SVM Ranking with Backward Search for Feature Selection in Type II Diabetes Databases

Sarojini Balakrishnan
Department of Computer Applications
K.L.N. College of Information Technology
Madurai, India
balakrishnan.sarojini@gmail.com

Ramaraj Narayanaswamy
Department of Computer Science & Engineering
G.K.M. College of Engineering & Technology
Chennai, India
ramaraj_tce@yahoo.co.in

Nickolas Savarimuthu
Department of Computer Applications
National Institute of Technology
Tiruchirappalli, India
nickolas@nitt.edu

Rita Samikannu
VIT Business School
VIT University
Vellore, India
ritasamikannu@gmail.com

Abstract—**Clinical databases have accumulated large quantities of information about patients and their clinical history. Data mining is the search for relationships and patterns within this data that could provide useful knowledge for effective decision-making. Classification analysis is one of the widely adopted data mining techniques for healthcare applications to support medical diagnosis, improving quality of patient care, etc. Usually medical databases are high dimensional in nature. If a training dataset contains irrelevant features (i.e., attributes), classification analysis may produce less accurate results. Data pre-processing is required to prepare the data for data mining and machine learning to increase the predictive accuracy. Feature selection is a preprocessing technique commonly used on high-dimensional data and its purposes include reducing dimensionality, removing irrelevant and redundant features, reducing the amount of data needed for learning, improving algorithms' predictive accuracy, and increasing the constructed models' comprehensibility. Much research work in data mining has gone into improving the predictive accuracy of the classifiers by applying the techniques of feature selection. The importance of feature selection in medical data mining is appreciable as the diagnosis of the disease could be done in this patient-care activity with minimum number of features. Feature selection may provide us with the means to reduce the number of clinical measures made while still maintaining or even enhancing accuracy and reducing false negative rates. In medical diagnosis, reduction in false negative rate can, literally, be the difference between life and death. In this paper we propose a feature selection approach for finding an optimum feature subset that enhances the classification accuracy of Naive Bayes classifier. Experiments were conducted on the Pima Indian Diabetes Dataset to assess the effectiveness of our approach. The results confirm that SVM Ranking with Backward Search approach leads to promising improvement on feature selection and enhances classification accuracy.**

Keywords— **Feature selection, false negative rate, predictive accuracy, backward search, SVM, classification accuracy**

I. Introduction

Now-a-days maintaining databases about demographic and pathological data for the patients are an essential task in healthcare industry to provide quality services. These databases provide descriptive features of patients and their respective disease diagnostics. Extracting knowledge from these data sources can lead to discovery of trends and rules for later diagnostic tools. Consequently, the predictability of disease will become more effective and early detection of disease will aid in increased exposure to required patient care and improved cure rates [2]. Through the manipulation of medical databases, the study of medical informatics supplies a knowledge base to the Medical industry. Medical informatics is the study of how to create, shape, share and apply medical knowledge [3]. This knowledge base provides a discovery tool used to create clinical guidelines, formal medical languages and information systems.

Data mining techniques are applied on the huge volumes of stored clinical data to discover the trends and patterns hidden within the data that could significantly enhance our understanding of disease progression and management of the disease [4].

A high degree of predictive accuracy is expected in the field of medicine. The Prediction accuracy of any data mining technique is based on the quantity and quality of the data [5]. The data gathered in medicine is generally collected as a result of patient-care activity to benefit the individual patient and as a result, medical databases may contain data that is redundant, incomplete, imprecise or inconsistent, which can affect the use of the results of the data mining techniques. So, mining the medical data may require more data reduction and data preparation than data used for other applications [6, 7]. Not all features used in describing data are equally important for all problems. The irrelevant or redundant features, the noisy data, makes knowledge discovery during training very difficult. The accuracy of the result produced by a classifier depends also on the number or interrelationships of the features of the instances used for learning. Also some robust algorithms, such as Naïve Bayes, which are quite robust with respect to irrelevant

1-4244-2384-2/08/$20.00 © 2008 IEEE

features, are known to degrade quickly, if correlated features are added [8].

Feature selection is the process of identifying and removing as much of the irrelevant and redundant features as possible [9]. Feature selection prior to learning can be beneficial. Reducing the dimensionality of the data reduces the size of the hypothesis space and allows algorithms to operate faster and more effectively [4]. Particularly, in medical data mining, feature selection may provide us with the means to reduce the number of clinical measures made while still maintaining accuracy and reducing false negative and false positive rates. It is undesirable to tell a healthy patient that he is sick(false positive), but it is worst, particularly with respect to life-threatening ailments, to tell a sick patient that he is healthy(false negative) and needs no further physical examination or treatment.

In the proposed approach, we use SVM based feature ranking and backward search for selecting the optimal feature set that enhances the predictive accuracy of the classifier. The features are ranked using feature-relevance and the feature subset is evaluated by applying to Naïve Bayes Classifier to select the optimum one which produces better predictive accuracy.

The remaining paper is organized as follows: Section II presents the details of the related research work on the problem; Section III gives the description of the Dataset. Section IV presents the Objective and Problem definition. Section V gives experimental results and Sections VI and VII deal with performance analysis and conclusion, respectively.

II. RELATED WORK

Various data mining techniques have recently been developed to extract knowledge from medical databases. Data mining is the search for relationships and global patterns that exist in large databases but are `hidden' among the vast amount of data, such as a relationship between patient data and their medical diagnosis. These relationships represent valuable knowledge about the database and the objects in the database [10].

Numerous applications of data mining in medical domain are found in the literature. Predicting breast cancer survivability using data mining techniques [11], application of data mining to discover subtle factors affecting the success and failure of back surgery which led to improvements in care [12], data mining classification techniques for medical diagnosis decision support in a clinical setting [13] and the techniques of data mining used to search for relationships in a large clinical database [14].

Supervised learning systems such as classification have been successfully applied in a number of medical domains, for example, in localization of a primary tumor, prognostics of recurrence of breast cancer, diagnosis of thyroid diseases, and rheumatology [15].

Researchers and practitioners realize that in order to use data mining tools effectively data preprocessing is essential to successful data mining [16, 17]. Feature selection is one of the important and frequently used techniques in data preprocessing for data mining [18]. It reduces the number of features, removes irrelevant, redundant, or noisy data, and brings the immediate effects for applications by speeding up a data mining algorithm, improving mining performance such as predictive accuracy and result comprehensibility [19]. The importance of feature selection in medical domain is found in [4].

Many authors have reported improvement in the performance of the classifier when feature selection algorithms are used [20, 21, 22]. In many pattern recognition applications, identifying the most characterizing features (or attributes) of the observed data, i.e., feature selection [23]-[27] is critical to minimize the classification error. It has also been reported that the combinations of individually good features do not necessarily lead to good classification performance. In other words, “the m best features are not the best m features” [23, 24]. Relevance of a feature does not imply that it must be in the optimal feature subset [28]. Various feature selection methods use different measures to evaluate the goodness of individual features. For example, information measures, distance measures and dependence measures. Features are ranked according to their values on this measure [29]. The feature ranking treats each feature as independent, and compares them to determine the order or importance [30]. One can simply choose the first X features as the selected feature subset; X is decided according to some domain knowledge or a user-specified threshold value [31].

Two approaches are proposed for the problem of feature selection. The filter model chooses features by heuristically determines “goodness/relevant” or knowledge, while the wrapper model does this by the feedback of classifier evaluation, or experiment. Research has shown the wrapper model outperforms the filter model comparing the predictive power on unseen data [32]. A wrapper [26], [27] is a feature selector that convolves with a classifier (e.g., Naive Bayes classifier), with the direct goal to minimize the classification error of the particular classifier. Wrappers use classification accuracy on training dataset as a measure of how well a subset of features performs, thus turn the problem of feature subset selection into an optimization problem. Many researchers have developed various feature selection algorithms using different evaluation criterions and searching strategies. Backward selection method may perform better than forward selection by eliminating the variables that do not provide the best separation [29].

Research [33] shows that the legitimate way of evaluating features is through the error rate of the classifier being designed. The classification error rate is used as a performance indicator for a mining task, for a selected feature subset; simply conduct the “before-and-after” experiment to compare the error rate of the classifier learned on the full set of features and that learned on the selected subset [16].

Naïve Bayes (NB), a special form of Bayesian Network has been widely used for data classification in that its predictive performance is competitive with state-of-the-art classifiers [30]. Research [34] shows Naïve Bayes performs well in spite of strong dependencies among attributes. It is proved that cross-validation avoids over-fitting [35] during performance evaluation of the classifier.

For supervised learning, the primary goal of classification is to maximize predictive accuracy; therefore, predictive accuracy is generally accepted and widely used as the primary measure by researchers and practitioners [36]. The performance of a classifier can be visualized by using a Receiver Operating Characteristic (ROC) curve [12, 13]. ROC graphs depict the tradeoff between true positive rates [37] and false positive rates [37] of classifiers [38]. ROC analysis has been extended for use in visualizing and analyzing the behavior of diagnostic systems [39]. The accuracy of the classifier is compared based on the area under the ROC curve, abbreviated **AUC** [2, 14].

III. Dataset

The Pima Indian diabetes dataset [1], includes 768 complete instances described by 8 features (labeled as number of times pregnant, glucose tolerance test, diastolic blood pressure, triceps skin fold thickness, 2-hour serum insulin, body mass index, diabetes pedigree function and age). The class distribution is class value 1 interpreted as "tested positive for diabetes" in 268 numbers of instances and class value 0 in 500 numbers of instances. There is no missing data present in the training dataset.

IV. OBJECTIVE / PROBLEM DEFINITION

The objective of the proposed work is to search for an optimal feature subset, using SVM-Ranking with backward search, which enhances the predictive accuracy of the naïve bayes Classifier.

The feature selection is done in two-phases. In the first phase, feature relevance weight is assigned to each individual feature using Support Vector Machine attribute evaluation. These weights are sorted to rank the features. In the second phase, backward search is used to remove the least ranked features one at a time and the effect of the feature in improving the performance of the accuracy of the classifier is studied. The feature subset which enhances the accuracy of the classifier is considered as the optimal feature subset.

Given the input data D as a table of N samples and M features $X = \{X_i,\ i = 1\ldots M\}$, and the target classification variable c, the feature selection problem is to find from the M-dimensional observation space, S^M, a subspace of m features, S^m. The total number of subspaces is 2^M.

SVM Attribute Evaluation with Ranker search method does the feature ranking. The features are ranked by the square of the weight assigned by the SVM.

Firstly, the classification is performed with the whole set of features FS_k where k=M …2. The backward search removes a least rank feature one at a time and is applied to the naïve bayes classifier with the constraint that the resultant feature set FS_{k-1} leads to an increase in the classification accuracy acc_{k-1} better than acc_k. In the consecutive iterations, for the removal of each feature, the respective classification accuracy is calculated. If the corresponding acc_{k-1} is larger than acc_k, then there is a gain in classification accuracy. This decremental selection procedure is repeated until the termination condition is satisfied. The feature subset, which gives highest classification accuracy, is chosen as the resultant optimal feature set.

V. EXPERIMENTAL RESULTS

We used Naïve Bayes classifier and the Support Vector Machine Attribute Evaluation feature selection technique available in the machine learning library with Java implementation "WEKA 3.4.2" [40] for our experiments. The experiments are performed on the PIMA dataset [1].

Classification is done on Naïve Bayes classifier and accuracy is evaluated using 10-fold cross validation test. Cross-validation involves breaking a dataset into 10 pieces, and on each piece, testing the performance of a predictor build from the remaining 90% of the data. The classification accuracy was taken as the average of the 10 predictive accuracy values.

The performance of the proposed approach is analyzed using two criteria: the accuracy of the classifier and the area under the ROC. The accuracy of the classifier is calculated using the formula [37].

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

Where TP: the number of true positives (number of 'YES' patients predicted correctly), TN: the number of true negatives (number of 'NO' patients predicted correctly), FP: the number of false positives (number of 'YES' patients predicted as 'NO') and FN: the number of false negatives (number of 'NO' patients predicted as 'YES')

The performance of a classifier is visualized by using a Receiver Operating Characteristic (ROC) curve. The area under the ROC curve (AUC) is used as a tool for comparing the performance of the classifier on the removal of each feature. The curve that has a larger AUC is better than the one that has a smaller AUC.

VI. PERFORMANCE ANALYSIS

The performance of the proposed feature selection approach is discussed in terms of the five main parameters of the classification results namely Accuracy, True Positive, True Negative, Correctly Classified Instances and Incorrectly Classified Instances.

Table I shows the experimental results taken. The accuracy of the classifier shows a steady rise and attains a peak value for the fourth iteration i.e. for the feature subset with last 3 least ranked features removed. After that the performance deteriorates. The True Positive and True Negative parameters also show a different pattern at this point. There is a notable increase in the True Negative parameter and decline in the True Positive parameter. That is the prediction of 'Yes' class is low compared to the previous iterations. But the overall predictive accuracy is high. In medical domain it is purely acceptable as it is desirable that the classifier to correctly classify positive instances but it is appreciable if the classifier reduces the false negatives. In other words a healthy patient may be identified as sick but not a sick patient as healthy, which makes the medical diagnosis a failure one. So, it makes a sense to sacrifice the precision of positive classifications in exchange for improving the precision of negative calculations. The accuracy of the classifier on the removal of each least rank feature is shown in Table I.

TABLE I CLASSIFICATION RESULTS

Iterations	Feature Subset	Classification results				
		Accuracy	*True Posi--tive*	*True Nega—tive*	*No. of correctly classi--fied insta--nces*	*No. of In-correctly classified instances*
1	All {a1..a8}	76.3021	164	422	586	182
2	A4 removed {a1,a2,a3,a5, a6,a7,a8}	76.8229	167	423	590	178
3	A5 removed {a1,a2,a3,a6, a7,a8}	76.8229	167	423	590	178
4	A8 removed {a1,a2,a3, a6,a7}	77.7344	157	440	597	171
5	A3 removed {a1,a2,a6,a7}	77.2135	155	438	593	175
6	A7 removed {a1,a2,a6}	75.6510	148	433	581	187
7	A1 removed {a2,a6}	76.4323	139	448	587	181

Fig. 1 shows the graphical representation of the performance of the classifier on each iteration. The predictive accuracy of the classifier for all features is 76.3021%. There is a slight increase in the performance of the classifier after the removal of the least rank feature a4. The value remains the same for the removal of feature a5. The peak performance of 77.7344% is attained for the feature subset {a1, a2, a3, a6, a7} after the removal of three least rank features a4,a5,a8. The performance degrades after that.

Fig. 2 shows the behaviour of the parameters, No. of Correctly and No. of Incorrectly Classified instances. There is an increase in the values of correctly classified instances for the first three iterations and a decrease in the number of instances incorrectly classified. After the fourth iteration, there is a decrease in the number of instances correctly classified and increase in the number of instances incorrectly classified. Hence, the feature subset obtained after the removal of three least ranked features is considered as an optimal subset and the searching process is terminated. The feature subset derived is the optimal feature subset

Another way of evaluating the performance of a classifier is by the analysis of the ROC curve. The two-dimensional ROC curve is defined by the False Positive Rate (FPR) on the x-axis and the True Positive Rate(TPR) on the y-axis [37] where TPR determines a classifier performance on classifying positive instances correctly among all positive samples and FPR, on the other hand, defines how many incorrect positive results occur among all negative samples.

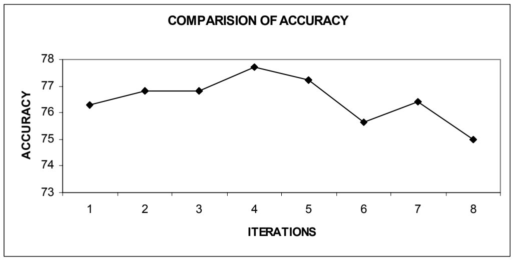

Figure 1: Accuracy of the classifier on each iteration

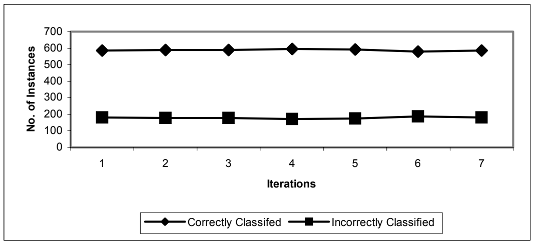

Figure 2: Performance of the classifier on each iteration

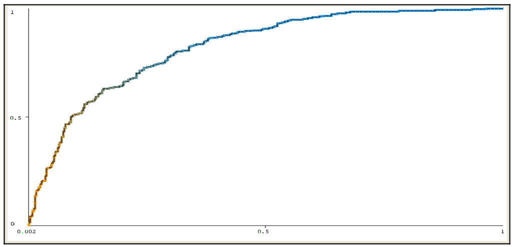

Figure 3: ROC – before applying feature selection

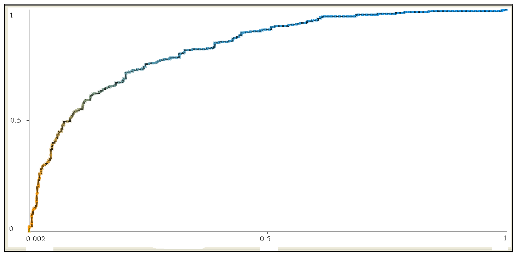

Figure 4: ROC – after applying feature selection

The area under ROC for the whole set of features is 0.8186 and for the optimal feature subset it is 0.8277. The improved area of the ROC curve proves that the proposed feature selection approach could enhance the predictive accuracy of the classifier with minimal number of features. It is depicted in Figure 3 and Figure 4.

VII. Conclusion

In this research work, we propose a novel feature selection approach that is experimented on the type II diabetes dataset. The performance of the approach is analyzed by comparing the classification accuracy of the Naïve Bayes classifier. Our approach with 37.5% feature reduction produces an increase of 1.88% of classification accuracy. The methodology uses minimal iterations to derive the optimal feature subset, which gives better classification accuracy. It helps the physician to diagnose the disease with less number of features. The approach is simple and effective and augments the argument simple methodologies are better for medical data mining.

The future work may focus on the other medical datasets with the objective of deriving maximum predictive classification accuracy.

References

[1] C.L. Blake, C.J. Merz, UCI repository of machine learning databases, Website: http://www.ics.uci.edu/~mlearn/MLRepository.html, 1998.

[2] Roshawnna Scales, Mark Embrechts, "Computational intelligence techniques for medical diagnostics".

[3] Coiera, E., 1997, Guide to medical Informatics, the internet and telemedicine, http://www.coiera.com/.

[4] Ranjit Abraham, Jay B.Simha, Iyengar S. "Medical datamining with a new algorithm for feature selection and Naïve Bayesian classifier", 10th International Conference on Information Technology.

[5] Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell FE Jr, Marks Jr, Winchester DP, Bostwick DG,"Artificial neural networks improve the accuracy of cancer survival prediction", Cancer 1997; 79:857—62.

[6] Lavrac N. "Selected techniques for data mining in medicine" Artif Intell Med 1999; 16:3—23.

[7] Cios KJ, Moore GW. " Uniqueness of medical data mining", Artif Intell Med 2002;26:1—24.

[8] R. Kohavi, and G. John. "Wrappers for Feature Subset Selection", AIJ special issue on relevance. http://robotics.stanford.edu/ ronnyk.

[9] Liu H., Sentino R, "Some issues on scalable Feature Selection, Expert Systems with Application", vol 15, pp 333-339, 1998.

[10] Dilly,Ruth.DataMining.2002 http://www.pcc.qub.ac.uk/tec/courses/datamining/stu_notes/dm_book_1.html.

[11] Dursun Delen*, Glenn Walker, Amit Kadam "Predicting breast cancer survivability: a comparison of three data mining methods" Artificial Intelligence in Medicine doi:10.1016/j.artmed.2004.07.002.

[12] Hedberg, SR. (1995) The data gold rush. Byte, 1995;Oct :83-88.

[13] Herron, "Machine Learning for Medical Decision Support: Evaluating Diagnostic Performance of Machine Learning classification Algorithms"

[14] Prather J. C., Lobach D. F., Goodwin L. K., Hales J. W.,Hage M. L., Edward Hammond W., "Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse", 1997.

[15] Richards G, Rayward-Smith VJ, Sonksen PH, Carey S, Weng C. "Data mining for indicators of early mortality in a database of clinical records", Artif Intell Med 2001; 22:215—31.

[16] H. Liu and H. Motoda "Feature Selection for Knowledge Discovery and Data Mining", Boston: Kluwer Academic Publishers, 1998.

[17] D. Pyle. "Data Preparation for Data Mining", Morgan Kaufmann Publishers, 1999.

[18] H. Liu and H. Motoda "Feature Extraction, Construction and Selection: A Data Mining Perspective.", Boston: Kluwer Academic Publishers, 1998. 2nd Printing, 2001.

[19] H. Liu and L. Yu, "Feature Selection for Data Mining".

[20] Almuallim, H., and Dietterich, T.G., "Efficient algorithms for identifying relevant features" In Proceedings of the Ninth Canadian Conference on Artificial Intelligence, Vancouver, BC: Morgan Kaufmann, 1992.

[21] Aha, D.W., and Bankert, R. L., **"A** comparative evaluation of sequential feature selection algorithms", In D. Fisher & J.-H. Lenz (Eds.), Artificial Intelligence and Statistics **V.** New York: Springer-Verlag. 1996.

[22] W Siedlecki and J. Skalansky, "On automatic feature selection," Int. J. Pattern Recog. Art. Intell. vol. 2, no.2.p~1.9 7-220. 1988.

[23] A. Webb, "Statistical Pattern Recognition", Arnold, 1999.

[24] A.K. Jain, R.P.W. Duin, and J. Mao, "Statistical Pattern Recognition: A Review," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 4-37, Jan. 2000.

[25] N. Kwak and C.H. Choi, "Input Feature Selection by Mutual Information Based on Parzen Window," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 12, pp. 1667-1671, Dec. 2002.

[26] P. Langley, "Selection of Relevant Features in Machine Learning," Proc. AAAI Fall Symp. Relevance, 1994.

[27] R. Kohavi and G. John, "Wrapper for Feature Subset Selection," Artificial Intelligence, vol. 97, nos. 1-2, pp. 273-324, 1997.

[28] W Siedlecki and J. Skalansky, "On automatic feature selection," Int. J. Pattern Recog. Art. Intell. vol. 2, no.2.p~1.9 7-220. 1988.

[29] Hassan Sabzevari , Mehdi Soleymani , Eaman Noorbakhsh "A comparison between statistical and Data Mining methods for credit scoring in case of limited available data",

[30] Duda and Hart. "Pattern classification and scene analysis" 1973, John Wiley and Sons, NY.

[31] Y, Liu, and M. Schumann, Data mining feature selection for credit scoring models", Journal of the Operational Research Society 56, 1099–1108, 2005.

[32] G.H. John, R. Kohavi, K. Pfleger, “Irrelevant features and the subset selection problem”, Proceedings of the Eleventh International Conference of Machine Learning, Morgan Kaufmann Publishers, San Francisco, CA (1994) 121-129.

[33] W. Siedlecki and J. Sklansky. ”On automatic feature selection”, International Journal of Pattern Recognition and Artificial Intelligence, 2(2):197.220, 1988.

[34] Domingos, P., Pazzani, M., “On the Optimality of the Simple Bayesian Classifier under Zero-One loss”, Machine Learning, 29(2/3): 103-130, November/December 1997.

[35] Dietterich T. G.: “Statistical Tests for Comparing Supervised Classification Learning Algorithms”, Tech. Report. Department of Computer Science. Oregon State University. (1996).

[36] G.H. John, R. Kohavi, and K. Pfleger “Irrelevant feature and the subset selection problem”, In W.W. asnd Hirsh H. Cohen, editor, Machine Learning: Proceedings of the Eleventh International Conference, pages 121–129, New Brunswick, N.J., 1994. Rutgers University.

[37] Rayner Alfred, “Knowledge Discovery: Enhancing Data Mining and Decision Support Integration”

[38] A. Swets, R. M. Dawes, and J. Monahan. “Better decisions through science”, Scientific American, 283:82–87, October 2000.

[39] J. Swets. “Measuring the accuracy of diagnostic systems. Science”, 240:1285–1293, 1988.

[40] Witten, I., Frank E. “Data Mining: Practical Machine Learning Tools with Java Implementations”, Morgan Kaufmann, San Francisco, 2000