Analysis of Data Mining Classification Based Algorithms in health care sector

Dr. Mrs. Y. V. Bhapkar

Asst. Professor,
Yashwantrao Mohite College of Arts, Science and Commerce
Pune 38

Dr. A. B. Nimbalkar

Asst. Professor,
Annasaheb Magar College
Pune28

Abstract

Data Mining is a way towards analyzing data findings covered up or obscure examples in extremely large datasets that are possibly helpful and logical. The objective of data mining is to extract meaningful data from tremendous informational repositories. Data mining provides different views and summary operations into useful information. Data mining is the process of discovering hidden or unknown patterns from the huge datasets , these patterns are potentially valuable and eventually understandable. The goal of data mining is to develop an understandable and structured model by applying different data mining techniques for future use. These techniques are based on statistics, machine learning and database management theory. Data mining plays important role in all the domains including science, commerce, health care industries, marketing, banking, telecommunication, government organizations, agriculture, educational sectors, weather forecasting , web applications and many more. This study is based on data mining case study in health care sector. In this sector data mining plays important role to predict a disease at early stage for future diagnosis. The main objective of this study is to predict diabetes depending on few given attributes. Diabetes is a unceasing disease caused due to the increased level of sugar in the blood. Various automated information systems were developed which utilizes the various classifiers to anticipate and diagnose the diabetes. In this case body does not properly process food for use as energy. The pancreas, make a hormone called insulin to help glucose get into the cell of our bodies. If diabetes is not processed and disclosed at earlier stage then many complications may occur. Early diagnosis can save individual's life and can manage over the diseases. Diabetics identifying processes results in visiting of a patient to a diagnostic center and consulting doctor. Using machine learning approaches we may solves this problem. The motive of this study is to design a model which can predict the likelihood of diabetes in various patients with maximum accuracy.
We have proposed the use of Naïve Bayes,J48 ,Random forest and Multi layer perseptron classifiers for developing diabetics detection models .And then compared the models for the best accuracy.

Keywords
Data Mining, Naive Bayes ,Multilayer perseptron,Decision Tree J48 ,Random Forest.

I. INTRODUCTION

The two most important techniques in data mining are Classification and Clustering , that are often called as supervised and unsupervised learning technique respectively. Supervised learning is the machine learning task of learning a function from labeled training datasets consisting of a set of training examples , that consists of input-output pairs, this function or a rule maps an input to desired output.
Classification is the data mining method to develop a set of models that describe and distinguish data classes and concepts, for the reason of being able to use the model to predict the class membership whose label is unknown. Classification is a two step process, first, it build

classification model on training dataset. Every object of the dataset must be pre classified so its class label must be known. In the second step the model generated in the previous step is tested by assigning class labels to data objects using test data set. Each sample is assumed to belong in a predefined class, as determined by the class label attribute. This classification based model presents classification rules, decision trees, or mathematical formulae. Next important step is model usage. It is for classifying future objects. This is use to estimates accuracy of the model. The known label of test sample is compared with the classified result derived from the model. Model construction describe a set of predetermine classes. Accuracy rate is the percentage of test set samples that are correctly classified by the model. Test set samples are independent of training set, otherwise problem of over-fitting may occur. If the accuracy is acceptable then the model is use to classify data samples whose class labels are not known.Classification algorithm is used to predict categorical label of a given data instance. Classification finds a function or a rule that assigns an item to a predetermined target categories or classes. Data instance for training is represented as attribute or feature vector is X and its label is y, represented as (X,y) ,where X=(x1,x2,x3,......,xn) such that x1,x2,x3,. ,xn are the values of attributes A1,A2,A3,.....,An in the dataset, and y represents the value of label. Classification Method focuses on a survey on various classification techniques that are most commonly used in data-mining.

Machine learning supports variety of algorithms for classification and prediction .But the selection of a particular algorithm is quite difficult. To predict the desired outcome ,algorithm processes a training set which consist of a set of attributes and the prediction attribute.

The algorithm analyses the input data set and produces a prediction. Accuracy of prediction is a criterion which defines how good the algorithm is.In case of medical science database training set contains patient data records and prediction attribute is whether or not the patient has a disease.

In this research we have study and compared classification based data mining algorithms. These are Naïve Bayes ,J48,Random forest ,Multi layer perseptron. Performance of the classifiers shows the strength and accuracy of each algorithm for classification in term of efficiency and time complexity.

II. METHODOLOGY

Following are the data mining classification algorithms used for this study

- **J48 decision tree learning algorithm**

J48 algorithm in WEKA environment that is newest form of previous ID3 and C4.5 algorithms . J48 is a standard algorithm for performing the partition that has been upgraded over the time and it is totally based on the perception of information-theory. The core idea behind this is to select the appropriate variable that can provide information needed to realize suitable partitioning in individual branch to other branches for classifying training set.

Basic Steps performed by J48 Algorithm are

i. If the instances belonging to the similar class for which tree represents a leaf , then it returned a leaf by giving a label for the same class.
ii. For every test feature information gain is calculated.
iii. From step 1 and step 2 best attribute for branching is selected.

Decision tree algorithm implemented on medical database is described below

Input D is data partition i.e. the training samples with their respected class labels, method to decide the splitting criterion which partitions the instances into respective classes.

Output: A decision tree Method

1. Construct a node N
2. If all the tuples in D are from same class, C then
3. Return N as a leaf node labeled with the class C
4. If attribute list is vacant then

5. Assign N as a leaf node labeled with the most of the class in the data set by mass choice.
6. For best splitting among the attributes apply Attribute_selection_method
7. Label node N with best split method
8. If selected attribute for splitting is discrete value having multiple splits then tree is not limited to binary tree
9. Attribute_list←attribute_list_spliting_attribute;//Remove splitting attribute
10. For every outcome m for the splitting criterion will partition the tuples and sub trees for each partition grow
11. From D if D_m is the set of data tuples which satisfy the outcome m
12. If D_m is empty then
13. For node N assign leaf label with majority of class in D
14. Otherwise attach a node which is returned by generated decision tree(D_m ,attribute list)
15. End loop
16. Return N

- **Random Forest algorithm**

Random Forest is of type ensemble learning classification. Random Forest construct a massive amount of decision trees while training time and yields the class as classification result. Random Forest is a collection of trees, all the trees are slightly different .Random Forest considers a randomly generated subsets from training dataset and build forest of trees. For decision tree building process at each node of a tree some of the variables from variable set are randomly selected and do the best partitioning of the dataset.

Steps performed by Random Forest algorithm are as follows

1. N is the number of training samples, and M is the number of variables .
2. m are the number *of* input variables used to determine the decision at a node of the tree; *m* must be less than *M*.
3. From all N samples training set is selected by replacement with 'n' times. Remaining cases are used to estimate the error.
4. For each tree node, *m* variables are randomly chosen on which decision is based at that node. Based on these *m* variables from the training set best split is calculated.
5. Allow each tree to grow fully.
6. A new sample prediction, label is assigned by taking average vote of all the trees .

Random Forest classification method is popular for giving highly accurate predictions and decision making. Random Forest runs efficiently on large databases also .It handles multiple of input variables. Also it gives estimations about important variables for the classification.

- **Classification using Naïve Bayes**

Naïve Bayes is a machine learning classification algorithm which is simple and probabilistic classifier , based on Bayes theorem. It assumes that the contributions of all attributes within dataset are independent. Here each attribute contributes equally for the classification. Naïve Bayes is based on Bayes rule which depends on conditional probability .Using independent attributes from database conditional probabilities are calculated. Naïve Bayes do the classification by combining the impact of the different attributes on which predictions are measured. Naïve Bayes assumes independence between the various attribute values within dataset.

For data value xi the probability that a associated tuple, t_i is classified in C_j class is given by $P(C_j| x_i)$. $P(x_i)$, $P(C_j)$ and $P (x_i| C_j)$ were calculated by using training dataset. Then the posterior probability $P (C_j | x_i)$ and $P(C_j | t_i)$ were calculated.

For each class in the dataset algorithm first compute the prior probability $P(C_j)$ from the training set. Then $P (x_i)$ is computed by counting number of occurrences of each attribute value x_i for every attribute. $P (x_i | C_j)$ is calculated by considering how frequently each value occur in the class in the training data. It is assume that there may be different attributes having many values in the training data for all attributes values this is performed. These are derived probabilities. When a new tuple is classified we used these derived probabilities. Naive Bayes act as an unambiguous and a predictive kind of performing algorithm. These probabilities are used to predict the class membership for a target tuple and are descriptive.

For the classification of a tuple, the conditional and prior probabilities are generated from the training set and are used for predictions making. Tuple ti has m independent attributes with the values $\{x_i1,x_i2, ……., x_im\}$. From the descriptive phase, $P (x_ik | C_n)$ calculated, for each class Cn and attribute x_ik. And then P (

t_i | C_n) is estimated . Prior probabilities P (C_n) for each class and the conditional probability P (t_i | C_n) were calculated.

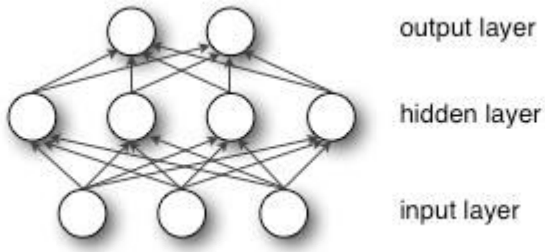
For calculating P (t_i) likelihood estimation is used. Likelihood is ti belong in a class. This is accomplished by finding the likelihood that the tuple is in each class. The probability of ti is belong in a class is calculated by finding product of the conditional probabilities for each attribute value. Posterior probability P(C_n | t_i) for each class is calculated. Then the tuple is classified into a class which is having highest probability.

Naïve Bayes algorithm takes only one scan of training dataset and not required multiple scans. Naïve Bayes approaches easily handles missing values and calculate likelihoods of membership in each class.

- **Multilayer perceptron**

Multi layer perceptron(MLP) is a feed forward artificial neural network. It consist of minimum three layers of nodes, input layer ,hidden layer and output layer. Each node except input node is neuron .It uses non linear activation function.MLP is a supervised learning technique which uses back propagation method for training. The architecture is of multiple layers. It's multiple layers and non linear activation bifurcates MLP from linear propagation. The performance of MLP is mainly depends on the activation function in the Multilayer Perceptron .Hence the selection of activation function is more important.

The computation is performed using set of simple units with weighted connections between them. Furthermore there are learning algorithms to set of values of the weights and the same basic structure having different weight values are able to perform many tasks. There can be multiple hidden layers for MLP.



The above is Artificial Neural Network with a single hidden layer.

One hidden layer MLP is a function $f : R^D \to R^L$, where D is the size of input vector x and L is the size of the output vector $f(x)$, such that, in matrix notation

$$f(x) = G(b^{(2)} + W^{(2)} (s(b^{(1)} + W^{(1)}x)))$$

with bias vectors $b^{(1)}, b^{(2)}$ weight matrices $W^{(1)}, W^{(2)}$and activation functions G and S.

III. PERFORMANCE OF THE CLASSIFIERS AND RESULTS GENERATED

For this study the classifiers have set of features as input values and using these features person is classified into the class they belongs. Classification techniques like Navie Bayes,J48, Random forest, Multi layer perseptron are implemented .Developed classifiers are discussed and used for classification into tested_positive and tested_negative. Result of applying the chosen classifier are tested by using two different testing methods using WEKA tool. Testing methods used in this study are cross validation and percentage split. Different accuracy measures of the classifiers and their classification accuracy is also studied for the comparison.

Diabetics database used in this study contains 200 records. Classifier is then evaluated on how accurately it predicts the class labels for data instances.

Evaluation methods

- **Cross-validation method :**

  This testing approach uses data instances equal number of times for training and testing .Cross validation of k-fold approach is very popular, here k is the number of subsets of the dataset. It is observed in the past researchers study that k= 10 gives best accurate results .

  In this study database is divided into 10 equal folds .Where 9 folds are utilized for training and remaining 1 is utilized for testing for each iteration. Finally average of all iteration is consider.

- **Percentage split :**

Here, database is partitioned into two datasets X% and Y%. X% of the data set is used for training and the remaining Y% data is used for testing. Here database is arbitrarily part into two disjoint datasets. Actually it is difficult to decide what is the proper percentage criterion for splitting but experimentally prove that to keep more data set for training then it is better for predictions and accuracy .WEKA tool has split the database randomly for the mining task into 2 parts. We have split health care database in such a way that 70% section is utilized as training and 30% section is utilized as testing purpose .Once classification model is build it is necessary to measure the performance of the model.

To measure the performance of the classifiers following criterions are used

- **Accuracy:**
  Accuracy of the classifiers is measured in terms of predicting the class labels into correctly classified and incorrectly classified samples.
- **TPR (True Positive Rate) :**
  It measures the actual positive classified instances in classification.
  It is calculated by TPR = TP / (TP + FN)
- **FPR (False Positive Rate) :**
  It measures the expectancy of false positive ratio of instances classified. It is calculated by FPR = FP / (FP + TN) .
  Once the classification model has been trained and tested, then there is a need to gauge the execution of the model. For this precision, recall and accuracy these measures are used.
- **Precision:**
  Precision is a proportion of predicted samples positively which are actual positive. It means positive predicted samples. It is calculated by
  Precision = TP / (TP + FP).
- **Recall :**
  It is a proportion of actual positive samples which are predicted positively. Recall is calculated by Recall= TP / (TP + FN).
- **F-measure :**
  F - measure is a harmonic mean of Precision and Recall. It is a weighted average of the precision and recall . F-measure is calculated by
  F-measure=2(Precision × Recall) / (Precision + Recall) .
- **ROC curve (Receiver Operating Characteristic curve) :**
  It is a curve of true positive rate against the false positive rate at various threshold settings. TP rate is plotted in function of FP rate for different cut-off points of parameter. The ROC is also known as a relative operating characteristic curve, as it is a comparison of two operating characteristics TPR and FPR.

Following are the results generated by applying classification algorithms like Naïve Bayes,J48,Random Forest ,Multilayer perseptron model using cross validation fold and percentage split implementation methods. These results are generated using data mining tool WEKA.

Following tables shows the results generated as per the different performance measuring criterion

Table1 Evaluation method - Cross validation fold

Algorithm	Correctly Classified	Incorrectly Classified	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area	Class
Naïve Bayes	70.5%	29.5%	0.560	0.208	0.618	0.560	0.587	0.785	Tested_positive
			0.792	0.440	0.750	0.792	0.770	0.785	Tested_negative
J48	73%	27%	0.538	0.147	0.737	0.538	0.622	0.761	Tested_positive
			0.864	0.493	0.745	0.864	0.800	0.715	Tested_negative
Random Forest	73.5%	26.5%	0.560	0.160	0.677	0.560	0.613	0.795	Tested_positive
			0.840	0.440	0.761	0.840	0.798	0.795	Tested_negative
Multi Layer Perseptron	65.5%	34.5%	0.427	0.208	0.552	0.427	0.481	0.710	Tested_positive
			0.792	0.573	0.697	0.792	0.742	0.710	Tested_negative

Table1 shows the results of applying data mining classification algorithms by using cross validation evaluation method. Results shows that number of samples correctly classified are more for Random Forest algorithm and are less in case of Multilayer perseptron.

Table2 Evaluation method - Percentage split (70% Training and 30% Testing)

Algorithm	Correctly Classified	Incorrectly Classified	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area	Class
Naïve Bayes	71.66%	29.5%	0.560	0.208	0.618	0.560	0.587	0.785	Tested_positive
			0.792	0.440	0.750	0.792	0.770	0.785	Tested_negative
J48	73%	27%	0.507	0.136	0.691	0.507	0.585	0.715	Tested_positive
			0.864	0.493	0.745	0.864	0.800	0.715	Tested_negative
Multi Layer Perseptron	63.33%	36.66%	0.500	0.265	0.591	0.500	0.542	0.699	Tested_positive
			0.735	0.500	0.658	0.735	0.694	0.699	Tested_negative
Random Forest	76.66%	23.33%	0.692	0.716	0.750	0.692	0.720	0.719	Tested_positive
			0.824	0.308	0.778	0.824	0.800	0.719	Tested_negative

Table2 shows the results of applying data mining classification algorithms by using Percentage split (70% Training and 30% Testing) evaluation method.Results shows that number of samples correctly classified are more for Random Forest algorithm whereas Multilayer perseptron gives less accurate results.

IV. CONCLUSION

Model constructed from the data mining algorithms could help to support decision making in different fields including health care field. The frequency of diabetes is increasing day by day among all age groups of people. This paper focuses that data mining algorithms can be useful in early prediction. Also it is significant for early safety measures before the diagnosis. The main goal of this paper is to study most suitable data mining classification based algorithms and provide a comparison. Based on the comparison this study also suggest best data mining algorithm which can be used for the pattern recognition and prediction in the study of diabetics prediction.

In this study Naïve Bayes, decision tree (j48),Random Forest, Multilayer Perceptron classification algorithms are applied for designing classification model which is use to predict diabetic patients database which consist of 200 records and 8 features values .The risk factors used for the model construction using data mining algorithms are pregnancy count, plasma, Blood pressure(mm Hg), TricepsSkTh, SerumI, BodyMI, DiabetesPF,Age. And last is prediction ie. class value tested _positive and tested_negative.

The algorithms used in this study have importance in medical datasets as these algorithms can be used as regular classification tools .This may help to the doctors or experts in this area for taking essential steps to overcome the disease. Algorithms used in this study can give high accuracy and efficiency based on type of data and features used .The tool used for testing and validation is WEKA .All algorithms are implemented and tested with 70 30 ratio for training and testing and by 10 folds cross validation method.

After the implementations of these data mining algorithms it is observed that for diabetics dataset model developed using Random Forest algorithm have most accurate results than model developed using Decision Tree(J48), Naïve Bayes and Multilayer perseptron .There is no large difference found between accuracies of model developed using Random Forest and J48 algorithms. This research can be utilized to create a control plan for diabetes.

The results shows that the suggested data mining algorithm and model can assist health care experts to make better lending decisions. In future, the results can be used to create a organized plan for diabetes as it has been seen that diabetic patients are normally not acknowledged at the first level and in the later stage of the disease it may become difficult.

References

1. N. Chandra Sekhar Reddy, K. Sai Prasad and A. Mounika ;2017;Classification Algorithms on Datamining: A Study; International Journal of Computational Intelligence Research ISSN 0973-1873 Volume 13, Number 8 (2017), pp. 2135-2142
2. Isha Vashi, Prof. Shailendra Mishra;2016;A Comparative Study of Classification Algorithms for Disease Prediction in Health Care; International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 9.
3. Supreet Kaur, Amanjot Kaur Grewal;2016; A Review Paper On Data Mining Classification Techniques For Detection Of Lung Cancer; International Research Journal of Engineering and Technology (IRJET), Volume: 03 Issue: 11 .
4. Haldurai Lingaraj, Rajmohan Devadass, Vidya Gopi, Kaliraj Palanisamy;2015; prediction of diabetes mellitus using data mining techniques":a review, Journal of Bioinformatics &Cheminformatics.
5. Parvez Ahmad, Saqib Qamar, Syed QasimAfser Rizvi;2015;Techniques of Data Mining in Healthcare : A Review; International Journal of Computer Applications (0975 – 8887) Volume 120 – No.15.
6. Vijayan, V. Veena, and C. Anjali;2015;Prediction and diagnosis of diabetes mellitus—A machine learning approach;Intelligent Computational Systems (RAICS), 2015 IEEE Recent Advances in. IEEE, 2015.
7. Butwall, Mani, and Shraddha Kumar; 2015;A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier; International Journal of Computer Applications.
8. Vivek Agarwal, Saket Thakare, Akshay Jaiswal;2015;Survey on Classification Techniques for Data Mining; International Journal of Computer Applications (0975 – 8887) Volume 132 – No.4.
9. S.Archana , Dr. K.Elangovan;2014; Survey of Classification Techniques in Data Mining‖; International Journal of Computer Science and Mobile Applications, Vol.2 Issue. 2, pg. 65-71 ISSN: 2321-8363.
10. V. krishnaiah, G. Narsimha, & N. Subhash Chandra;2013;A study on clinical prediction using Data Mining techniques; International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR) ISSN 2249-6831 Vol. 3, Issue 1, 239 248.

11. Mohammed Abdul Khalid, Sateesh kumar Pradhan, G.N.Dash, F.A.Mazarbhuiya;2013; A survey of data mining techniques on medical data for finding temporally frequent diseases; International Journal of Advanced Research in Computer and Communication Engineering Vol.2, Issue 12.
12. Delveen Luqman Abd Al.Nabi, Shereen Shukri Ahmed;2013;Survey on Classification Algorithms for Data Mining: (Comparison and Evaluation)(ISSN 2222-2863),Vol.4, No.8.
13. SMs. Aparna Raj, Mrs. Bincy, Mrs. T.Mathu;2012;Survey on Common Data Mining Classification Techniques; International Journal of Wisdom Based Computing, Vol. 2(1).
14. ShwetaKharya;2012;Using Data Mining Techniques ForDiagnosis And Prognosis Of Cancer Disease; International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2.
15. N. AdityaSundar, P. PushpaLatha and M. Rama Chandra;2012;Performance Analysis of Classification Data Mining Techniques Over Heart Disease Data Base;International Journal of Engineering Science & Advanced Technology.
16. scope;Padhy,N.Mishra,P.,Panigrahi,R;2012; The survey of data mining applications and feature ;Computer science,Engineering & information technology,2(3).
17. Patil, B. M., R. C. Joshi, and Durga Toshniwal; 2010;Association rule for classification of type-2 diabetic patients; Machine Learning and Computing (ICMLC) Second International Conference on. IEEE.
18. Han, J., & Kamber, M; 2006; "Data Mining: Concepts and Techniques" (2nd ed.). Morgan Kaufmann Publishers.
19. Barros, R. C., Basgalupp, M. P., Carvalho, A. C., &Freitas, A. A. ;2010; "A Survey of Evolutionary Algorithms for DecisionTree Induction ; IEEE Transactions on Systems,Mans and Cybernetics, Vol. 10,No. 10,pp. 1-22.
20. Thair Nu Phyu ; 2009;Survey of Classification Techniques in Data Mining‖; IMECS,18-20, vol 1,hong kong.