	Attribute selection	Data sets	Before selection correctly classified instances (%)	After selection correctly classified instances (%)	Time build before selection	Time build after selection
Bayes	Apriori	Balloon	100	100	0 Sec	0
J48	Apriori	Balloon	100	100	0 Sec	0.16
Multi-perception	Apriori	Balloon	100	100	0.02 Sec	0.02
Bayes	PCA	Car	85.53	80	0 Sec	0 Sec
J48	PCA	Car	92.36	80	0.02 Sec	0 Sec
Multi-perception	PCA	Car	99.53	81	9.84 Sec	3.22 Sec

both PCA and Apriori for classification. To prove our goal we use WEKA 3.6.0 which includes all the algorithms for association, classification and feature selection etc concepts of data mining.

Algorithm:

- Select the data set from UCI machine learning repository
- Save the data set with the extension .csv or .arff
- Open WEKA 3.6.0 and in explore open data set file which was saved as .csv or .arff
- Go to Classify Select all classification algorithms and note down results and note how correctly classified instances, accuracy and time take to execute it.
- Then after noting the results apply PCA Principal Component Analysis on the current data set and select the attributes which came as result of PCA
- Then apply classification algorithm on selected attributes and note how correctly classified instances, accuracy and time take to execute it.
- Then after noting the results apply Apriori on the current data set and select the attributes which came as result of Apriori.ie most frequently used attribute and most associated to the class label.
- Then apply classification algorithm on selected attributes and note how correctly classified instances, accuracy and time take to execute it
- Compare noted result of PCA and Apriori (Fig. 2)

RESULTS AND DISCUSSION

The Table 1 shows the summary of accuracy obtained in percentage before and after selecting certain features using PCA and Apriori algorithm. Different classification algorithms like J48, Bayes and Multi Layer perceptron are used for obtaining the correctly classified instances using 10 fold cross validation.

Figure 3 gives the graph of accuracy of correctly classified instances for car data set with all attributes and selected attributes using PCA. It is understood for the data sets taken, namely car and balloon, apriori gives 100% accuracy with selected closely associated features with the class label.

CONCLUSION

Hence we need to first apply PCA or Apriori based association rule mining technique as a feature selection technique to select the significant subset attributes. Then apply any classification technique to classify the

test data set. We get more accurate results in less computation time. Apriori selected features give 100% accuracy whereas PCA selected features give a compatible accuracy with more time taken for building the model.

REFERENCES

Agrawal, R., M. Heikki, S. Ramakrishnan, T. Hannu and A. Inkeri Verkamo, 1996. Fast discovery of association rules. Adv. Knowl. Discov. Data Mining, 12: 307-328.

Atlas, L., J. Connor, D. Park, M. El-Sharkawi, R. Marks *et al*., 1989. A performance comparison of trained multilayer perceptrons and trained classification trees. Proceeding of the IEEE International Conference on Systems, Man and Cybernetics, pp: 915-920.

Han, J. and F. Yongjian, 1995. Discovery of multiple-level association rules from large databases. Proceeding of the International Conference on Very Large Data Bases. Institute of Electrical and Electronics Engineers, pp: 420-431.

Holte, R.C., 1993. Very simple classification rules perform well on most commonly used datasets. Mach. Learn., 11(1): 63-90.

Nguyen, T.T.T. and A. Grenville, 2008. A survey of techniques for internet traffic classification using machine learning. IEEE Commun. Surv. Tutor., 10(4): 56-76.

Phyu, T.N., 2009. Survey of classification techniques in data mining. Proceedings of the International MultiConference of Engineers and Computer Scientists. Hong Kong, March 18-20, Vol. 1.

Quinlan, J.R., 1993. C4.5: Programs For Machine Learning. Morgan Kaufmann, Los Altos, CA.

Rajeswari, K. and V. Vaithiyanathan, 2012a. Mining association rules using hash table. Int. J. Comput. Appl., 57(8): 7-11.

Rajeswari, K. and V. Vaithiyanathan, 2012b. Attribute selection using artificial neural networks-A case study of ischemic heart disease. J. Theoret. Appl. Inform. Technol., 46: 510-515.

Rajeswari, K. and V. Vaithiyanathan, 2012c. Improved apriori algorithm based on selection criterion. Proceeding of IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp: 1-4.

Sprinkhuizen-Kuyper, I.G., 2008. Data Mining Algorithms for Classification. B.Sc. Thesis, Artificial Intelligence, Radboud University Nijmegen.

Thames, H., D. Kuban, L. Levy, E.M. Horwitz, P. Kupelian *et al*., 2003. Comparison of alternative biochemical failure definitions based on clinical outcome in 4839 prostate cancer patients treated by external beam radiotherapy between 1986 and 1995. Int. J. Radiation Oncol. Biol. Phys., 57(4): 929-943.

Usama, F., G. Piatetsky-Shapiro and P. Smyth, 1996. From data mining to knowledge discovery in databases. AI Mag., 17(3): 37.