Fused Features Classification for the Effective Prediction of Chronic Kidney Disease

Mohammed Siyad B
Assistant Professor
Department of Computer Science and Engineering
TKM College of Engineering Kollam, Kerala

Manoj M
Assistant Professor
Department of Computer Science and Engineering
UKF College of Engineering and Technology Kollam, Kerala

Abstract

The paper presents an application of data mining for improving the accuracy of prediction of a disease state by selecting the most relevant features associated with it. The experiments are performed on chronic kidney disease (CKD) data. The basic idea in this study is that use of a number of methods instead of a single one increases the probability of selecting features which are more closely related to the disease. Multiple feature selection methods have been applied independently on the CKD data set and the results integrated into a final optimal set of features. These data have been applied to the classifiers to identify CKD from reference cases. Various classification methods are compared to select the best model over 10-fold cross-validation in the training data set. Random Forest classifier is chosen as the best model with superior performance.

Keywords: Chronic kidney disease, Classification, Feature selection, Fusion, Prediction

I. INTRODUCTION

Chronic kidney disease (CKD) is a condition where the function of the kidney gradually loses over time. If the disease gets worse, the level of wastes in the blood increases to a great extent and that makes us feel sick. It will lead to complications like high blood pressure, anemia, weak bones, heart problems and blood vessel disease. Finally it may eventually lead to kidney failure and requires dialysis or kidney transplantation to maintain life. The major causes of CKD may be diabetes, high blood pressure and other disorders. Early detection and treatment can prevent chronic kidney disease from getting worse[6]

The end stage of the CKD, known as end-stage renal disease(ESRD) is a permanent kidney failure and cannot live without dialysis or kidney trans-plantation. The estimated glomerular filtration rate (eGFR) is an important bio-marker for the CKD, where eGFR < 60mL=min=1:73m. In ESRD (or stage 5 CKD), eGFR < 15mL=min=1:73m and it is the severe form of CKD. In 2010, 50965 adult patients in UK received renal replacement therapy reflecting a UK ESRD prevalence of 832 per million population and a 63% increase in renal replacement therapy population over the past decade [4]. Thus there is an urgent need to identify the bio-markers that help to find individuals who are at high risk of developing CKD so that proper treatments can be applied to prevent from getting worse.

Developments in data mining in recent years have thrown lights into the discovery of attributes which are the best associated with a particular disease. Although there have been a lot of works attempted in this arena, there still need better understanding and improvements in the research, particularly in the medical area. To such examples belongs to CKD data. In most of the cases, particular methods are applied and one of them is selected as the best and considered as the appropriate method for attribute selection.

This paper mainly concerned with the selection of attributes which are more closely associated with the disease. Applying a classifier on the selected data should lead to the improved accuracy in identifying CKD and reference (non-CKD) cases. In the experiments, a ranking method is applied on the attributes. Different selection methods may provide different results for the same data set. These features can be fused together to obtain the final set of features. This fusion based feature selection is the main contribution of the paper and it will increase the probability of getting more relevant features. The trained classifier system may then be used to predict the CKD or non- CKD class of the newly acquired data.

The rest of this paper is organized as follows. Section II reviews the various researches that have done in this area. Section III describes the feature selection methods applied in this study. The detailed algorithm for the proposed work in given in section IV. Section V describes the experimental set up including the data set chosen for the illustration of idea. Section VI analyzes the results of the experiment. Section VII concludes the study with a remark on the future enhancements that can be made.

II. RELATED WORKS

Recent studies conveys that several data mining techniques can be applied in the arena of disease diagnosis and prognosis. Identifying the most suitable attributes which can lead to the proper recognition of the disease is a frequent issue that researchers would face. In this regard, a few computational works have been done in the recent past.

Jefery Li et al.[2] examined a large set of CpG methylation data, histone modification data and genome data for predicting the differential expression of RNA-Seq transcriptome. The methylated features in the promoter region are more important in the prediction of gene expression.

Tomasz Latkowski et al.[3] proposed a two stage gene selection approach and classification. Using the set of methods instead of single one will increase the probability of finding the globally optimal set of genes which are the best associated with the particular disease.

Shoon Lei Win et al.[10] proposed a machine learning approach called the averaged one dependence estimator with subsumption resolution (AODEsr) for achieving more accuracy in prediction. The proposed method is a semi Naive Bayesian approach that retains the strengths of Naive-Bayes while the errors are minimized by reducing the attribute independence.

In most of the cases, different methods have been tried but the best method is treated as the final solution. Different feature selection method may produce different feature sets. So the accuracy of prediction varies based on the method applied. If we fuse the results of a number of selection methods, we may able to achieve features that are most closely related to the disease.

III. APPLIED FEATURE SELECTION METHODS

Feature selection is a commonly employed dimension reduction technique in machine learning. It aims to select a small subset of the relevant features from the original set based on certain evaluation criterion. Feature selection leads to better learning performance, lower computational cost, and better model interpretability[1]. Here a number of feature selection methods are applied independently and the results are integrated into a single final feature set. "Using the set of methods instead of single one will increase the probability of finding the globally optimal set of genes which are the best associated with the particular disease"[3]. The following feature selection methods were applied in this study.

A. Information Gain

It measures the number of bits of information obtained for category prediction by knowing the presence or absence of a term [9]. Let Ci (for i=1,2,...,m) be the set of categories in the target space. The information gain of the term t is defined to be

$$G(t) = -\sum_{i=1}^{m} P(C_i)\log P(C_i) + P(t)\sum_{i=1}^{m} P(C_i|t)\log P(C_i|t) + P(\bar{t})\sum_{i=1}^{m} P(C_i|\bar{t})\log P(C_i|\bar{t})$$

B. Gain Ratio

Gain ratio (GR) is a modification of the information gain that reduces its bias. When choosing an attribute, Gain ratio considers the number and size of branches. It takes the intrinsic information of a split into account. Intrinsic information is entropy of distribution of instances into branches (i.e. how much info do we need to tell which branch an instance belongs to). Attribute value decreases as the increase in the intrinsic information [5].

$$\text{Gain Ratio(Attribute)} = \frac{\text{Gain(Attribute)}}{\text{Intrinsic - Info(Attribute)}}$$

C. One R(OR) Attribute Evaluation

This is one of the most primitive schemes in which one rule is produced for each attribute in the training data and the rule with the smallest error is selected. It treats all numerically valued features as continuous and uses a straightforward method to divide the range of values into several disjoint intervals. Missing values are taken as a reasonable value. It produces simple rules based on a single feature. It can be useful for determining a baseline performance as a benchmark for other learning schemes [7].

D. ReliefF

ReliefF is a simple yet powerful procedure to estimate the quality of attributes in problems with strong attribute dependencies. For a random instance Im from class L, this method finds the k nearest neighbors from the same class (hits, H) and k nearest neighbors from each of the other classes (misses, M) and updates the quality estimation Wi for ith attribute based on their values for Im, H and M. For the ith attribute, if the instance Im and those in H have different values, then Wi decreases. Similarly if Im and those in M have different values, then Wi increases. This procedure is repeated n times [3].

E. Fusion of the selection methods

To identify the most relevant features, a fusion based the feature selection approach is used. The above feature selection methods are applied independently on the data set in association with a ranking method. The position of each attribute is noted after each run. A global weight w is assigned for each attribute f and is obtained by

$$w(f) = \sum_{i=1}^{K}\sum_{j=1}^{N_r} w_{ij}(f)$$

where, K is the number of applied selection methods, Nr indicates the number of repeated runs of selection and wij is the position of features in i^{th} method of selection and j^{th} run.

IV. ALGORITHM

The proposed algorithm takes N_r , K and an n-dimensional vector D as inputs. It returns a subset of features in D , as output. Vector R_i contains randomly chosen rows from D on which the the feature selection method Fi is applied. This set of final features are used to train the classifiers which produce the class label of the samples.
F ={Information Gain, Gain Ratio, One R(OR) Attribute Evaluation, ReliefF}

A. Algorithm 1. Fused Feature Selection algorithm

Inputs: Vector D =< f1,f2, f3,.... fn >, Nr and K
Output: A subset of features < f1, f2,.... fn >
1) procedure Select_Fused_Features(D, N_r, K)
2) begin:
3) Initialize:
4) w(f)=0, i=1, j=
5) repeat for j
6) repeat for i
7) 6: R_i = random rows from D
8) Apply method F_i on R_i.
9) Store the position of each feature.
10) I = i + 1
11) until i = K
12) j = j + 1
13) until j = N_r
14) 13: Set i=1, j =1
15) for every feature f do
16) repeat for i
17) repeat for j
18) w(f) = w(f) + w_{ij}(f)
19) j = j + 1
20) until j = N_r
21) i = i + 1
22) until i = K
23) end for
24) Arrange the features in the decreasing order of w(f).
25) Return a set of least significant features.
26) end
27) end procedure

V. EXPERIMENTAL SETUP

A. Data Set

The numerical experiments have been done on the data set of CKD. The data set is downloaded from UCI Machine Learning Repository [8]. There are 400 instances and 25 attributes. The database consists of two classes: the first one is related to persons with CKD (number of such observations n = 250) and the second to the control group of healthy children (n = 150). The samples are collected from the hospital nearly 2 months of period.

B. Main stages of Experiments

This section describes the numerical experiments of attribute selection for the CKD data. In the first stage, four feature selection methods were applied independently on the data set to identify the attributes and their order. The attributes are ranked using a ranker search method that ranks the attributes by their individual evaluations, based on 10-fold cross validation. The position of the attribute in each run is noted. Then a global weight value w(f) for each feature f is calculated as the sum of the positions of those features in each application of the selection method. Then these w(f) values are arranged in the decreasing order. The least valued attributes were chosen as the final feature set.

In the next stage, these features are used to train the classifiers using 10-fold cross-validation on the training data set. Here four classification strategies were applied and their performance was compared. The applied classification techniques were Naive Bayes, Random Forest, J48 classifier and Logistic Regression.

VI. RESULTS AND DISCUSSION

The main contribution of this paper is the fusion based feature selection for accurate prediction. The ranker searching method ranked the features based on their relevance. Four selection methods are applied independently as explained in the above section.

Even though the contents of the selected feature sets were different in different methods, few of them produced a large percentage of same attributes. In most of the cases the features HEMO, SG, PCV, HTN, SC, DM were obtained higher rank compared to others. This implies that these features have an important role in determining the disease state. Also the features BA, SU, CAD, AGE etc were less relevant.

Four classification techniques (NB - Naive Bayes, RF - Random Forest, J48 classifier and LR - Logistic Regression) were applied on the fused feature set. The results of the comparison are summarized in Table 1.

Table - 1
Summary of the Comparison

Characteristic	*Method*			
	NB	*RF*	*J48*	*LR*
Mean Absolute Error	*0.0278*	*0.0377*	*0.0435*	*0.0194*
RMS Error	*0.01426*	*0.0848*	*0.01204*	*0.137*
Relative Absolute Error	*5.93%*	*8.038%*	*9.267%*	*4.144%*
Root Relative Squared Error	*29.45%*	*17.50%*	*24.86%*	*28.29%*
Coverage of cases	*99%*	*100%*	*99.75%*	*98.25%*

Among the four classification techniques that considered, the Random Forest show superior performance with an accuracy of 99.75%. The performance accuracies of the four classification methods are listed in Table 2. The accuracy chart is also presented in the Fig. 1.

Table 2
Accuracy of the classification

Method	*Correctly classified*	*Incorrectly classified*	*Accuracy*
NB	*390*	*10*	*97.5%*
RF	*399*	*1*	*99.75%*
J48	*392*	*8*	*98%*
LR	*392*	*8*	*98%*

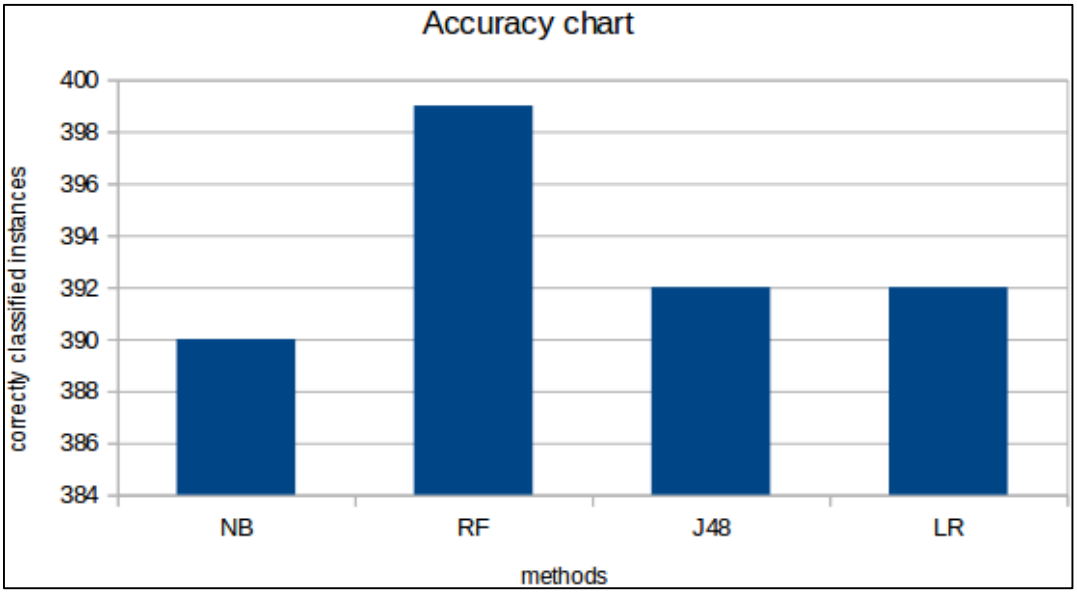

Fig. 1: Accuracy chart

VII. CONCLUSION

Several data mining techniques have been used in this paper for selecting the most relevant features associated with CKD. A two stage feature selection approach is examined here. In the first step, four attribute selection methods have been applied independently and then the results were fused together to obtain the final set in the second step. The selected features have been used as the input attributes for the classifiers, which were responsible for recognition of CKD cases from the reference ones.

The main contribution of the paper in the research of CKD is the fusion of many selection methods with different principle of operation into final system for the accurate prediction of the disease. Also a special approach for accessing the quality of attributes have been applied through ranking them during selection process.

Further, the idea presented in this paper needs to be continued in few directions. The main point is that more selection methods can be used in the fusion for getting more accurate features. Also more efficient alternative methods can be applied for fusing the independently selected features. Application of genetic algorithms was successful in selecting the optimal set of feature in some other recognition problems[3]. Another enhancement can be in ranking of attributes where they can be ranked based on the frequency of appearance in each run.

REFERENCES

[1] M. Dash and H. Liu. "Feature selection for classification". Intelligent Data Analysis(Elsavier), 1(1-4):131156, Marche 1997.
[2] Jeffery Li , Travers Ching , Sijia Huang and Lana X Garmire. D. "using epigenomics data to predict gene expression in lung cancer". BMC Bioinformatics, 16(S2):1471{2105, March 2015.
[3] Tomasz Latkowski and Stanislaw Osowski. "Data mining for feature selection in gene expression autism data". Expert Systems with Applications(Elsavier), 42:864{872, February 2015.
[4] Laura J Smyth , Gareth J McKay , Alexander P Maxwell and Amy Jayne McKnight. "DNA hypermethylation and DNA hypomethylation is present at different loci in chronic kidney disease". Epigenetics, 9(3):366{376, March 2014.
[5] R. Praveena Priyadarsini , M.L.Valarmathi and S. Sivakumari. "Gain ratio based feature selection method for privacy preservation". ICTACT Journal On Soft Computing, 1(4):201{205, April 2011.
[6] National Kidney Foundation. "About chronic kidney disease". https://www.kidney.org/kidneydisease/aboutckd `.
[7] Jasmina Novakovi , Perica Strbac and Dusan Bulatovi. "Toward optimal feature selection using ranking methods and classification algorithms". Yu-goslav Journal of Operations Research, 21(1):119{135, March 2011.
[8] UCI Machine Learning Repository. "Chronic kidney disease data set". https://archive.ics.uci.edu/ml/datasets/Chronic Kidney Disease.
[9] J. O. Pedersen Y. Yang. "A comparative study on feature selection in text categorization". In Proc. Of the 14th Intl Conf. on Machine Learning (ICML97), pages 412{420, New York, NY, USA, 1997. Morgan Kaufmann Publishers,.
[10] Shoon Lei Win , Zaw Zaw Htike , Faridah Yusof and Ibrahim A. Noor-batcha. "Gene expression mining for predicting survivability of patients in early\