International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Co- Disease prediction in Diabetic Patients using Ensemble learning for Decision Support System

Shahebaz Ahmed Khan, M A Jabbar

Abstract: ***The methods of classification that are available in the data mining concepts along with Ensemble methods of data prediction in data mining and machine learning gradually helps to predict the data for the by building the various classification models for future analysis in a better as well as accurate way. The Ensemble learning method algorithms can be used to build the classifiers by taking the weighted vote of the classifiers in order to construct the new data predictions and points. Two or more different data models are taken into consideration for running the process to predict the results in Ensemble Prediction System. In this paper, the the research work carried out by us on diabetic medical data using various classification models like Naive Bayes, Random Forest, Zero R etc. are compared and analyzed with the Ensemble prediction models to prove the efficiency of the used method so as to predict the diabetic syndrome possibility in the patients of various health symptoms. The algorithm used for voting and their uses as well as application on such data to predict the diseases is discussed. The rules developed in this work can be helpful to predict and find the co-disease in the patients of diabetes for decision making and these rules developed have been then ranked according to the final classifier for better form of the disease prediction. The classification methods that are proposed can not only effectively but also can accurately predict the datasets in the various context of disease analysis by improving the accuracy of the classifiers.***

Keywords: ***Co-disease, Ensemble Prediction, Zero R, classification methods, Naive Bayes and Random Forest, Diabetes and Associative Classification.***

I. INTRODUCTION

Data mining is to analyze the data from various perspectives by applying intelligent methods and combining it by summarizing it into useful information by considering the associations, patterns and relationships among all the given data or data sets and information. The process of discovering new patterns, trends with meaningful correlations is referred as data mining and this is by transforming large amounts of data which is stored and found in data repositories, using various mathematical, statistical and pattern recognition technologies[9]. Data is generally considered as some form of fact, or any fact, text or numbers that can be easily and accurately processed by a computer. Data mining software is now considered as the most influential tools used for the purpose of analyzing data from different corners [8]. Data mining allows users to analyze data from different dimensions as well as many angles, allows the users for categorization and summarization of the identified relationships. Today, we can see that a number of organizations are accumulating vast and growing amounts of data which is of different formats and databases. In machine learning and data mining, Ensemble methods are treated as powerful and influential components. The Ensemble methods combine multiple models into one model usually which is considered to be the more accurate than the normal and simple models for analysis. Ensembles in data mining are mostly used for marking the timings to drug discovery[10], disease prediction and fraud detection to recommendation systems due to its vital functionality in accurate prediction than its normal model interpretability. Ensembles are useful with all modeling algorithms, our work focuses on Naive Bayes and Random Forest methods for voting the Ensemble System of prediction. The Ensemble rules derived in our work focuses on comparison and analysis of the previous work in the context of the diabetic disease prediction. It reveals classical ensemble methods like bagging, random forests, Naive Bayes etc to be special cases of a single algorithm, in order to show how to improve their accuracy and speed. Generally in an ensemble method, a learning algorithm s used and then it is combined with the predictions of other models so that the efficiency and robustness can be improved within a single model scenario. In this Ensemble method bagging is used for improving classification schemes and unstable predictions. In our research work, the prediction of the future diseases and also the symptoms of those diseases are predicted with the classification model like Naive Bayes and Random Forest. This is carried out so that the possibility for the prediction of the co-disease can be made in the diabetic patients. Any disease found or any disease that can occur along with the already existing or occurred disease in a patient can be considered as a Co- disease. In order to find and conclude the Co-disease detection, both the syndromes and other diseases are taken into count.. The classification models for ensemble prediction system used in this research work can easily explain the nature of the data for decision support systems. The work is also compared to the previously predicted data by us so as to maintain the accuracy and to improve the classifiers accuracy for better estimation and prediction. Experimental results in our work have demonstrated and shown that the ensemble model of data mining can be a superior approach in the context of providing high performance and predictive accuracy for disease diagnosis

Manuscript published on 30 September 2019
* Correspondence Author
Shahebaz Ahmed Khan*, Research Scholar, Department of CSE, JJTU, Jhunjhunu, Rajasthan, India. sakklkyr@gmail.com
M A Jabbar, Professor at Department of CSE, Vardhaman College of Engineering, Hyderabad, India. jabbar.meerja@gmail.com

II. CLASSIFICATION IN DATA MINING

The aim of classification in data mining is to derive and to produce an abstract model with set of classes called classifiers. The classifiers form the set of training set and labeled data, [5]. The class label is known to approximately classify and group the given dataset. To build the accurate classifiers various methods have been proposed and some of these said methods like Naive Bayes, SVM, Decision trees etc. can be used. The combination of association rules in data mining and the classification techniques is considered as a associative classification. It is recently a big trend and need in the various situations of data prediction in data mining tasks. From a given set of training data, the Bayesian networks are used to derive and construct the classifiers with class labels and this is nothing but classification. Classification techniques in data mining theories can be used to classify the data without any complex task involving, at the same with accuracy [2]. Preprocessing method is used for cleansing the data in order to gain effective classification.

III. ENSEMBLE LEARNING

In predictive modeling and data analytics, a classification technique which is of single model based when applied on data sample can have the possibility of biases; can be found with outright inaccuracies that can cause variations highly and can affect the actual reliability of the deep analytical findings of data. To maintain and improve the accuracy one idea in data mining is to combine the various models for prediction in order to reduce the limitations and risks that are met during the prediction. In Ensemble method model, two or more analytical models which are of different or related classification models are combined to run them. Then the results are synthesized after running into a single score system so that the accuracy of the predictive analytics can be improved.

In Ensemble system using the same learning algorithms multiple models are trained and also it is a machine learning concept [11]. We use bagging technique to reduce and lower the variance in the process of prediction; this is possible by selecting and generating some sort of extra data from dataset available for the purpose of training, This is done by using different permutations and combinations to produce multi-sets of the actual data. Ensemble prediction method is a well-proved and efficient method which is used in works of research and deep studies for obtaining highly accurate data classification. This is done by hybridizing and mixing various classifiers. Due to the combination of various classifier models in Ensemble approach an enhanced prediction and improved performance is a good and favorable in-built characteristics A vote-based classifier which is weighted one is used for ensemble here, which will overcome the drawbacks of conventional methods available in data mining. The ensemble or combination of two important well accurate heterogeneous classifiers like Random forest and Naive Bayes are used to improve the accuracy in Co- disease prediction criteria. In health and medical industry, the predictive tasks to model the solutions for Co-disease prediction in diabetic patients are risk estimation and also challenging trouble. The attempt which is made to screen the health care databases clinically for accurate modeling can be done using Ensemble Prediction approaches [12]. In our work, it is proposed that Ensemble model approach can be for better prediction accuracy, we have combined the predictive ability of various classifier models. In this approach, ensemble learning combines two highly accurate classifier model approaches which include Naive Bayes and random forest algorithms to estimate and diagnose occurrence of other additional diseases in the diabetic patients.

IV. DIABETES AND CO-DISEASES IN RECENT SCENARIO

Many people consider diabetes as a disease but it is a syndrome in where blood glucose or levels of sugar in the human body are too high. As per the medical diagnosis any diabetic human body can become a victim of type 1 diabetes or diabetes of type 2. Type 1 diabetes the pancreas produces no insulin in the human body The problems with diabetic disorder and syndrome are very severe which can cause psychological issues like dementia, , dental diseases, neurological problems, vision complications, cardiac attacks, thrombocytopenia, and brain strokes etc. When there is more than enough levels of glucose that too much quantity of glucose levels in human body then it can lead towards unexpected and serious health problems. Due to this, there is a finite probability of damaging the vision, excretion system, nervous system and also can create cardiac arrests by causing the severe heart pain. Moreover, diabetes can be considered as a metabolic disorder and due to this disorder the symptoms of polyuria, polydipsia etc. are found to be the common problems. More or less, 10 percentage of the total population is found with type 1diabetes and 90 % of the people are victims of type 2 diabetes. Type one diabetes is a chronic disorder. It takes the roots of health to become vulnerable with other complicated and dangerous diseases.. In this regards, this makes a need to find the By-diseases at the very first stage , that occur with diabetes[7]. To safeguard the health and to avoid the health problems the need and necessity for prediction of diabetic disease and its co diseases arises.. Ii is estimated that there are more than 400 million people who suffer with diabetic syndrome all over the world.

Large scale of research and expertise work is being done for medical diagnosis of diabetic disease and heart diseases. There were good results in predicting the early heart diseases using various data mining algorithms [1]. In our work, an attempt is made to predict probability of heart problems in the people who suffer with diabetic syndrome using Ensemble prediction methods by applying the Naive Bayes and Random Forest algorithm classification. All the previous work which was based on the Zero R and simple Naïve Bayes in the same context with future subset selection. Feature subset selection in data mining is an active , interesting form of enhancing the results for decades in expert systems, data mining , pattern matching systems. Sellappan Palanniappan has proposed Intelligent Heart Disease Prediction System by using Naïve bayes and neural networks algorithms [11][6]. The present work and the previous work is compared and then analyzed for the co-disease prediction.
Here, the results are compared to prove the reliability as well as accuracy of the ensemble prediction system in the picture process of the By- disease prediction in people suffering with diabetes.

The idea of the heart disease classification and prediction using the approach of associative classification for the purpose of Knowledge discovery was earlier proposed by M.A. Jabbar et al [4].

V. PROPOSED WORK AND RESULTS

The aim of the proposed work is to design a decision support systems by compare the results of the previous predictions of our work on diabetic data sets in the terms of Co-disease prediction and then analyze the results in order to show the Co-disease predictions with better accuracy and speed. The data classifiers which were built on the Naïve Bayes and ZeroR are compared to the Ensemble approach by combining the weighted voting of the basic and complex algorithms of Naive Bayes and Random Forest so that accurate decisions can be taken in this context of future risks.

A. Naive Bayes and Random Forest approaches

Aggregation of results gives better prediction than the individual result predictors. This can be done by Ensemble learning. A random forest algorithm in data mining is a powerful and superior approach of estimating and random forest on different samples of data sets can fit more and more number of decision tree classifiers [14]. Random Forest Classifier in data mining theory is treated as an ensemble algorithm. Random forest is a classifier that takes the subsets which are randomly selected in order to produce a group of decision trees. And later this algorithm to find and derive the final class for the given test elements in dataset, it aggregates the weighted votes from various decision trees. The sub-sample size and original input sample size are always same in random forest classifiers. Random Forest chooses and selects a random subset and from these subsets it builds many decision trees [13]. These are combination of tree predictors. This model averages the results of all the decision trees obtained [3]. It has the parameters which can be changed so as to improve the prediction generalization. Random forest model can be applied to find out significant health issues and parameters in data mining. Random forest algorithm improves the accuracy of the prediction and for this it uses the idea of averaging. It also takes the same idea of averaging for the control of over-fitting.

Naive Bayes model in data mining is a classifier model supported with Bayesian theorem, this model is an easy and effective classification technique in data extraction theory [15]. It performs more sophisticated classification though it is simple to implement. Naive bayes allows the idea of independent assumption between the predictors of the data set. Conditional independence is the basic theme of the naïve bayes algorithm; it is assumed that the values of some other attributes are independent on an attribute value in a given class.

The Bayes theorem is ruled as : Let us say some set of n attributes called $X=\{x1, x2, ..., xn\}$. Here X is treated as evidence according to Bayesian then, H is considered as means of hypothesis. The data of set represented by X belongs to a specific class called C. P (H|X) is determined, and the probability that the hypothesis H holds given evidence which is data sample X. According to Bayes theorem, $P(H|X) = P(X|H)\,P(H) / P(X)$.

With Naïve Bayesian classifier, training is very easy and fast[16].The training data for Naive bayes algorithm can be of less amount for the estimation of the parameters that are really important in classification.

Our approach of ensemble method reduces unnecessary and redundant properties of data by generating exclusive rules for dataset. This method improves the accuracy and effectiveness of the results compared to the other classifying methods due to the ensemble approach. The original medical record sheets of the diabetic patients were diagnosed to predict and find the possibility of the other diseases which occur along with sweet urine disease which is also referred as diabetes mellitus. Various attributes selected were age, gender, different cholesterol levels in diabetic body, triglycerides, ldl and hdl cholesterols etc. For every 100 instances of data the results found were that twenty to twenty two patients were vulnerable to suffer with heart problems and other symptoms.

The pre-condition set for the prediction of the disease was

if
diabetes = ‘yes’ and age>=50
gender=female||male
total cholesterol >high
HDL=good
LD L =bad
VLDL =bad
TTL =bad,
Triglycerides >high
then
by-disease=’yes’
else
by-disease=’no’

Correctly Classified Instances	91	88.3495 %
Incorrectly Classified Instances	12	11.6505 %

Table 1 Instance accuracy DS1

The other absolute errors and the statistical co relations related to the experiment are given below.

Kappa statistic	0.6686
Mean absolute error	0.1165
Root mean squared error	0.3413
Relative absolute error	29.2446 %
Root relative squared error	76.6544 %
Total Number of Instances	103

Confusion Matrix for the classifiers achieved for our experiment is a follows

```
 a  b   <-- classified as
74  1 |  a = no
11 17 |  b = yes
```

It was found with the 74-11 classification theme.
The detailed accuracy by the class is as follows

TP Rate	FP Rate	Precision	Recall	F Measure	ROC Area	Class
0.987	0.393	0.871	0.987	0.925	0.797	No

0.607	0.013	0.944	0.607	0.739	0.797	Yes
Weighted Avg	0.883	0.29	0.891	0.883	0.874	0.797

Table 2 Accuracy by the class DS1

The other data set with some other attributes was tested to find the Co diseases. The attributes were age, gender, hypertension, diabetes, co-disease. The following was the result tested with Ensemble approach. It was to find the various Co-diseases that were added as an attribute to class.

Correctly Classified Instances	30	32.967%
Incorrectly Classified Instances	61	67.033%

Table 3 Instance Accuracy DS2

Kappa statistic	-0.0227
Mean absolute error	0.1676
Root mean squared error	0.4094
Relative absolute error	89.0568 %
Root relative squared error	134.2898 %
Total Number of Instances	91
Ignored Class Unknown Instances	23

Detailed Accuracy obtained by Class is given below

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0	0.075	0	0	0	0.471	Other
0	0	0	0	0	0.5	Thyroid
0.273	0.116	0.429	0.273	0.333	0.593	Cad
0	0.024	0	0	0	0.491	Sob
0.632	0.755	0.375	0.632	0.471	0.421	Nil
0.	0.06	0	0	0	0.477	Poor vision
0	0	0	0	0	0.496	Poor vision
0	0	0	0	0	0.491	Cad
Weighted Avg	0.33	0.358	0.26	0.33	0.277	0.483

Table 4 Accuracy by the class DS2

Confusion Matrix derived is a follows

```
a  b  c  d  e  f  g  h   <-- classified as
0  0  0  0 11  0  0  0 |  a = other
0  0  2  0  2  1  0  0 |  b = thyroid
0  0  6  0 15  1  0  0 |  c = cad
0  0  0  0  4  2  0  0 |  d = sob
6  0  6  1 24  1  0  0 |  e = nil
0  0  0  1  6  0  0  0 |  f = poor vision
0  0  0  0  1  0  0  0 |  g = poor vision
00  0  0  1  0  0  0 |  h =  cad
```

Prediction using simple Naive Bayes for data set 1 (Instances 103)

Correctly Classified Instances	88	85.4369%
Incorrectly Classified Instances	15	14.5631%

Table 5 Naive Bayes for data set 1

Prediction using simple Naive Bayes for data set 2 (Instances 114)

Correctly Classified Instances	88	85.4369%
Incorrectly Classified Instances	15	14.5631%

Table 6 Naive Bayes for data set 2

Prediction using Zero R (Instances 103)

Correctly Classified Instances	39	37.8641%
Incorrectly Classified Instances	64	62.1359%

Table 7 Zero R Data set 1

The above all results predict that the class derived as a Co- disease or By- disease for the diabetic patients is found with heart disease and poor vision. It can also be seen from the derived tables that the accuracy of the class is improved than the older method.

VI. CONCLUSION

From our work, the results predicted shows that most commonly seen health issue in the diabetic patients is heart disease and next to it poor vision according to the predicted results. In this context, the techniques of classification implemented using Ensemble approach shows the accuracy of the algorithms has been improved when compared to the traditional classification methods used in our work. To detect and predict the other diseases in diabetic patients at the early stage, our work can be a future application in the health care industry with decision support system. The combination of Naïve Bayes and Random Forest algorithms in Ensemble approach has predicted the results with greater accuracy when compared to the separate classification models. The difference in the old and new results can be found with the accuracy gap of 3-4 percent in data set one and nearly 20-30 percent in data set two.

The proposed work can be implemented for precautionary measures and care to diabetic patients where the probability of the occurrence of heart diseases is found. The chance of occurrence of the heart diseases and poor vision in the victims of diabetes mellitus can be avoided by the pre care and prediction of the disease present using the various data mining techniques like the above said.

REFERENCES

1. Long W.J, et al. (1997) : Reasoning requirements for diagnosis of heart disease. Artificial Intelligence in Medicine, 10(1), pp. 5–242
2. Thabtah F., Mahmood Q., McCluskey L., Abdel-jaber H (2010). A newClassification based on Association Algorithm. Journal of Information and Knowledge Management, Vol 9, No. 1, pp. 55-64. World Scientific.

3. Chen J and Greiner R (1999): Comparing Bayesian Network Classifiers. In Proc. of UAI-99, pp. 101–108.
4. M.A.Jabbar, B.L.Deekshatulu, Priti Chandra," anevolutionary algorithm for heart disease prediction, CCIS pp 378-389 Springer Verlag 2012.
5. Liu, B., Hsu, W., Ma, Y. 1998. Integrating Classification and Association rule mining. KDD'98, pp. 80-86.
6. Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science.
7. Definition and diagnosis of Diabetes Millitus and Intermediate Hyperglycemia, report of WHO/IDE,2006 P:21 ISBN 978-92-4-159493-6
8. Hastie, Trevor. Tibshirani, Robert. Friedman, Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, NY, 2009.
9. Labib NM. Data Mining for Cancer Management in Egypt Case Study: Childhood Acute Lymphoblastic Leukemia, Int J Med Health Pharm Biomed Eng. 2007;1(8)
10. Lung Cancer Survival Prediction using Ensemble Data Mining on Seer Data Scientific Programming Volume 20, Issue 1, Pages 29-42
11. Ali Safiyari, and Reza Javidan, "Predicting lung cancer survivability using ensemble learning methods," 2017 Intelligent Systems Conference (IntelliSys), pp. 684–688.
12. Dietterich, T. (1998). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, Boosting, and Randomization. Machine Learning, 40, 139-158
13. http://nymetro.chapter.informs.org/prac_cor_pubs/RandomForest_SteinbergD.pdf
14. Simon BP, and Eswaran C. 1997. AnECG classifier designed using modified decision based neural networks.
15. Rennie, J.D., et al.: Tackling the poor assumptions of naive bayes text classifiers. In: Machine Learning-International Workshop then Conference, vol. 20(2) (2003)
16. Chai, K.; H. T. Hn, H. L. Chieu; "Bayesian Online Classifiers for Text Classification and Filtering", Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval, August 2002, pp 97-104

AUTHOR'S PROFILE

Shahebaz ahmed Khan is a Research Scholar at Shri Jagdish Prasad Jhabarmal Tibrewala University of Jhunjhunu. He has completed his B.Tech and M.Tech in CSE from Bharat Institute of Engineering and Technology, Hyderabad. He has published more than 13 research papers in various journals and conferences of both national and international levels. Shahebaz is doing his PhD in Computer Science and Engineering with Data Mining Specialization. His areas of interest are Data Mining, Machine Learning, Computer Programming, Information Retrieval Systems and Operating Systems. He has also written and published four books on various issues in the society which are of general nature. He is a member of ISTE, IAENG, CSTA, UACEE and SDIWC. The research work carried out by Shahebaz includes the prediction of diseases using data mining techniques with special reference to diabetes and thyroid.

Dr. M.A.JABBAR is a Professor and Centre Head at the Computer Science and Engineering Department, Vardhaman College of Engineering, Hyderabad, Telangana, India. He has been teaching for more than 19 years. He obtained Doctor of Philosophy (Ph.D.) from JNTUH. He published more than 50 papers in various journals and conferences. He is
Reviewer for Scopus and SCI journals like Springer, Elsevier, and IEEE Transactions on Systems Man and Cybernetics, Wiley. He served as a technical committee member for more than 40 international conferences. He published 5 patents (Indian) in machine learning and allied areas**.** He is a senior member of IEEE, also a member of ACM, Life member of CSI, ISCA and currently he is Vice-chair, Computer Society Chapter of IEEE Hyderabad Section