Data Mining Using R Tool

Nalini.N[1], Sri Subhashini.P[2], Suruthi.R[3], Suvetha.K[4]
[1]Assistant Professor, Dept of Computer Science and Engineering
[2, 3, 4]Dept of Computer Science and Engineering
[1, 2, 3, 4] Sri Shakthi Institute of Engineering and Technology, Coimbatore

Abstract- *Data mining is the practice of examining large pre-existing data base in order to generate new information. The overall goal of data mining process is to extract information from a dataset and transform it into an understandable structure for further use. R tools is used to mine the data to make sense out, it is used to predict based on dataset. In this paper we have applied the classification algorithm namely Decision tree, Logistic regression, Support vector machine to compare the accuracy of two different datasets.*

Keywords- Mining, Algorithm, Prediction, Accuracy.

I. INTRODUCTION

Data mining is the process used to extract usable data from a large set of any raw data. It is a cross disciplinary field that focuses on discovering the properties of datasets. R is a language and it has a wide storage capacity. It holds large data set. It has over 4800 inbuilt packages. R provides facilities for analyzing and many operators can be used. There are different types of algorithm used in data mining. By applying these algorithms to the data set we can find the accuracy of the algorithm. The accuracy can be found by using Decision tree algorithm, logistic regression and support vector machine. We need to import the data set into R. Dataset is a collection of related items that can be accessed individually or in combinations. Here we have used two datasets.

In bank dataset we have ten fields (i.e. age, sex, region, income, married, children, car, save act, cur act, mortgage, and pep). In this we have chosen pep as the predictor field. In medical dataset we have eight fields (i.e. glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age and outcome). In this BMI is the predictor field.

II. MINING ALGORITHM

Algorithm is a set of heuristics and calculations that creates a model from data. The different types of algorithm are Classification algorithms, Regression algorithms, Segmentation algorithm, Association algorithms, and Sequence analysis algorithms. Here we have chosen classification algorithm based on our dataset.

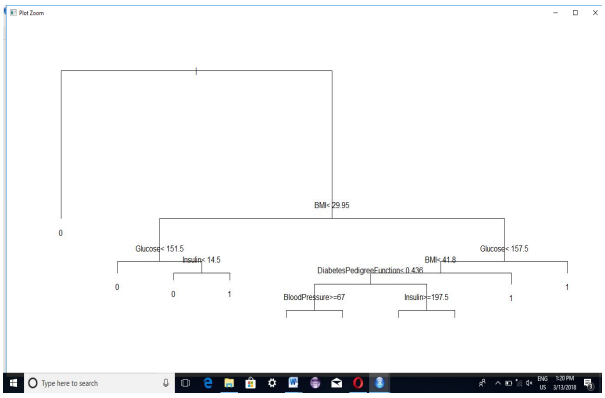
The classification algorithms analyze one or more discrete variable based on other attributes in the dataset. The classification algorithm predicts the category from the input variables. It receives the input data and provides the output. It is one of the supervised task in which desired output will be obtained. The goal is to match the moral output to achieve its target. When the data's are taken we can apply many types of algorithm in the training phase. The different types of algorithm are Decision Tree algorithm, Logistic Regression, Support Vector Machine, Naïve Bayes, Random Forest, K-Nearest Neighbors and Stochastic Gradient Descent.

The accuracy is found by the confusion matrix. The accuracy is found by true positive added with true negative and divided by the total value. True positive is the number of correct predictions that the occurrence is positive. True negative is the number of correct predictions that the occurrence is negative.

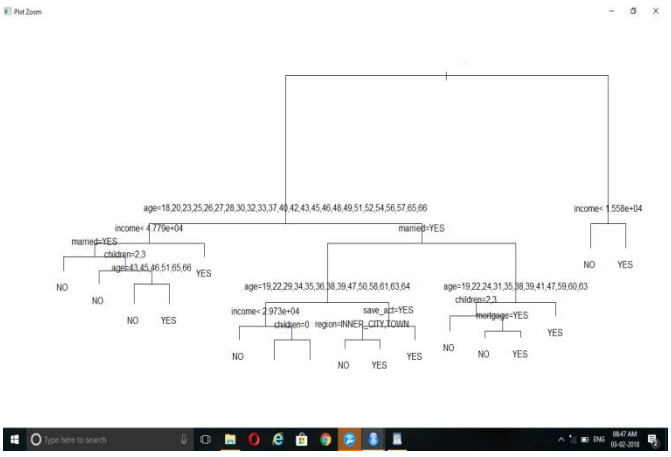
A. Decision tree

Decision tree has a root node, branches and leaf node and it forms structure. In decision tree internal node represents attributes and branches represents outcome and leaf represents label. The top node is the root node. The in-build packages are installed such as library (MASS), library (r part). The numerical values are identified and converted to categorical value. These value are changed by using the as factor function. The data set is split into training data set and test data set. The percentage is fixed to our assumption. By using the training data set we predict the test data set. The accuracy is predicted by using the result column. The confusion matrix will be formed and the diagonals are used to calculate the accuracy.

Decision tree graph (medical dataset)



Decision tree graph (bank dataset)



Advantages

It reduces ambiguity. It is the best predictive model. It is transparent in nature. It shows all possible alternatives in single view. It is easy to compare. It has the ability to assign specific decision tree has a definite conclusion or end values or it needs repetition for new issues.

Disadvantages

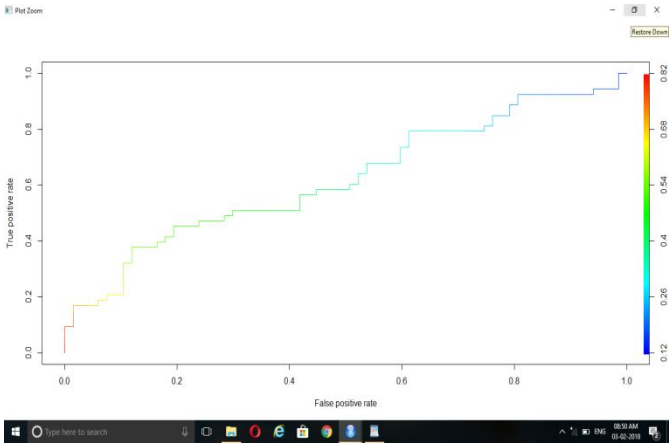
Even a small change in input values it will reflects in large values in the tree. When there is a large data set it has many branches which consume lot of time.

B. Logistic regression

The one or more independent values are analyzed by using the statistical methods. The outcome is based on only two values. We use dummy variables to represent the binary or categorical outcome. It mostly forms an S shape curve and takes values between 0 and 1. Here the build in packages are installed. The data set is imported and the numerical values are stored in a separate object. The data is split into training and test set by the user. Many functions are performed and the S shaped curve is obtained. The value lies between 0 and 1. By using the confusion matrix the accuracy of the algorithm is predicted.

Logistic regression graph (medical dataset)

Logistic regression graph (bank dataset)



Advantages

It is useful for understanding the independent variables on a single outcome variable.

Disadvantages

It works when the predicted value is binary. It assumes all predictors are independent.

C. *Support vector machine*

Support vector machine is data classification method in which the data's are separated by using hyper planes (three dimensions). It is easily understandable by using support vector machine the data can be segmented into two. The separated lines are used to classify the data. There can be more data points that could lie anywhere in the subspace. If the line is too close to the data set, noisy test data gets classified in wrong segment. We need to import the data set and the

packages are installed. The entire data set is stored in an object. The data is split into two as training and test data set. The training data set is predicted by using function. The outcome of the prediction is used to predict the testing data set. The confusion matrix is formed and the accuracy is evaluated.

Advantages

It is effective in high dimensional spaces. It works well with clear margin of separation.

Disadvantages

When the dataset is large it does not perform well. When there is noise in the data the algorithm does not perform well.

III. CONCLUSION

In this paper the various classification algorithm like decision tree, logistic regression and support vector machine are applied to two different datasets namely bank data set and the medical data set to predict the accuracy of the algorithm. The accuracy is predicted using the confusion matrix. The same algorithm has different accuracy for different data sets. The accuracy varies according to the algorithm.

Decision tree confusion matrix

Bank	No	Yes
No	69	5
Yes	4	42

Medical	No	Yes
No	92	26
Yes	13	23

Logistic regression confusion matrix

Bank	No	Yes
No	51	16
Yes	29	24

Medical	No	Yes
No	87	10
Yes	21	35

Support vector confusion matrix

Medical	No	Yes
No	42	32
Yes	39	40

Bank	No	Yes
No	60	21
Yes	10	29

Accuracy table

Algorithm used	Bank data set accuracy	Medical data set accuracy
Decision tree algorithm	75	74
Logistic regression algorithm	76	78
Supportvectormachine algorithm	74	72

REFERENCES

[1] Shwet Kharya,"Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease", International Journal of Computer Science Engineering and Information Technology(IJCSEIT), Vol.2 No2,April 2012

[2] P. Hariharan,“ Design an Disease Predication Application Using Data Mining Techniques for Effective Query Processing Results”, ISSN 0973-6107 Volume 10, Number 3 (2017) pp. 353-361.

[3] Tipawan Silwattananusarn and Dr. KulthidaTuamsuk,” Data Mining and Its Applications for Knowledge Management : A Literature Review from 2007 to 2012”, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.5, September 2012.