A Wrapper Based Feature Subset Evaluation Using Fuzzy Rough K-NN

Dr.B.SAROJINI

Department of Computer Science
Avinashilingam Institute for Home Science and Higher education for Women University,
Coimbatore, Tamil Nadu, India.
dr.b.sarojini@gmail.com

***Abstract*— Application of data mining techniques on medical databases is a challenging task considering the high volume, complexity, and poor quality of the medical databases. Data mining in medical domain could greatly contribute in the discovery of disease associations and provide the physicians with valuable and previously unavailable knowledge. Among the hundreds or thousands of features in the medical databases only very few features predominantly contribute for medical decision making. The small subset of informative features, selected from a whole set of features, may carry enough information to construct reasonably accurate prognostic or diagnostic models. The objective is to find the optimal feature subset of the medical databases that could enhance the Accuracy, the Sensitivity and the Specificity of the classification algorithms. In this paper, a wrapper based feature subset selection approach with Fuzzy Rough K-Nearest Neighbor (Fuzzy Rough KNN) classification algorithm is used to select the discriminatory features of the Indian Liver Patient Dataset. The empirical results show that the feature selection approach could achieve a feature reduction of 70% and enhance the performance of the classifier Fuzzy Rough KNN by 7%.**

Keyword-Data mining, Feature selection, Accuracy, Sensitivity, Specificity, Classifier Subset Evaluation, Fuzzy Rough KNN

I. INTRODUCTION

The diagnosis of disease is a vital and intricate job in the field of medicine that needs to be executed accurately and efficiently. The application of data mining in medicine has proved successful in the areas of diagnosis, prognosis and treatment [1,2]. The discovered patterns may represent valuable knowledge for medical discoveries, for example identification of combinations of features that would lead to diagnosis of the disease. Studies show that improved medical diagnosis and prognosis may be achieved through automatic analysis of patient data stored in medical records i.e. by learning from past experiences [3]. The application of data mining techniques on medical databases is a challenging process considering the high volume, complexity, and poor quality of the medical databases. The medical data is collected as a patient-centric activity to benefit the individual patient without any specific research purpose [4,5] and so the data collected may include missing, corrupted, inconsistent, or non-standardized data [6]. The quality of prognostic or diagnostic prediction models depend on the quality of the data and researchers realized that in order to use data mining tools on these medical databases effectively, data preprocessing is essential [7,8,9]. The application of efficient and sound data preprocessing procedures could reduce the amount of data to be analyzed without losing any critical information, improve the quality of the data, enhance the performance of the actual data mining algorithms and reduce the execution time of mining algorithms [10].

Feature selection is an efficient data preprocessing technique that aims to identify and remove as much of the irrelevant and redundant information as possible. Research has proven feature selection to be an effective measure in reducing dimension to improve the classification and predictive accuracy of the diagnostic models [11]. Feature selection identifies the optimal features of the dataset that could improve the performance of the classification algorithms. The feature selection techniques preserve the original semantics of the variables, offering the advantage of interpretability by a domain expert, which is very much required in the medical domain. 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 clinical tests, each with different financial cost, diagnostic value and associated risk, reducing the cost and time. The integration of feature selection with data mining techniques could help in non-invasive diagnosis and decision support. For example, conducting a biopsy in women to detect cervical cancer is an invasive, costly and painful process. Thangavel et al., (2006)[12] found a set of interesting attributes that could be used by doctors as additional support on whether or not to recommend a biopsy for a patient suspected of having the cervical cancer. Though a number of feature selection methods that enhance the performance of the mining algorithms are available, still the research goes on to identify more informative features of the dataset. The objective of this research work is to prove that a small subset of features may carry enough information to construct reasonably accurate diagnostic models. In medical domain even

very little improvement in diagnosis accuracy is significant as it means that one more patient is correctly diagnosed.

In this research work, a wrapper based feature subset selection method Classifier Subset Evaluation with Best Search method is used to identify the optimal feature subset of the Indian Liver Patient Dataset. The feature selection approach is validated by studying the performance of the classifier with the reduced feature subset. The enhanced accuracy of the classification algorithm proves that the selected feature subset carry enough information for accurate classification. The performance of Fuzzy Rough K-NN Classifier is analyzed in terms of the accuracy of the classifier, Sensitivity and Specificity.

II. LITERATURE SURVEY

A Majority of research works have been carried out in the area of feature selection and predictive classification with the goal of improving accuracy. Many authors have reported improvement in the performance of the data mining algorithms when feature selection algorithms are used. The survey of literature shows that improved classification accuracy is attained by applying various feature selection algorithms to different medical datasets of UCI Machine Learning Repository. Much research efforts have recently been made to the development of effective feature evaluation criteria, and the development of efficient search methods.

Hongmei Yan et al., (2008) [13] proposed a real-coded genetic method to select 24 critical features from 40 original features that are essential to the heart diseases diagnosis. R.E.Abdel-Aal (2005) [14] used the group method of data handling (GMDH) to reduce the dimensionality to 22% and 54% for the breast cancer and heart disease data, respectively, leading to improvements in the overall classification performance. Wang et al., (2009)[15] extracted 20 critical features are selected from original 105 features from the liver cirrhosis dataset. Polat et al., (2006)[16] diagnosed heart disease diagnosis using a hybrid expert system combining AIRS classifier and fuzzy weighted pre-processing and achieved an enhanced accuracy of 96.39% with 10-fold cross-validation. Jelonek et al.,(1993)[17] used rough sets to select attributes for classification of histological images with neural networks. The approach reduced dimensionality to about 11% of the original set Empirical results showed an improvement in the performance of the classifier also. The highest classification accuracy of 82.05% was attained by Polat et al., (2008) [18] who presented a cascade learning system based on Generalized Discriminate Analysis (GDA) and Least Square Support Vector Machine (LSSVM) to the diagnosis of Pima Indian Diabetes disease which otherwise gave the accuracy of 72%. Classification of ophthalmological data also showed that the decrease of classification parameters, from 14 to 3, noticeably increased accuracy from 70% to 80%[19]. Cheng et al., (2006) [20] compared the expert judgment and Correlation-based Feature Selection (CFS) strategies and showed that the CFS strategy outperformed expert judgment; however results of both approaches delivered more accurate predictions than that with full data set. Li et al., (2004) [21] explored a novel analytic cancer detection method with different feature selection methods and selected proteomic patterns and reported improvement in terms of detection performance. Su et al., (2006) [22] used four different data mining approaches to select the relevant features from the data to predict diabetes. Piramuthu (2004) [23] showed through empirical results the enhancement in classification algorithms when feature selection algorithms are used. Sarojini et al.,(2008,2009,2011) [24,25,26] proved that the feature selection approaches indeed improve the performance of various classification algorithms through a number of empirical results. Tsang-Hsiang Cheng et al., (2006) [20] adopted correlation feature selection (CFS) method to obtain the feature subset about cardiovascular disease. The empirical results show that the selected the feature subset improved the predictive power of the classifier.

III. DATASET DESCRIPTION

The Indian Patient Liver data set [27] contains 416 liver patient records and 167 non liver patient records. The data set was collected from north east of Andhra Pradesh, India. Selector is a class label used to divide into group (liver patient or not). This data set contains 441 male patient records and 142 female patient records. The attributes are (i).Age of the patient (ii).Gender of the patient (iii) Total Bilirubin (iv).Direct Bilirubin (v).Alkphos Alkaline Phosphotase (vi).Sgpt Alamine Aminotransferase (vii).Sgot Aspartate Aminotransferase (viii) Total Protiens (ix). ALB Albumin (x).A/G Ratio Albumin and Globulin Ratio and (xi) Selector field used to split the data into two sets (labeled by the experts). There is no missing data in the dataset.

IV. PROPOSED METHODOLOGY

This research work focuses on selecting the optimal feature subset with informative and discriminatory features of the medical datasets that is required for quality medical diagnosis. The criterion for the selection of the feature subset is that the selected feature subset should enhance the performance of the classifier. The proposed feature selection is evaluated based on the percentage of improvement in the performance of the classifier and the percentage of feature reduction.

The K-NN classification is suitable for data that is only partially exposed to the system prior employment [28] and Rough sets theory is based on the theory that the misclassification is due to imperfect learning space, i.e. imperfect feature vector description about the elements in the universe. The rough uncertainty is integrated into the fuzzy K-NN classifier forming the Fuzzy Rough k-NN classifier. Research shows fuzzy-rough NN classifier could perform better under partially exposed and unbalanced domain [29]. Also the fuzzy-rough NN approach contains not only upper but also lower membership degree which helps in drawing more meaningful interpretation from the output which in return provides the decision maker more valuable information. This characteristic makes fuzzy-rough NN a suitable classifier for medical data classification.

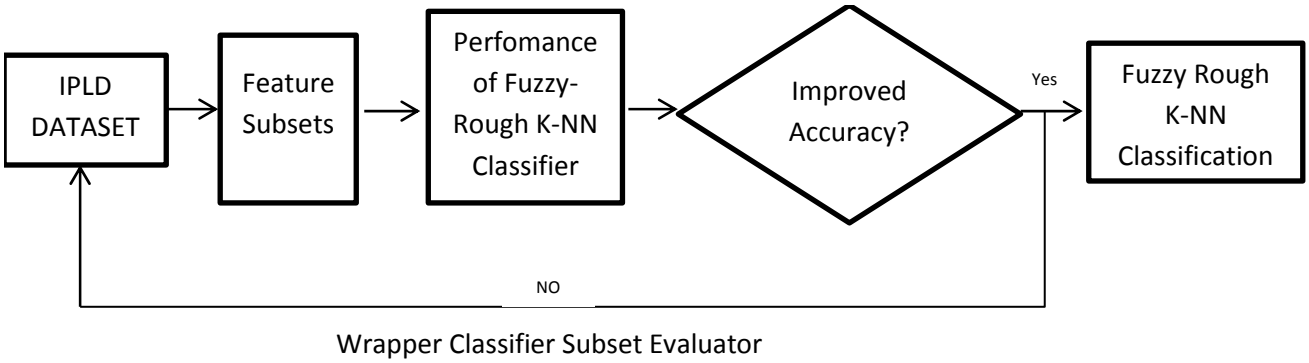

FIGURE1. THE FRAMEWORK OF THE PROPOSED WRAPPER BASED SUBSET EVALUATION

Figure 1 shows the framework of the proposed wrapper based subset evaluation. The classification algorithm is used as the subset evaluator and so the best feature subset that enhances the performance of the classification algorithm is chosen. The method considered the 311 possible feature subsets of the IPLD and an optimal feature subset with 3 features is selected.

V. EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

The experiments are performed using the machine learning library with Java implementation WEKA [30]. The experiments are performed on the Indian Liver Patient Dataset. The performance of the proposed approach is analyzed using two criteria: the increase in the accuracy of the classifier and reduce in the number of features. The accuracy of the classifier is calculated using the formula.

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

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

The Classifier Subset Evaluator uses Fuzzy Rough NN classifier to estimate the 'merit' of the possible feature subsets of the dataset. The 'merit' considered is the minimum classification error. The Best-first search method searches the space of attribute subsets and evaluates all possible feature subsets of the dataset against the Fuzzy-Rough NN classifier before selecting the best optimal feature subset of the dataset. A total of 311 feature subsets of the dataset are considered before selecting an optimal feature subset with 3 features. The experimental results show the feature subset with three features namely Age of the patient, Alkphos Alkaline Phosphotase, Sgot Aspartate Aminotransferase {1,5,7} is selected as the optimal feature subset. There is a feature reduction of 70%. In medical domain, feature reduction is appreciable as reduce in the number of features means the diagnosis of the disease with less number of tests or symptoms.

The Fuzzy Rough NN classification algorithm is applied to the reduced feature set of IPLD. The accuracy is evaluated using 10-fold cross validation. The performance of Fuzzy Rough NN classification algorithm is analyzed 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 1 shows the experimental results of Fuzzy Rough NN classifier before and after feature selection. The results prove that the proposed feature selection has improved the performance of the classifier. The performance has improved by 7.4%.

TABLE I
Classification Results

Feature Subset	Classification Results				
	Accuracy%	True Positive	True Negative	Correctly classified instances	Incorrectly classified instances
Whole feature set	66.038	298	87	385	198
Reduced feature {1,5,7}	73.413	322	96	428	155

Sensitivity and Specificity indicate how well the classifier discriminates between case with positive and with negative class. The discriminating ability of the classifier measured as Sensitivity and Specificity is very important in medical domain. Table II shows the comparative performance of the classifier analyzed in terms of other performance metrics like True Positive Rate or Sensitivity and True Negative Rate or Specificity. Sensitivity and Specificity are measured using the following formulae.

$$\text{Sensitivity} = (TP)/(TP+FN)$$

$$\text{Specificity} = (TN)/((TN+FP)$$

The Sensitivity measures the proportion of actual positives which are correctly identified as such (the percentage of sick people who are identified as having the disease (True Positive rate)); and the Specificity measures the proportion of negatives which are correctly identified (the percentage of healthy people who are identified as not having the disease (True Negative rate)).

TABLE III
Sensitivity and Specificity

Feature Subset	Classification Results			
	False Negative	False Positive	Sensitivity %	Specificity %
Whole feature set	118	80	71.63	52.10
Optimal feature subset {1,5,7}	94	71	77.40	57.48

It is evident that after feature selection there is a significant decrease in the number of False Negatives and False Positives. An improved sensitivity and Specificity reduces the probability of mistake in diagnosis (a sick patient as healthy) and the probability of unnecessary medical resource wasting (diagnose a healthy patient as sick).

VI. Conclusion

Healthcare to common man could be achieved by bringing both computer technocrats and medical professionals together to develop knowledge-driven systems to improve the quality of life. In this research work, the features, which are less influential on the predicted output, are removed and an optimal feature subset that enhances accuracy of the classifier is obtained. The empirical results prove that the performance of the classification algorithm is enhanced with small number of discriminatory features.

Acknowledgment

The author wholeheartedly thanks the UGC, New Delhi for its financial assistance under UGC-start-up research project grant No. F.20-1(13) /2012 (BSR)/20-1(13)/2012 (BSR) under which this research work is carried out.

References

[1] Cheng T.S.,J.Chen and Y.H.Kao.(2010). A novel hybrid protection technique of privacy-preserving data mining and anti-data mining. Inform. Technol. J., 9 :500-505.
[2] Lin.,R-H.(2009). An intelligent model for liver disease diagnosis. Artif. Intell.Med., 47: 53-62.
[3] Lavrac N, (1999). Selected techniques for data mining in medicine. Artificial Intelligence in Medicine 16(1), 3-23.
[4] Li, J., Fu, A. W.-C., He, H., Chen, J., &Kelman, C. (2005). Mining risk patterns in medical data. In Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining. pp. 770–775.
[5] Tsumoto, S. (2000). Problems with mining medical data. In The twenty-fourth annual international conference on computer software and applications. pp. 467–468.
[6] Cios K J and Moore G W, 2002. Uniqueness of Medical Data Mining. Artificial Intelligence in Medicine 26(1-2), 1-24.
[7] Wu X, Holmes G and pfahringer B (2008). Mining arbitrarily large datasets using heuristic k-nearest neighbor search. In Wobcke W and Zhang M, (Eds) Proc. of Twenty-First Australian Joint conference on Artificial Intelligence, Advances in Artificial Intelligence(AI 2008). LNAI 5360. Auckland, NZ: Springer, pp.355-361.
[8] Thongkam, J., Xu, G., Zhang, Y., & Huang, F. (2008a). Breast cancer survivability via AdaBoost algorithms. In The Australiasian workshop on health data and knowledge management, Vol. 80, pp. 1–10.
[9] Thongkam, J., Xu, G., Zhang, Y., & Huang, F. (2008b). Support vector machines for outlier detection in cancers survivability prediction. In International workshop on health data management, pp. 99–109.

[10] ParaskevasOrfanidis and David J. Russomanno, (2008). Preprocessing enhancements to improve data mining algorithms. International Journal of Business Intelligence and Data Mining 3(2), pp.196-211.
[11] Günter S and Bunke H, 2004. Feature selection algorithms for the generation of multiple classifier systems and their application to handwritten word recognition. Pattern Recognition Letters 25(11), 1323-1336.
[12] Thangavel, K., Jaganathan, P.P. and Easmi,P.O. (2006). Data Mining Approach to Cervical Cancer Patients Analysis Using Clustering Technique. Asian Journal of Information Technology (5) 4, 413-417.
[13] Hongmei Yan, Jun Zheng, Yingtao Jiang. (2008). Selecting critical clinical features for heart diseases diagnosis with a real-coded genetic algorithm, Applied soft computing, Vol.8, pp.1105-1111.
[14] R.E. Abdel-Aal. (2005) GMDH-based feature ranking and selection for improved classification of medical data, Journal of Biomedical Informatics, Vol.38, pp.456-468.
[15] Yan Wang, Lizhuang Ma, and Ping Liu. (2009). Feature selection and syndrome prediction for liver cirrhosis in traditional Chinese medicine. Comput. Methods Prog. Biomed. 95, 3, pp 249-257.
[16] Polat, K., Tosun, S., &Günes, S. (2006). Diagnosis of heart disease using artificial immune recognition system and fuzzy weighted preprocessing. Pattern Recognition, 39(11), pp.2186–2193.
[17] JacekJelonek, Krzysztof Krawiec, Roman Słowi´nski, Jerzy Stefanowski, and JanuszSzyma´s. (1993). Neural networks and rough sets: Comparison and combination for classification of histological pictures. Proc. International Workshop on Rough Sets and Knowledge Discovery. pp. 426–433.
[18] Polat, K., Gunes, S., &Aslan, A. (2008). A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. Expert Systems with Applications, 34(1), pp. 214–221.
[19] J.Bernatavièiene, G. Dzemyda, O. Kurasova, V.Barzdžiukas, D. Buteikiene, A. Paunksnis.(2008) Rule Induction For Ophthalmological Data Classification. Proc. of 20th EURO Mini Conf. Continuous Optimization and Knowledge-Based Technologies. pp. 328-334.
[20] Cheng, T.H., Wei, C.P., Tseng, V.S. (2006). Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches. Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06).
[21] L. Li, H. Tang, Z. Wu, J. Gong, M. Gruidl, J.Zou, M. Tockman, and R. Clark. 2004. Data mining techniques for cancer detection using serum proteomic profiling. Artificial Intelligence in Medicine, Vol. 32, No. 2, pp. 71–83.
[22] C.T. Su, C. H. Yang, K. H. Hsu, and W. K. Chiu. (2006). Data mining for the diagnosis of type II diabetes from three- dimensional body surface anthropometrical scanning data. Computers & Mathematics with Applications, Vol. 51, No. 6–7, pp. 1075–1092, 2006.
[23] S.Piramuthu. (2004). Evaluating feature selection methods for learning in data mining applications. European Journal of Operational Research. Vol.156, Issue 2, pp.483-494.
[24] Sarojini Balakrishnan, Ramaraj Narayanasamy, Nickolas Savarimuthu, Rita Samikkannu.(2008). SVM Ranking with Backward Search for Feature Selection in Type II Diabetes Databases. Proc. IEEE International Conference on Systems, Man and Cybernetics, SMC 2008. pp. 2628 – 2633. DOI: 10.1109/ICSMC.2008.4811692.
[25] Sarojini Balakrishnan, Ramaraj Narayanasamy, Nickolas Savarimuthu.(2009). Enhancing the performance of LibSVM Classifier by Kernel F-score Feature Selection. S.Ranka et al. (Eds.) : IC3 2009, CCIS 40, pp.533-543. Springer-Verlag Berlin Heidelberg 2009. ISSN: 1865-0929.
[26] Sarojini Balakrishnan, Ramaraj Narayanaswamy. (2010). A Hybrid Prediction Model with F-score Feature Selection for Type II diabetes Databases. Proc. 1st ACM-W Celebration on Women in Computing in India, A2CWiC 2010. ACM, New York, NY,2010. 1-4. ISBN: 978-1-4503-0194-7.
[27] C.L.Blake, C.J. Merz, UCI repository of machine learning databases, Website: http://www.ics.uci.edu/~mlearn/MLRepository.html, 1998.
[28] Keller J.M., Gray M.R. and Givens J.A.(1985). A Fuzzy K-Nearest Neighbor Algorithm. IEEE Transactions on Systems, Man and Cybernetics, 15(4), 1985.
[29] Bian, H., Mazlack, L.(2003). Fuzzy-Rough Nearest-Neighbor Classification Approach. In: Proceeding of the 22nd International Conference of the North American Fuzzy Information Processing Society (NAFIPS), pp. 500–505.
[30] Witten, I., Frank E. "Data Mining: Practical Machine Learning Tools with Java Implementations", Morgan Kaufmann, San Francisco, 2000.