International Conference on Recent Innovations in Science, Engineering & Management
R.R. INSTITUTE OF TECHNOLOGY, Bengaluru (India)
10th - 11th May 2019 www.conferenceworld.in ISBN : 978-93-87793-87-3

Development of Predictive Model to Improve Accuracy of Medical Data Processing using Machine Learning Techniques

Likitha V[1], Mrs. Sowmya Naik[2], Prof. Manjunath R[3]

[1]*PG Student, Department of Computer Science & Engineering,*
City Engineering College, Bengaluru,India,
[2]*Assistant Professor, Department. of Computer Science & Engineering,*
City Engineering College, Bengaluru, India,
[3]*Professor, Department of computer Science & Engineering, RRIT, Bengaluru, India*

ABSTRACT: *Data mining is nothing but the process of viewing data in different angle and compiling it into appropriate information. Out of the many software tools used for data evaluation, the one which is widely used is the data mining. Technically the data mining can be considered as the sequence of steps followed for searching patterns or identifying correlations between large numbers of fields within a huge relational database. Recent improvements in the area of data mining and machine learning have empowered the research in biomedical field to improve the condition of general health care. Within the medical data, the medical data mining searches for patterns and relationships which can provide useful information for appropriate medical diagnosis. Data mining techniques are applied to different medical domains to improve the medical diagnosis. Improving the accuracy of the classification and improving the prediction rate of medical datasets are the main tasks/challenges of medical data mining. Since the wrong classification may lead to poor prediction, there is a need to perform the better classification which further improves the prediction rate of the medical datasets. When medical data mining is applied on the medical datasets the important and difficult challenges are the classification and prediction. In this proposed work we evaluate the data mining techniques like Logistic Regression (LR), Artificial Neural Networks (ANN), Support Vector Machines (SVM) and Random Forest (RF) with Feature Selection Methods (FSMs) and Percentage Split (PS) as test option on Diabetes Datasets. The performance of the proposed hybrid model is measured in the form of classification accuracy.*

Keywords: BE: Backward elimination; CA: Classification Accuracy; EE: Entropy Evaluation; FSM's: Feature Subset Selection Methods; FS: Forward selection.

I. INTRODUCTION

Recent improvements in the area of data mining and machine learning have empowered the research in biomedical field to improve the condition of general health care. In many parts of the world the tendency for maintaining long-lasting records consisting of medical data is becoming an accepted practice. In addition to this,the newer medical equipment's and the techniques used in diagnosis, produces composite and huge data. Therefore, to handle these ill-structured biomedical data, intelligent algorithms for data mining and machine learning are required in order to take logical reasoning from the saved raw data, which is considered as medical data miningthe newer medical

equipment's and the techniques used in diagnosis, produces composite and huge data. Therefore, to handle these ill-structured biomedical data, intelligent algorithms for data mining and machine learning are required in order to take logical reasoning from the saved raw data, which is considered as medical data mining. Within the medical data, the medical data mining searches for patterns and relationships which can provide useful information for appropriate medical diagnosis [3]. Data mining techniques are applied to different medical domains (health care databases or medical datasets) to improve the medical diagnosis. To check for any invisible patterns inside the medical datasets, medical data mining is strongly recommended. In medical data mining, the actual tasks (challenges) are the classification and prediction of medical datasets. To manage these tasks the following methods are used most often.

i. LR: LR is one of the data mining methods used for analysing problems where the outcome is determined based on one or more independent variables. A dichotomous variable is used to measure the outcome. In LR, the non-independent variable is dichotomous or binary i.e., it consists of data represented as 0 (FALSE, failure, etc.) or as 1 (TRUE, success, etc.) [4]. In various biomedical fields such as cancer analysis, survival forecast, kidney transplant etc. [5] [6], LR has been widely used. Even in statistics, it is a well-established and a powerful method. It is suggested that LR has to be compared to data mining techniques while performing medicinal data mining [7].

LR is implemented on the health care databases for detecting the patterns which are useful for either forecasting or determining the diseases along with take the remedial measures for handling such diseases [8].

ii. ANN: ANN is one among the various fields of Artificial Intelligence. The human brain architecture is the main inspiration behind the development of the model. ANNs are successfully used in various disciplines such as environmental science, study of human mind, study of numbers, study of medicine, study of computers etc. ANNs are also being used in many business areas like accounts and audits, funding, managing and decision making, promotion and manufacture etc. ANNs have turned out to be a well-liked model and recently they are used to identify diseases and to forecast the patients' survival proportion [9]. ANN models or "neural nets" are also called by different names. Whatever the name is; each one of these models tries to give good performance through compact interconnection of uncomplicated computational elements. For many years these models have been studied with a hope of achieving the performance like humans in the field of speech and image recognition [10].

iii. SVM: In machine learning SVMs [1] are the models used for supervised learning accompanying with other learning algorithms which can analyse data used for regression and classification. For any set of training examples given, each of them is marked as fitted to one or other group, an SVM

training algorithm constructs a model that allocates new examples to a single category or the other, constructing it a non-probabilistic binary linear classifier.

iv. RF: RM [1] [2] are an ensemble learning method for regression, classification, and other jobs, that functions by making an assembly of decision trees at training time and generating the class that is the classification or regression of the distinct trees.

II. LITERATURE SURVEY

For optimizing the parameter for SVM, an Adjusted Bat algorithm (ABA) is proposed. The experiments are conducted on the diabetes disorder dataset. The experimental result was compared with the Grid-SVM and other approaches. Based on the result, ABA-SVM is considered as a better classifier than Grid-SVM and compared to other approaches like PSO-SVM and PTVSPSO-SVM, the ABA-SVM achieved better classification accuracy [12]. A method similar to PCA was used to select the important attributes was developed. These attributes are given as an input to the feed forward ANN. The result achieved by the method is measured up with other methods of the feature selection like Tarr's, RUCK's, PCA and t-test. The new model was applied on the diabetes disorder dataset. Testing is done using 20% of data and remaining 80% is used for training. The proposed method achieved good classification accuracy with less number of attributes [11]. A unique algorithm is presented for the induction of full oblique decision trees (EFTI). The algorithm depends on single and special evolutionary algorithm, which generates a full decision tree by altering the node coefficients and structure of the complete tree at the time of evolution. EFTI algorithm is often used in embedded applications, since it uses small resources for computation when compared with decision tree inference algorithm. The algorithm is implemented on diabetes disorder dataset and the result was compared with other approaches based on decision tree. The proposed algorithm generates better result the other [13]. A growing-pruning spiking neuron network (GPSNN) consisting of 2 stage learning algorithms is developed for handling the problems of pattern classification. The GPSNN consisted of three layers and two stages of learning algorithm. The GPSNN was experimented on diabetes disorder dataset. The outcomes are evaluated with batch and online spiking neuron. From the result, it was identified that GPSNN achieved better accuracy that the other [14]. EUCAFES is a robust filter which works on feature weighing approach. Feature weighing approach is used to calculate the weights of the binary feature and gives the detailed information related to feature based on continuous weight. The technique is applied on diabetes disorder dataset. Based on the result we can see that RBF-DDA achieved good accuracy with a smaller number of attributes [15].

In addition to earlier kernel Fisher Discriminate (KFD), by employing heterogeneous kernel model, an iterative algorithm is proposed for KFD. The new KFD selections of kernels are automatic. The

proposed method wasimplemented on diabetes disorder dataset. From the experiment it was observed that the new KFD gives better classification than the earlier KFD [16]. The kNN classifiers are delicate to noise and the outliers present inside the training dataset. Two approaches of depuration algorithm are employed to edit training data. For kNN classifier, to edit the data neural network ensemble is made use of. The method was implemented on diabetes disorder dataset. From the result, we can see that kNN is much better than the two methods of depuration algorithm [17]. For data reduction or compression, a method based on multidimensional scaling is proposed. This method can produce shorter vectors from data vectors of high dimension, but with some loss of information. The formal model for data reduction in Bayesian framework is the Bayesian networks. The method was applied on diabetes disorder datasets. The result of kNN is compared with Naïve Bayes.

III. FRAMEWORK

In our hybrid framework selection of important attributes, reduction attributes for the classification and prediction of diseases for a given medical data set is proposed.The proposed model is shown in the following Figure 1. Proposed model consists of the following steps:

1. First step is the collection of BUPA diabetes dataset.
2. Pre-processing is done for any missing values
3. For pre-processed data we apply Entropy Evaluation method.
4. Based on the entropy value we apply the FSMs like FS and BE. This results in generating different subsets of attributes.
5. For each attribute we evaluate the performance of LR, SVM and RF using PS as test option.

Finally, we identify the method that achieves the best CA as the best method for the prediction of Diabetes data.

ISBN : 978-93-87793-87-3

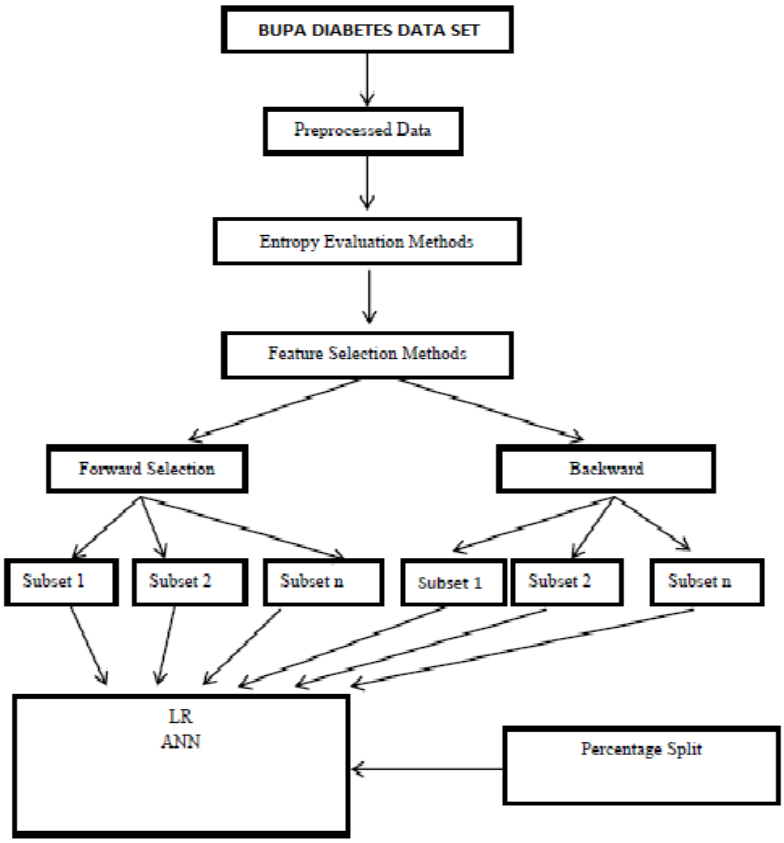

Fig 1: Proposed model for prediction

IV. RESULTS AND DISCUSSION

The below is the table for the accuracy attained for the full attribute set of PIMAIndian diabetes dataset.

Table 1:Classification Accuracy for full attribute set of PIMA Indian diabetes dataset

Techniqueused for finding CA	Percentage Split				
	50%	**66%**	**70%**	**75%**	**80%**
LR	83.8	83.2	82	81.2	79.9
NN	83.29	83.11	82.88	82.8	83.4
NN with 10 Fold	83.70	**84.52**	83.35	83.55	83.8
SVM	72.9	69.9	64.7	66.4	65.3

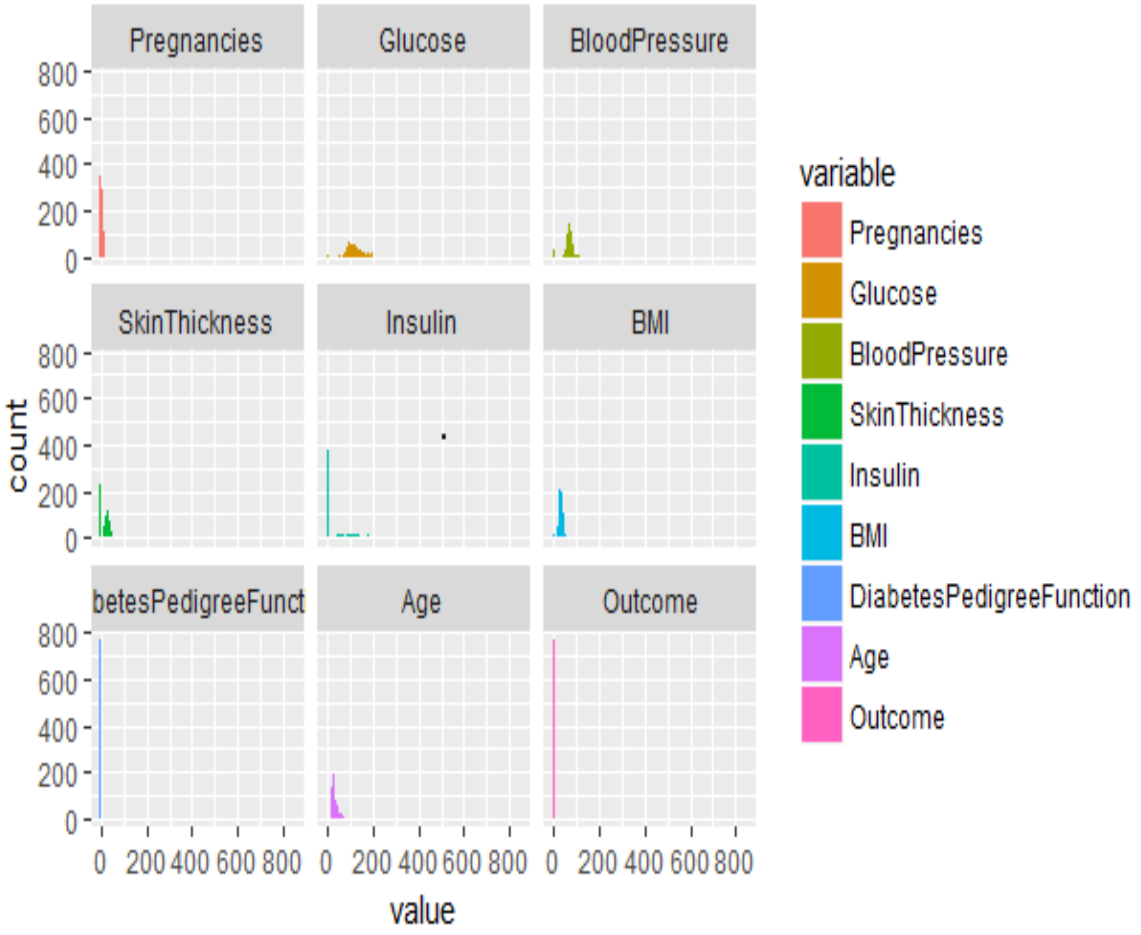

Fig 2: Count of each attributes with value

V. CONCLUSION

The proposed research work uses feature selection methods like forward selection for Diabetes disorder medical dataset. LR, NN and NN with 10 fold CVS are applied on feature selection methods using Cross Validation Sample and Percentage Split as test options. From the experimental results it is identified that for Reduced Diabetes Disorder dataset with NN using percentage split of 66%, prediction accuracy of 84.52% is achieved. For all datasets used in the research work gives better classification accuracy with reduced subset of features. From the experimental results it is observed that the reduced subsets of attributes gives more efficient results than that obtained by using full set of attributes.

REFERENCES

[1] S. Sumathi and S. N. Sivanandam. "Introduction to Data Mining and its Applications", Studies in Computational Intelligence, Volume 29, Springer, 2006.

[2] R. Tamilselvi and S. Kalaiselvi. "An Overview of Data Mining Techniques and Applications", International Journal of Science and Research (IJSR), Volume 2, Issue 2, pages 506-509, 2013.

[3] Siri Krishan Wasan, Vasudha Bhatnagar and Harleen Kaur. "The Impact of Data Mining Techniques on Medical Diagnostics", Data Science Journal, Volume 5, pages 119-126, 2006.

[4] Raghavendra B. K. and Dr. Jay B. Simha. "Evaluation of Logistic Regression Model with Feature Selection Methods on Medical Datasets", ACS-International Journal on Computational Intelligence, Volume 1, Issue 2, pages 35-42, 2010.

[5] J. Chhatwal, O. Alagoz, M. J. Lindstorm, C. E. Kahn, K.A. Shaffer and E.S. Burnside. "A Logistic Regression Model Based on the National Mammography Database Format to Aid Breast Cancer Diagnosis", American Journal of Roentgenology, Volume 192, Issue 4, pages 1117-1127, 2009.

[6] H. Khedmat, G. R. Karami, V. Pourfarziani, S. Assari, M. Rezailashkajani and M. M. Naghizadeh. "A Logistic Regression Model for Predicting Health-Related Quality of Life in Kidney Transplant Recipients", Transplantation Proceedings, Elsevier, Volume 39, pages 917-922, 2007.

[7] Riccardo Bellazzi and Blaz Zupan. "Predictive Data Mining in Clinical Medicine: Current Issues and Guidelines", International Journal of Medical Informatics, Elsevier, Volume 77, Issue 2, pages 81-97, 2008.

[8] Raghavendra S. and Dr. Indiramma M. "Performance Evaluation of Logistic Regression and Artificial Neural Network Model with Feature Selection Methods using Cross Validation Sample and Percentage Split on Medical Datasets", Proceedings of the 2nd International Conference on Emerging Research in Computing, Information Communication and Applications, Volume 2, Elsevier Publication, pages 750-755, 2014.