Available Online at http://www.recentscientific.com

CODEN: IJRSFP (USA)

International Journal of Recent Scientific Research
Vol. 9, Issue, 6(B), pp. 27340-27344, June, 2018

ISSN: 0976-3031

International Journal of Recent Scientific Research

DOI: 10.24327/IJRSR

Research Article

DIABETES DATA PREDICTION USING DATA CLASSIFICATION ALGORITHM

Ammulu K*

Department of Computer Science Dravidian University Kuppam, A.P

DOI: http://dx.doi.org/10.24327/ijrsr.2018.0906.2239

ARTICLE INFO

Article History:

Received 17th March, 2018
Received in revised form 21st April, 2018
Accepted 05th May, 2018
Published online 28th June, 2018

Key Words:

Data Mining, Classification, Modified Random Forest Algorithm, Diabetes Dataset.

ABSTRACT

Diabetes is the major disorder occurs due to the lack of produce insulin in the body among human being. All types of diabetes can lead to complications in many parts of the body and can increase the overall risk of dying prematurely. The risk of diabetes is increasing day by day and is found mostly in women than men.

The very dangerous disease in medicinal field is diabetes disease which is affected for many peoples in popular countries like India. The diagnosis of diabetes is a tedious process. So with improvement in science and technology it is made easy to predict the disease. The purpose is to diagnose whether the person is affected by diabetes or not using Random Forest (RF) Classification Algorithm. The diabetes dataset is a taken as the training data and the details of the patient are taken as testing data. The training data are classified by using the RF classifier and secondly the target data is predicted. RF algorithm used here would be more efficient for both classification and prediction. The results are analyzed with different values for the parameter k.

INTRODUCTION

The target of the information extracting method is to extract data from a dataset and make over it into a clear construction for additional use. This is a diagnostic method planned to scrutinized the information in seek of reliable patterns or organized associations connecting variables, and then to confirm the findings by applying the detected patterns. The focal point of this document is to concern a variety of categorization methods such as J48, kNN, CART and SVM. As the commonness of diabetes is on the rise, there is a proportionate rise in the complications that are associated with diabetes and the illness has been the most deadly disease in the United States with no imminent cure in sight. The diabetic sickness has number of side effects like eye disease, kidney failure, and additional complications. However, early detection of the disease and proper care management can make a difference. It is the reasons sugar to build up in blood leading to complications like heart disease, stroke, blindness, kidney failure, nerve damage, and death. Regular Symptoms of Diabetes are increased thirst, increased urination, Weight loss, unsettled stomach or vomiting - Blurred vision, Slow-healing infections and weakness in men. There are a variety of research work is carried out by many researchers based on the observed medical diabetes data. Some of such works are discussed hereafter. Arvind Sharma and P.C. Gupta described that data mining can add by means of necessary benefits to the blood stockpile division. J48 method and WEKA software has been utilized for the entire research work. Classification rules performed well in the classification of blood donors, whose accuracy rate reached 89.9% . The research is aimed at finding out the characteristics that determine the presence of diabetes and to track the maximum number of men and women suffering from diabetes with 249 population using WEKA tool. Asha Gowda Karegowda, M.A. Jayaram, A.S. Manjunath , used cascading k-Mean and kNN algorithm for cataloging of diabetic patients in their paper. They classified diabetic patients by proposing results using kNN and k-Mean. Accuracy achieved by the proposed system is 82%. Hardik Maniya, Mosin I. Hasan, Komal P. Patel , have done the relative study of Naive Bayes Classifier and kNN for Tuberculosis, and justify the effectiveness of results using kNN can be further improved by increasing the number of data sets and for Naïve Bayesian classifier by increasing attributes or by selecting weighted features. W. Yu, and W. Zhengguo , have gave the investigational result shows the classification using traditional kNN algorithm produce normal evaluation value, with fulfillment rate of 75%. Y. Angeline Christobel, P. Sivaprakasam, the concert of classification calculate regarding sensitivity, specificity and accuracy has been increased

Corresponding author:* **Ammulu K
Department of Computer Science Dravidian University Kuppam, A.P

significantly in the case of proposed CkNN method. In a research work of, Estebanez, Alter and Valls used genetic programming for classification tasks. The error rate for SVM is 22%, Simple Logistics is 22.14% and Multilayer perceptron is 23.31%. In another work some of the classification algorithms are compared by utilizing matrix and classification accuracy. The 10-fold cross validation method was used by three different types of breast cancer databases and calculated the accuracy. Jianchao Han used type 2 diabetes data for his effort and the decision tree using WEKA has been used to put up the prediction model. The main element for his research was predicting the disease is the models of Plasma Insulin. Asma A. Aljarullah in her research work J48 decision tree classifier was used. Using Diabetic data set was used to implement Association rule. B.M. Patil finds out different range of accuracies using some of classification techniques on the diabetes dataset. Weighted least squares support vector machine based on quantum particle swarm optimization algorithm is used to development in the prediction accuracy. E.G.Yildirim in his research work the type 2 diabetes data set is used to predictive data mining and applied in dosage planning. Adaptive Neuro Fuzzy Inference System and Rough Set theory methods are coated by him. The main objective of G. Parthiban et al. in their research work the prediction and changes of diabetic patient getting heart related problem. In their research they used Naive Bayes classifier method which gives the best possible prediction model.

The organization of this paper is formed as follows. Section II states some basic concepts of classification algorithms and its applications. Section III mentions the proposed algorithm. Section IV mentions the experimental results of the classification algorithm Random Forest algorithm for diabetes data set. Finally, the conclusion of this research work is given in section V.

Datamining Classification Algorithms

Generally, the classification is the primary part of the data mining concepts which is used to identify the group membership for data instances. It is also called as the supervised machine learning technique where the availability of labeled data in prior. The part of the dataset is used for training sets to predict the future data. The way of predicting the class to which the data belongs to is referred to as prediction. The training is based on the samples provided for it. Classically there are two types of attributes available, they are

- Input or independent attribute, and
- Output or dependent attribute.

The mapping of input datasets to the limited set of different set of class labels takes place in supervised classification. Consider the input dataset as Y € D_i, I is the dimensionality of input space and X is the discrete set of class labels, belongs to 1 to m, such as X € 1….m where m is the sum of class types. X=X(y,u), u is the vector of the adjustable parameters.

The data mining classification methods are as follows:

- Decision tree induction:

The decision tree is constructed based on the class labels. The structure of decision tree is based on the flowchart and contains leaf node, branch, and internal node. The leaf node represents the labels of class, branch is used to denote the test outcome, and internal node is used to detect the attribute test. The basic advantage of the decision tree is simple and fast. The primary goal is to predict the solution for the linear attributes, where the decision tree is less applicable for the task estimation. The complexity of the tree is occurrence of error and the number of leaves leads to expensive construction of tree. However the splitting of node also happens at each level.

- Rule – based classification:

The Rule based classification is based on the IF-THEN rules. Generally, the number of rules are examined to identify how these rules are constructed, how to develop it from the decision tree, how the rules are developed from the training data. Rules based classification is expressed as follow:

IF condition THEN conclusion

Now we define accuracy and coverage of S by following expression

Coverage (R) = $Ntotal/$I$\boldsymbol{D}$I

Coverage (R) = $Ncorrect/Ntotal$

Lazy learners

Lazy learner is completely contrast to the eager learner. In eager learner the system generalize the training data earlier the tuple is received whereas in lazy learner the generalization of training data is stored until the tuple is given to the system. This is suitable for the large dataset having countable attributes and it also support incremental learning. The lazy learning provides good influence in case based reasoning and K- nearest neighbor classifiers.

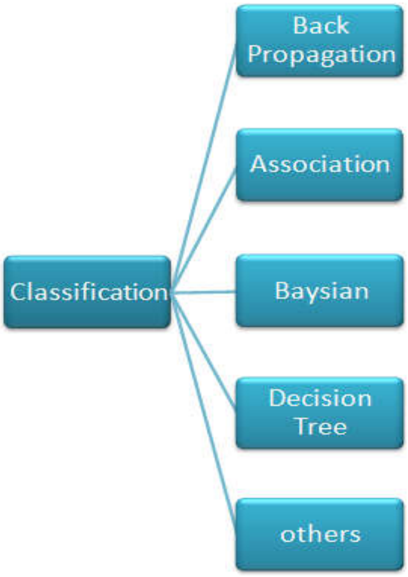

Figure 1 The types of classification in data mining

Classification by back propagation

The classification by back propagation is a neural network learning algorithm. The neural network learning algorithms have number of input and output units which are interconnected with each other and each connection have weight associated to it. This is also called as connectionist learning due to number of weighted connections. The algorithm is processed by executing the data linearly and learns by comparing the outcomes to the target value assigned earlier.

Bayesian classification

It is statistical classifier used for predicting the class membership probabilities it predict whether the tuple belongs to the particular class. It is also known as naïve baiyes due to the prediction of independence between diverse attributes values .The naïve baiyes algorithm estimates the prior probability based on the count of how often the classes occur in the training data. To classify a target tuple the conditions and prior predictions are used from the training data to make the prediction. The Bayesian classification is easy to handle and uses only one scan of training data. The baiyes can easily handle the missing value by omitting the particular probability calculation. The baiyes is straightforward but does not provide a satisfactory result. The algorithm cannot handle continuous data.

Decision tree based algorithm

It is most commonly used in classification. In this approach a tree is constructed to model the classification process. When the tree creates the tuples in the database will be applied and the result classification for the tuple will be provided. The two essential steps in this process is constructing the tree and administering the tree to the database. The approach of this method in classification is to divide the space in to rectangle regions. The tuples will be classified based on the region it falls.

K-Nearest neighbors

This algorithm is based on the distance based classification. The KNN uses the entire data set not only the data but also the desired classification for each. The training data is taken as the model for performing a new classification for new item set the distance of the training set as to be calculated only the k-nearest training data will considered in the new item set.

Iterative dichotomiser 3

This algorithm is utilized to create a decision tree based on the information theory and to minimize the expected number of comparison. It does not provide an optimal value. The algorithm uses backtracking for searching the optimal decision tree. It is harder to use in continuous data. It is based on Occam's razor it prefers smallest decision tree over larger.

Classification and regression tree

This algorithm is used for generating a binary decision tree. The carts consist of only two child node and sub divided based on the best spilt value. The spilt value is calculated based on the tuples in the sub tree and the tuples in the training set. The chart handles the missing data by simply ignoring that record and calculating the splitting value will improve the performance. The CART is best for training data and not possible for all data. It consists of pruning strategy.

Neural network based algorithm

It is based on the decision tree algorithm to classify any given database tuples and construct a tree. The input is given into the directed graph in the corresponding source node. This method will access all training data or until the classification is accurate or adequate.

Proposed Algorithm

The main approach in the decision tree learning is decision tree pruning. The process of removing the tree that present poor voting at the final classification output by reducing the tree size is defined as the tree pruning. The main advantage of this is reducing the tree size, the complication of the build tree, and the reduction of over-fitting. In most of the cases, random forest algorithm will not use this approach. Alternatively, random forest classifier use n tree as parameter that responds to the number of decision trees which is created during the ensemble bagged forest classifier. The random forest algorithm describes the datasets training for each decision tree.

Algorithm: Random Forests Training

```
Step 1: Set: Number of classes = N, Number of features = M
Step 2: Let: m determine the number of features at a node of decision tree, (m < M)
Step 3: for each decision tree do
Step 4: Select randomly: a subset (with replacement) of training data that represents the N classes and use the rest of data to measure the error of the tree
Step 5: for each node of this tree do
Step 6: Select randomly: m features to determine the decision at this node and calculate the best split accordingly. /* No tree pruning used */
Step 7: end for
Step 8: end for
```

Random forest algorithm takes more time to build multiple classifiers even it is highly parallel algorithm. This can't be processed in a usual core (need multiple core) and also not without any graphical processing unit. Initially, the dataset needs to be trained in order to build the classifier at the testing phase. The large number of trees will take more time for the real-time prediction of the classifier. However the more number of trees will provide more accuracy in final. Once it is trained, the remaining dataset is treated as a testing dataset where the prediction of class is not available at shorter duration. The list of parameters used in Random forest algorithm is listed in Table I.

Table I. Random Forest Parameters Optimized

Parameters	Range	Mode
Ntree	1 – 100	Discrete
Sampsize	100 – 10000	Discrete
Mtry	1 - max. no. of predictors in dataset	Discrete
node	size 1 - 100	Discrete
Number of features for finding best split node (m)	Min:2 – Max:100	Discrete

Algorithm: Modified Random Forests Training

Step 1: Initialize Population. The population size is set to pop size, the max iteration time is set to maxgen, the position of the binary particle is Xk= {Zk,1, Zk,2, …}, k = 1, 2, … popsize, the velocity is V, the learning factors are c1, c2, and the weight is w.

Step 2: Combine the PSO with RF classification and calculate the fitness function F = max(1/f), gen = 1.

Step 3: WHILE (till the ending criterion)
Step 4: FOR p=1 to n particles

Choose attributes
Divide training data and testing data using K-fold
Cross-validation
Train the data available for it
Classify the remaining data available for test
Store the estimated rate in an array

Step 5: For NEXT p

Step 6: Update particle's velocity and position Vk + 1 velocity of particles and Xk + 1 as position of particles. Let Pk be the optimal position of an individual particle, Pgk be the optimal position of all particles, and rand be a random number uniformly distributed in the range (0,1)

Step 7: If gen>maxgen, the algorithm will terminate; otherwise, return to Step 6.

Step 8: NEXT generation ends until it reaches the stopping criterion

Modified Random forest algorithm is one of the best authentic learning algorithms available for enormous data sets where it constructs more accurate classifier. There is adequate approach to obtain the missing data and it also maintains the accuracy even there is huge amount of data is missing. This provides an internal impartial estimation of error occurred due to generalization as the forest built prior. The prototypes are processed by the data given for the interaction between the classification and variables.

EXPERIMENTAL RESULT

List of notations used in the proposed method of classification:
D = Dataset
N= Number of instances
TP = True Positive
FP = False Positive
TPR = True Positive Rate
FPR = False Positive Rate
P = positive instances
N* = negative instances
A = Accuracy
P = Precision
R = Recall
F1 = Score
r = range of misclassified instances

The confusion matrix illustrates the more accuracy of the solution to a classification problem. The classification result can be obtained from the true positive, false positive, true negative and false negative rate. The term true and false refer to whether the predicted corresponds to the trusted external judgments. Regard P as positive instances and N* as negative instances then the true positive rate and false positive rate is TPR and FPR respectively. Precision is defined as the range of true positives over the number of true positives plus the number of false positives. Recall is defined as the number of the true positives over the number of true positives plus the number of false negatives. The F1 score is a measure of the test's accuracy where both the precision and recall are used to compute the rate. Accuracy is explains as the number of all correct predictions split by the total number of the dataset.

The dataset contains four fields per record as date, time, code, value and the dataset type is multivariate, time-series with no missing value. The table shows the true and false positive count. The following Table I represents the confusion matrix values for both the test and the train data. The accuracy of random forest algorithm and the modified random forest algorithm for the Diabetes dataset is measured by varying number of trees, results are shown in Table II. Precision, recall, F1 score, Accuracy of RF and MRF is evaluated and corresponding graph is shown in figure 2. Accuracy with number of trees is compared and corresponding graph is depicted in figure 3.

Table I Confusion Matrix

	Predicted Positive	Predicted Negative
Observed Positive	TP 941	FN 16
Observed Negative	FP 27	TN 173

$P=TP/(TP+FP) = 0.9721$
$R=TP/(TP+FN) = 0.9832$
$F1 = 2x[(PXR)/(P+R)] = 0.9775$
$A = (TP+TN)/N = 0.9628$

The accuracy measure of random forest algorithm and the modified random forest algorithm for the Diabetics dataset.

Table II Accuracy of RF and M**RF** for the Diabetes dataset

Number of Trees	Accuracy of RF	Accuracy of modified RF
10	0.485604	0.527493
20	0.518294	0.555933
30	0.550984	0.584373
40	0.583674	0.612813
50	0.616364	0.641253
60	0.649054	0.669693
70	0.681744	0.698133
80	0.714434	0.726573
90	0.747124	0.755013
100	0.779814	0.783453

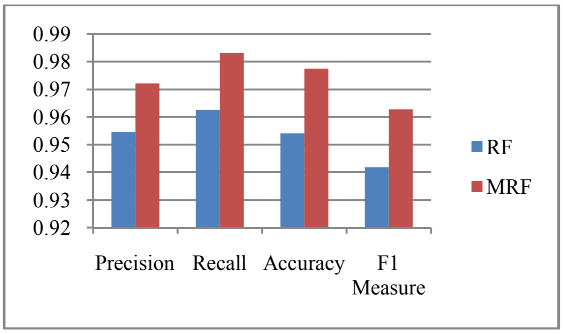

Figure 2 Performance comparison of RF and MRF algorithms

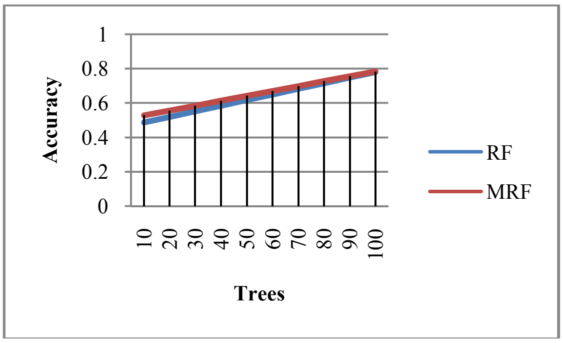

Figure 3 Accuracy Vs. Number of trees in RF and MRF algorithms

CONCLUSION

Data mining is a tool to extract interesting patterns. In our paper we focus on data mining classification technique. Diabetes dataset with 4 variables and 1157 instances of data is supplied as a input to classification algorithm. Random forest algorithm is a traditional algorithm and facing accuracy problem. We proposed modified random forest algorithm to resolve the problems faced by the random forest algorithm. The experimental work is implemented by using the java language in Net beans IDE. Proposed modified random forest algorithm shows better results when compared with existing algorithm

References

1. M. Z. Islam, and H. Giggins, "Knowledge Discovery through SysFor-a Systematically Developed Forest of Multiple Decision Trees", Proceedings of the 9th Australian Data Mining Conference, pp. 195-204, Ballarat, Australia, December 1-2, 2011.
2. R. Polikar, "Ensemble Based Systems in Decision Making," IEEE Circuits and Systems Magazine, vol. 6, pp. 21-45, Third Quarter, 2006.
3. M. F. Amasyali, and O. K. Ersoy, "Classifier Ensembles with the Extended Space Forest", IEEE Transaction on Knowledge and Data Engineering, vol. 26, pp. 549-562, March 2014.
4. L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, "Classification and Regression Trees", Wadswart, Belmont, 1984.
5. Sarika Pachange, Bela Joglekar and Parag Kulkarni, "An ensemble classifier approach for disease diagnosis using Random Forest", Annual IEEE India Conference (INDICON), pp. 1-5, 2015.
6. T. Sousa, A. Silva, and A. Neves, "Particle Swarm based Data Mining Algorithms for classification tasks," *Parallel Computing,* vol. 30, no. 5-6, pp. 767-783, May/June 2004.
7. Alireza Aminsharifi; Shima Pouyesh; Hamid Parvin, "Building a Diverse Ensemble for Classification", Fourteenth Mexican International Conference on Artificial Intelligence (MICAI), pp. 145-151, 2015.
8. Jiawei Hanl, Yanheng Liul and Xin Sun, "A Scalable Random Forest Algorithm Based on MapReduce", published in 4th International Conference on Software Engineering and Service Science, IEEE, pp. 849-852, 2013.

How to cite this article:

Ammulu K.2018, Diabetes Data Prediction Using Data Classification Algorithm. *Int J Recent Sci Res.* 9(6), pp. 27340-27344. DOI: http://dx.doi.org/10.24327/ijrsr.2018.0906.2239

****