Attribute Selection to Improve Accuracy of Classification

Shailaja V. Pede
Assistant Professor
Pimpri Chinchwad College of Engineering, Pune

Swati Chandurkar
Assistant Professor
Pimpri Chinchwad College of Engineering, Pune

Suyoga Bansode
Assistant Professor
College of Engineering, Osmanabad

ABSTRACT
Nowadays the use of computer technology in the field of medical diagnosis and prediction of disease has increased. In these fields the computers are used with intelligence such as fuzzy logic, artificial neural network and genetic algorithms. Many techniques of data mining are useful in the field of medicine and many algorithms have been developed. The main objective of this work is to find out the important attributes which are highly important for accuracy of the classifier and reduce the dimensionality of dataset for classification of disease dataset. The other objective of this work is to classify the dataset in cost effective manner. As many tests are redundant and also are highly expensive. We have used various approaches for feature selection as using Brute force approach and correlation based approach. We have also proved that accuracy of classifiers are improved using feature selection.

General Terms
Disease Prediction, Data mining,

Keywords
Feature Selection, Disease Prediction, Correlation, Classifier, Association Rule

1. INTRODUCTION
Feature selection for classification in medical field attempts to save computational cost in terms of time and space. Dimensionality reduction is one of the major problem in data mining and feature selection is the way to reduce dimensions also computational cost in terms of storage. Classification of reduced dataset indicates less execution time in terms of time complexity. The performance of the algorithms is affected due to redundant and irrelevant data. In medical domain many tests need to be performed for prediction of disease, to avoid number of tests feature selection helps to find out important symptoms of any disease.

In this paper, feature selection is applied on three datasets of different diseases. Heart Disease, Breast cancer and Diabetes disease are considered for classification which are downloaded from standard benchmark UCI repository.

we have considered Heart Cleveland dataset contains 14 attributes including class attributes and the number of instances are 303. The available dataset is initially discretized based on assumptions studying different papers. Breast cancer is another problem observed in females. This dataset contains total 10 attributes including class attribute and number of instances are286. Diabetes diagnosis dataset contains 9 attributes including class attribute. Numbers of instances are 768 where last attributes shows sick or healthy.

2. RELATED WORK
Researchers have been investigating the use of statistical analysis and data mining techniques to help healthcare professionals in the diagnosis of heart disease. Statistical analysis has identified the risk factors associated with heart disease to be age, blood pressure, smoking habit [1], total cholesterol [2], diabetes [3], hypertension, family history of heart disease [4], obesity, and lack of physical activity [5]. Knowledge of the risk factors associated with heart disease helps health care professionals to identify patients at high risk of having heart disease. Researchers have been applying different data mining techniques such as decision tree, na£ŧve bayes, neural network, bagging, kernel density, and support vector machine over different heart disease datasets to help health care professionals in the diagnosis of heart disease [6]. In[7] showed correct classified accuracy of approximately 77 % with logistic regression. The another model R-C4.5 is applied which is based on C4.5 shows better results , rules created by R-C4.5 CLASSIT conceptual clustering system[8] achieved a 78.9% accuracy on the Cleveland database. In [9] Neural network based method obtained 89.01% .

3. METHODOLOGY
One of the primary objectives of the feature selection system to reduce the size of dataset by removing redundant attributes and improve the performance of the classifier and computational cost. Similarly disease classification is done with important attributes only further this will useful to make the decision support system and predict any disease in cost effective manner.

The system is divided into three parts.

a The collection of different dataset
b Construction of knowledge base by selecting important subset.
c Classification with reduced dataset.

The first part involves the collection of disease dataset in any format. Then preprocessing i.e. discretization is done. Then the discretized dataset is given for knowledge extraction. , in second part the knowledge base is constructed using different techniques of feature selection such as brute force approach with classifiers, feature selection using association rule and existing method namely genetic search and in third part classification of disease dataset with reduced set of attributes and accuracy is measured and performance is evaluated based on accuracy of different classifiers. The System Diagram is shown in Figure 1.

3.1 Feature selection using Brute Force Approach
Three discretized datasets such as Cleveland Heart disease, Brest Cancer and diabetes downloaded from UCI repository and then by trying all possible combinations of features i.e. to select k features from N number of features, 2k subsets of features are created by removing one attribute from actual dataset and accuracy of the classifier is measured and it is observed that accuracy of classifier is better than actual one. This process leads to find out the subset of attributes which really affect the performance of the classifier. But the major

problem with this approach is it usually takes too many try and also danger of over fitting. But using this work we conclude that removal of certain attributes from the dataset improves the accuracy of classifiers

3.1.2. Correlation based associated feature selection

Correlation shows how two itemsets are closely related to each other which can be used for generation of association rule. It shows the dependence of two itemsets or correlation of two itemsets. If Corr(X, Y) > 1 then two items are correlated otherwise considered as independent. With these terms and using association rule generation algorithm [10] the features can be selected.

Following are the steps for this approach of feature selection using correlation and association

1. Consider Raw dataset D, Initialize minsupp (Minimum Support), minconf (Minimum Confidence) with threshold values.
2. Discretize dataset D.
3. Apply association rule generation algorithm and generate set of rules R.
4. For each rule , calculate Corr(r), if Corr(r) < 1 and rule r does not contain class attribute discard r from R.
5. Sort all remaining rules based on confidence and then support in descending order.
6. For each rule r check the minsupp and minconf as per the threshold value.
7. Select the attributes from satisfied rules and store in result set.
8. Display final result set with reduced set of attributes.

Above algorithm works in three different steps.

1 Generate the association rules using apriori algorithm.

2 Select the feature subset using correlation and association rules.

3 Test the dataset with reduced set of attributes by classifier.

In this way attributes will get selected and further used for classification.

3.2. Database Classification : The reduced dataset after feature selection process is given for classification to check the accuracy of classifier. In this case two different classifiers are applied for evaluating performance such as neural network and decision tree. It displays the accuracy in percentage that how much data is classified correctly which shows that feature selection is an important process before classification.

3.3 Performance Measure

Performance of system is measured through accuracy of classifiers. Once feature selection is done on the selected medical dataset the attributes get selected

using different approaches feature selection methods and then classification is done on the reduced dataset of any medical dataset. Accuracy of the classifier is the main performance measure of the given system.

1 Accuracy of actual dataset is compared with accuracy of the reduced dataset. Which indicates system is performing better than existing methods.

2 Confusion matrix shows the performance by giving the correctly classified records.

3. In this way attributes will get selected and further used for classification.

Figure 1 shows overall procedure for classification of reduced data set.

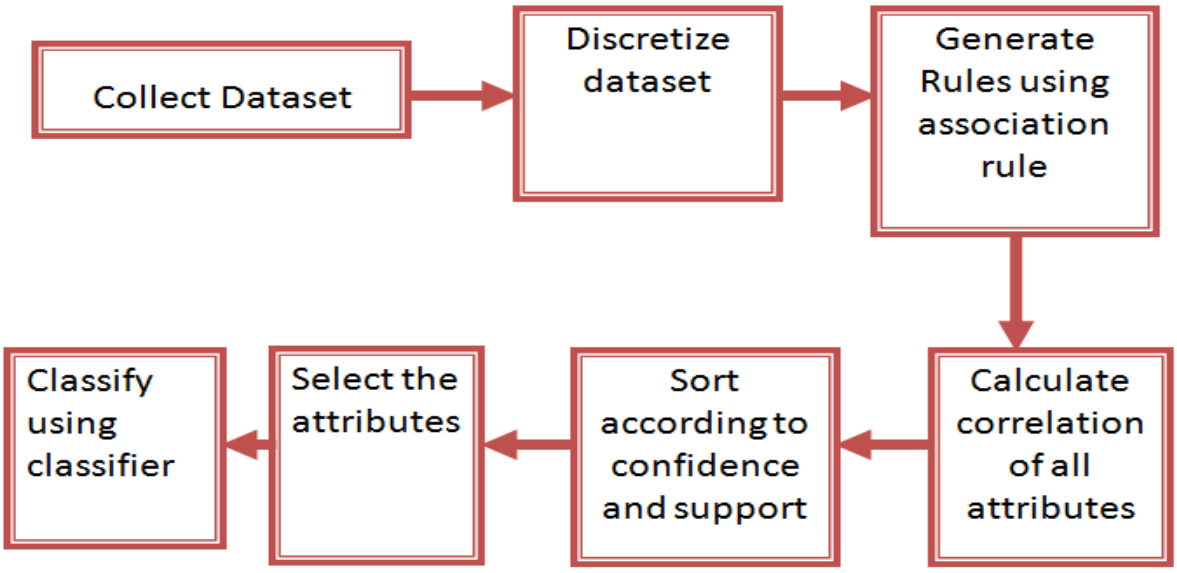

Figure 1. System Diagram of Classification using Feature Selection

4. RESULT & DISCUSSIONS

In this work experiments were done using data mining tool weka 3.6.3, Java, Mysql, Intel core i5 processor on Heart disease (Cleveland), Breast cancer , Diabetes downloaded from UCI standard repository .

The results are tabulated in Table 1 , Table 2 and Table 3 and discussed. We have used widely popular data mining tool WEKA 3.7.7 for checking the accuracy of classifiers on different datasets. Initially data is preprocessed while preprocessing data discretization is done and then classified using multilayer perceptron, J48 decision tree. Following table 1 shows results of classifiers among all three datasets considering all attributes. Table 2 shows results for features selected by genetic search method and accuracy of classifiers namely neural network and decision tree on reduced dataset. By using new approach of feature selection, we have selected the important attributes among the all given datasets. Table 3 shows the results of Correlation based feature selection on different datasets.

Table 1: Classification with all attributes

Classifier->	Neural Network	Decision Tree
Heart Cleveland	80.53%	77.89%
Brest Cancer	64.69%	75.52%
Diabetes	75.13%	73.83%

Table 2: Classification with Reduced data set using Genetic Search Method

Dataset (Total features)	No. of Selected Significant features	Selected Attributes	Neural Network	Decision Tree
Heart Cleveland(14)	08	3,5,8,9,10,11,12,13	78.20 %	73.87%
Breast Cancer(10)	05	3,4,5,6,9	77.34%	71.82%
Diabetes : Pima Indian (09)	03	2,6,8	73.95%	73.82%

For Heart Cleveland dataset out of 13 attributes 7 attributes are selected. The dataset with those attributes is given to both the classifier as neural network and decision tree and the performance is evaluated by checking and comparing the accuracy of reduced dataset against actual dataset and existing methods of feature selection. In Cleveland Heart disease dataset it is analyzed that for using genetic search 3,5,8,9,10,11,12,13 and using Correlation based Association rule mining algorithm 2,6,7,9,11,12, 13 are significant attributes. Thus from this three comparison we say that 9,11,12,13 i.e. exang, slope, ca and thal are most significant attributes for heart disease prediction. Experiments shows that Feature selection plays an important role.

Table 3: Classification using Correlation based Feature Selection

Dataset	No. of Selected Significant features	Selected Attributes	Neural Network	Decision Tree
Heart Cleveland(14)	07	2,6,7,9,11,12, 13	80.20 %	78.87%
Breast Cancer(10)	05	3,4,5,6,9	78.34%	73.82%
Diabetes : Pima Indian(09)	04	2,4,6,8	74.95%	73.9%

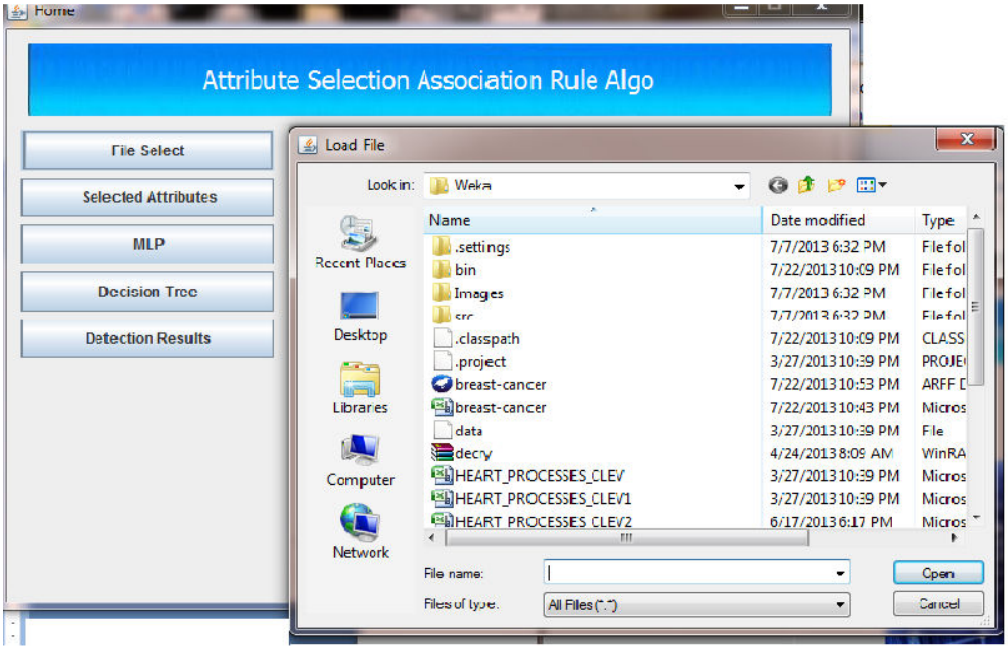

Figure 2. Selection of Database

To improve accuracy of classification which may further lead to form a decision support system to predict a disease and helps medical practitioner for decision making. Snapshots of Implementation shown in Figure 2, 3, 4.

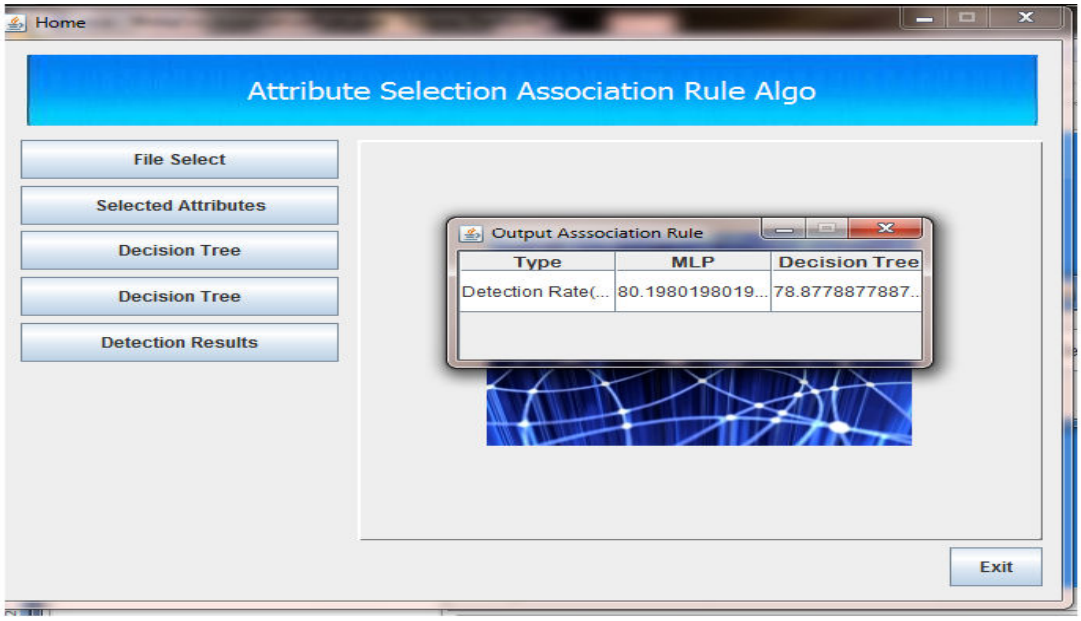

Figure 3. Results using Attribute Selection using Association Rule

5. CONCLUSION

Attribute selection method was implemented with association rule mining and correlation. The average accuracy was 78% with reduced features. Computational time was of the order O(n2). The accuracy of these dataset is compared with existing method as Genetic search. The average accuracy using Genetic Search was 76% with reduced features. The number of significant features selected by both of the above methods was half the number of features in original dataset. Hence the space requirement is reduced by a minimum of 50%.

It is found that accuracy of classifiers can be improved using attribute selection which also helps in minimizing space complexity.

This work can be further extended as a Decision Support System to help medical practitioners to make decisions in medical domain. Further extension to reduce time complexities can be done by improving the performance of algorithms.

6. REFERENCES

[1] L. Shahwan-Akl, "Cardiovascular Disease Risk Factors among Adult Australian-Lebanese in Melbourne," International Journal of Research in Nursing, 2010.

[2] J. Xie, J. Wu, and Q. Qian, "Feature Selection Algorithm Based on Association Rules Mining Method," 2009 Eighth IEEE/ACIS International Conference on Computer and Information Science, pp. 357–362, 2009.

[3] P. Andreeva, "Data Modelling and Specific Rule Generation via Data Mining Techniques," International Conference on Computer Systems and Technologies - CompSysTech, 2006.

[4] V. A. Sitar-Taut et al., "Using machine learning algorithms in cardiovascular disease risk evaluation," Journal of Applied Computer Science and Mathematics, 2009.

[5] K. Srinivas, B. K. Rani, and A. Govrdhan, "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks," International Journal on Computer Science and Engineering (IJCSE), vol. 2, no. 2, pp. 250-255, 2010.

[6] M. C. Tu, D. Shin, and D. Shin, "Effective Diagnosis of Heart Disease through Bagging Approach," Biomedical Engineering and Informatics,IEEE, 2009.

[7] H. Yan, et al., "Development of a decision support system for heart disease diagnosis using multilayer perceptron," in Proc. of the 2003 International Symposium on, vol. 5, pp. V-709- V-712.

[8] Y. Kangwanariyakul, et al., "Data mining of magnetocardiograms for prediction of ischemic heart disease," EXCLI Journal, 2010.

[9] Detrano, R.; Steinbrunn, W.; Pfsterer, M., "International application of a new probability algorithm for the diagnosis of coronary artery disease". American Journal of Cardiology, Vol. 64, No. 3, 1987, pp. 304-310.

[10] R. Agrawal, T. Imielinski, and A. N. Swami. "Mining association rules between sets of items in large databases." In Peter Buneman and Sushil Jajodia, editors, Proceedings of the 1993 ACM SIGMOD Intl. Conference on Management of Data, pages 207–216, Washington, D.C., 26–28 ,1993.

[11] Yao, Z.; Lei, L.; Yin, J., "R-C4.5 Decision tree model and its applications to health care dataset". proceedings of International Conference on Services Systems and Services Management 2005, pp. 1099-1103.

[12] Gennari, J., "Models of incremental concept formation".Journal of Artifcial Intelligence, Vol. 1, 1989, pp. 11-61.