Ensemble Learning with Variable Selection to Predict the Diseases with Improved Accuracy

Shahebaz Ahmed Khan[1], Dr Santosh Kumar Yadav[2]

Research Scholar, Dept of CSE, Shri Jagdishprasad Jhabarmal Tibrewala University, Jhunjhunu, Rajasthan, India[1]

Research Dean and Professor, Shri Jagdishpr asad Jhabarmal Tibrewala University, Jhunjhunu, Rajasthan, India[2]

ABSTRACT: The solvers in Data Mining are intelligent enough to provide meaningful information for decision making. Data Mining is capable of recommending the health industry to take necessary measures in consideration with disease diagnosis. Disease prediction can be easy and efficient, if we use Data Mining techniques. In this research paper, Ensemble learning is implemented to find the disorders and other diseases in diabetic patients. The results of the clinical analysis using the Ensemble models with feature selection have been discussed in this paper. Ensemble models have the ability to incorporate other classifiers to improve performance of solvers with efficiency. The results of this implementation are further compared to the simple classification techniques and methods. The Ensemble Learning used in our work has combined some effective classification models to show the accuracy gain compared to the traditional methods. The idea of this work is to design a valid decision support source using the methods mentioned.

KEYWORDS: Ensemble Learning, solvers, feature selection, classifier, diabetic, decision support source.

I. INTRODUCTION

The Data Mining theory can be useful in demands of data analysis, prediction and estimation. The mining and extraction concepts are widely applied and used in medical domains for disease data classification and prediction. Some large amounts of clinical data is used to know the hidden information and knowledge with application of Data Mining algorithms to help the HCI. The recent trends of Data Mining concepts make use of advanced methods, well formed and intelligent algorithms in medical domain to achieve more and more transparent results. The results obtained by various Data Mining techniques like Classification, Clustering, Association Mining, Associative Classification etc have proved that, those are the trusted outcomes with accuracy and efficiency which can help the medical domain for better and precautionary steps in treatment of diseases [1].

Many research works on the health issues have mentioned where the point for the need of Data Mining and Machine Learning are mentioned as a necessary factor to solve and reduce the obstacles that are found in the disease diagnosis and treatment. Data Mining has been a supportive concept for medical domain to diagnose and predict complex diseases using medical data. Data Extraction or Mining serves as a suitable tool that can serve the needs of heath industry in predicting diseases and interpreting hidden patterns.

Diabetes is a major health problem all over the world now-a-days. Health care expenses and expenditure are high for problems related to the syndrome of diabetes Mellitus. Diabetes can lead to very severe and serious health complications like brain strokes, cardiac arrest, short of breath, poor eye vision, liver failure etc. This work tends to predict the chances of occurrence of other major health issues in diabetic patients using various Data Mining techniques like Classification and Ensemble Learning. The potential benefits of Ensemble Learning are combined with feature subset selection to improve the accuracy and prediction rate. Various Classification algorithms like Random Forest, Naïve Bayes, ID3, Multilayer Perceptron etc used in this research to compare the results with ensemble model. The experiments were done for data classification using Ensemble Learning model with feature subset selection.

II. RELATED WORK

In a research, Han et al [2] have used the Decision tree algorithms and ID3 for the classification of diabetes data sets in order to derive an effective predictive model. The model gave an accuracy rate of eighty and seventy two percentages respectively. Various Data Mining concepts like Neural Networks were also used to interpret and estimate factors of risk in diabetes index process and it was found that in this context, many of the indices were additive in nature [3]. In some works on disease prediction using Data Mining techniques, Apriori and FP Growth algorithms of Association Rule Mining technique were used for the prognosis of diabetes syndrome. Chang-Shing Lee and Mei-Hui [4] Wang proposed a semantic system for decision making that can be used for diabetic data, these researchers used fuzzy expert systems with a 5 layer approach that can extract the information from the uncertain data sets. Wenxin Zhu and Ping Zhong [5] tried to predict the unknown and meaningful patterns using the model of one class Support Vector Machines. The objective of this work was to classify the data based on a single class because of the limited number of diabetes syndrome data sets of training. Here, in this advanced SVM model called SVM+ technique, the accuracy of classifiers was improved. This was done by embedding the optimization methods using the technique of SVM + model. This method gave a good performance of prediction than earlier methods. A Data Mining system which used nearly 30 Machine Learning algorithms was capable of providing a highly effective decision support system for diabetes data prognosis. This system was made effective with a combination of new approach named as rotation forest [6]. The techniques of bagging and boosting of Ensemble Learning used by the researchers have shown a way to take the criteria of disease prediction of diabetes into a well mannered and organized way which gave desired outcomes as results for future trends and data analysis.

III. PROPOSED METHOD AND TECHNIQUE

The work was intended to predict the presence of Co-diseases which means the diseases other than the original disease (here it is diabetes) in diabetic patients. The methodology selected was Ensemble Learning. It is one of the powerful mining techniques which can be applicable for data classification. This technique can be used to improve the accuracy of Classification technique in Data Mining. Even, this learning model has another advantage like it can reduce bias and over fitting towards a specific class. Ensembles are considered as meta-solvers which are a combination of various Machine Learning techniques to give one predictive model that can reduce bias (boosting), decrease variance (bagging), or improve predictions (stacking).

The Ensemble Learning model was built with a combination of Naïve Bayes and Random Forest algorithm form the single model. These two classification models were selected and ensemble for the results due to the reason that, in the previous works these two models of classification have quite good better accuracy than the other used models of Classification in Data Mining. Besides these two classification models, ID3, Multilayer Perceptron, Multi Class classifier were used for classification, but these proved to be less accurate in terms of prediction of the selected data set. To improve the performance of Classification using Ensemble Learning, our work has used the meta vote classifier with a combination of majority voting rule.

The raw diabetic data sets were used for training the data, which had 12 attributes in count. The Ensemble model was executed based on the feature subset selection which is also sometimes called variable selection. The proposed model used the K-folds cross validation to form the classifiers using Ensemble Learning. The aggregation of Random Forest and Naive Bayes algorithms for Ensemble approach was forced to give the results of improved accuracy than in comparison classification models.

The basic criteria selected to predict and conclude the presence of some kind of diseases like heart problems, poor vision, brain strokes etc is as follows.

A class label was defined as Co-disease based on the condition as

If,

Diabetic >= 110, HDL Cholesterol >=60 as high (for male) and >= 70 as high (for female), LDL Cholesterol >130 as bad and <= 130 good, VLDL Cholesterol > 30 as high and <= 10 as low, TTL Cholesterol > 5.0 = high, >= 8.0 = very high, range between 4.5 and 5.0 as border line, Total cholesterol < 200 as normal, range between 200 and 230 as border line, > 250 as very high, Cholesterol = 130 as normal, >130 as high and > =250 as very high.

Then,

Co-disease = yes.

This means if the above condition is true then, the cardiac problems can arise as a prediction outcome.

The below given figure 1 shows the flow of proposed model of disease prediction.

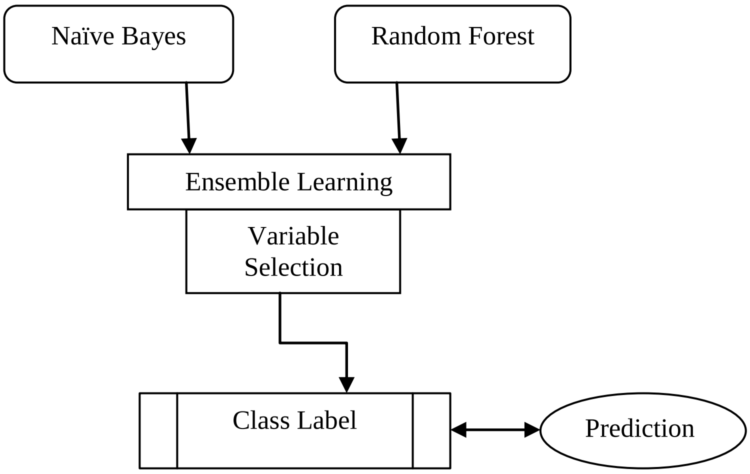

Figure 1: Proposed Model of Classification

IV. Experimental Results

The total data instances were more than 114 and attributes were 12.The experiment was carried out using 10 folds cross validation technique and the model was build in 0.02 seconds of time. The model combined algorithms of Naïve Bayes and Random Forest classifiers to form an efficient Ensemble Learner with majority voting combination rule. The accuracy obtained with this model 94.7368 percentage. Anyhow, without feature subset selection, an experiment was done which showed the below results of accuracy as shown in confusion matrix. The confusion matrix of the method is as follows in table 1.

A	**b**	**Classified as**
77	2	a = no
4	31	b = yes

Table 1: Confusion Matrix for Ensemble Method (without subset selection)

The experiments performed using feature subset selection criteria gave the improved accuracy than normal Ensemble Learning without subset selection. This is a process of automatic selection of variables of the data which are considered to be most relevant to the predictive modeling problem on which the users pay attention. The selected features to improve the rate of accuracy of Classification were age, gender and these two features showed impact on the other attributes of data set. The accuracy gained with this idea of Ensemble method was 99.1228 percent. The confusion matrix for this is as follows in table 2.

A	**b**	**Classified as**
77	0	a = no
1	34	b = yes

Table 2: Confusion Matrix for Ensemble Method (with subset selection)

The other accuracy details are as follows in table 3.

TP Rate	1	0.971	**Weighted Average**
FP Rate	0.029	0	0.991
Precision	0.988	1	0.02
Recall	1	0.971	0.991
F-Measure	0.994	0.986	0.991
ROC Area	0.986	0.986	0.991
Class	No	Yes	0.986

Table 3: Accuracy Details of Proposed Technique

Results with Naïve Bayes and Random Forest are found to be less accurate than proposed model. The results in comparison with Ensemble Learning method have clearly shown that, Ensemble models can provide better form of accuracy gain and prediction rate. The Classification accuracy with Naïve Bayes was 93.8596 percent and with Random Forest model was 92.982. The confusion matrix of these two classifiers is given below in table 4.

Classifier	**A**	**B**	**Classified as**
Random Forest	77	2	a = no
	6	29	
Naïve Bayes	75	4	b = yes
	3	32	

Table 4: Confusion Matrix for NB and RF

The results of our Classification of diabetic data set have shown that the presence of other diseases and heath issues like cardiac problems and poor vision are more common in diabetic patients. The majority population of diabetic syndrome is prone to get heart diseases in future as per our work and experiments. Based on probability distribution it was found that, merely 42 to 45 percent of diabetic patients can have the chances to get cardiac problems. The prediction accuracy of this proposed method has also shown that, there was a difference of 2 to 3 percent in result prediction of presence of diseases in diabetic data.

V. CONCLUSION AND FUTURE WORK

The results of our work have shown that, the proposed method of Ensemble Learning can be effective for prediction and data Classification. Health sector can make use of the included methods of Data Mining for decision making. The accuracy of Classification can be improved to some greater extent, which or where can be a recommendation to the medical field to design a decision support source for medical diagnosis and cost effective treatment. At the same time, we can also say that possibilities of cardiac diseases in diabetic patients are danger alarms, which should be paid at most care by medical professional and diabetic patients. It can also be concluded once again that, Data Mining possesses a capability to provide useful data patterns to medical industry.

REFERENCES

1. Milovic, B. (2011). Usage of Data Mining in Making Business Decision. YU Info 2012 & ICIST 2012, (pp. 153-157)..
2. Han, J., Rodriguze, J.C ., Beheshti, m., " Diabetes data analysis and prediction model discovery using rapidminer", Second International Conference on Future Generation Communication and Networking.96-9 (2008) .
3. Juliyet, L. Cynthiya, and Mr K. Mohamed Amanullah. "The Surveillance on Diabetes Diagnosis Using Data Mining Techniques.
4. Lee, Chang-Shing, and Mei-Hui Wang. "A fuzzy expert system for diabetes decision support application." IEEE Transactions on Systems, Man, and Cybernetics, Part B a(Cybernetics) 41.1 (2011): 139-153.
5. Zhu, Wenxin, and Ping Zhong. "A new one-class SVM based on hidden information." Knowledge-Based Systems 60 (2014): 35-43.
6. Ogunyemi O., Kermah D. Machine learning approaches for detecting diabetic retinopathy from clinical and public health records. AMIA Annu Symp Proc. Nov 5 2015;2015:983–990.