Comparative Study of Various Classification Techniques using XLMiner

Ashish Gupta[1], Harish Kumar M[2], Rajeshwari C[3]
[1, 2, 3] Department of information technology and engineering
[1, 2, 3] VIT University, Vellore, India

Abstract- *This paper compares various classification techniques in data mining using XLMiner. XLMiner is a comprehensive data mining add-in for Excel. Data mining is a discovery-driven data analysis technology used for identifying patterns and relationships in data sets. Steps involved- Data Cleaning, Data Integration, Data Selection, Data Transformation, Data mining, pattern evolution, Knowledge evolution. The different types of classification techniques are, Bayes, neural nets, KNN, Discriminant Analysis, and trees etc... This papers shows the accuracy for three different Data Sets using five different techniques namely, Naive Bayes, Neural Networks, Discriminant Analysis and Random Tree.*

Keywords- Data Mining, Data Cleaning, Data Integrating, Data Selection, Data Transformation, Pattern Evolution, Knowledge Evolution.

I. INTRODUCTION

Data mining is a collection of methods to gather data from information and transform into important patterns and guidelines to enhance your understanding. The essential standards of data mining are to break down the information from various bearing, arrange it lastly to condense it. Today we are living in advanced world where information expanding step by step, to get any data from heap of database is troublesome. To manage this colossal information, we require data mining methods. The data mining Process I.e. Selection, Pre-Processing, Transformation, Data Mining, Interpretation and Evaluation are used for performing the mining operation optimally. There are many different tools for performing this process namely, Orange, Clementine, XLMiner, Weka etc. In this paper XLMiner is used for performing the classification operation for different data sets and compare the performance of different classification techniques.

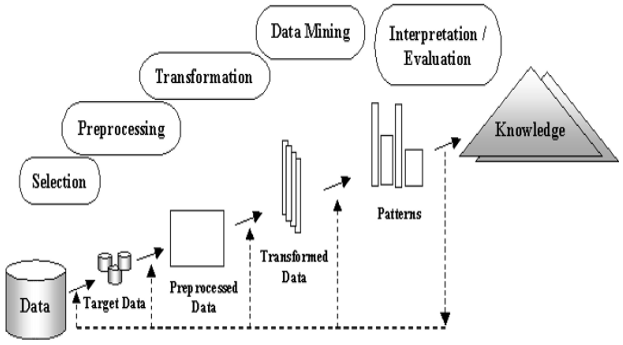

Figure 1. Data Mining Process

II. TOOL - XLMINER

XLMiner is a comprehensive data mining add-in for Excel. The algorithms can either be applied directly to a dataset or called from your own Java code. XLMiner contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

Figure 2. XLMiner View

III. DATA PROCESSING

1. Missing data handling

While pre-processing, we treated missing data that are represented by null values by replacing them through the function of mean, in all the datasets that are categorized by different time duration. After that, we collaborate all the datasets, into 1 dataset, that contains all the features of each location and time period.

2. Converting categorical to numerical

Then we convert the categorical data into numerical data by analyzing the levels and respectively providing labels. So, to perform operations on our dataset. Locations and State attributes have been converted.

3. Converting numerical to Binomial

Taking the water quality label, and specifying min and max value to be true, if the quality is good, else false. So, it'll provide the data into categorical form, in terms of True and False.

4. Feature Selection

	F Test		
Feature Identifier	F-Statistics	F: P-Value	Fisher Score
age	4.8872	0.000890226	0.1003
sex	1.3532	0.251745592	0.0278
cp	3.7226	0.006062748	0.0764
trestbps	1.9275	0.107374782	0.0395
chol	1.0123	0.402191463	0.0208
fbs	0.4883	0.744318637	0.01
restecg	0.8178	0.515178919	0.0168
thalach	0.3615	0.835813116	0.0074
exang	2.3048	0.059764262	0.0473
oldpeak	10.5203	9.73895E-08	0.2158
slope	3.0187	0.019103338	0.0619
ca	N/A	N/A	N/A
thal	0.6691	0.614140753	0.0137

Figure 3. Features of Heart Disease are selected according to F Test Statistical filter for analysis of variance.

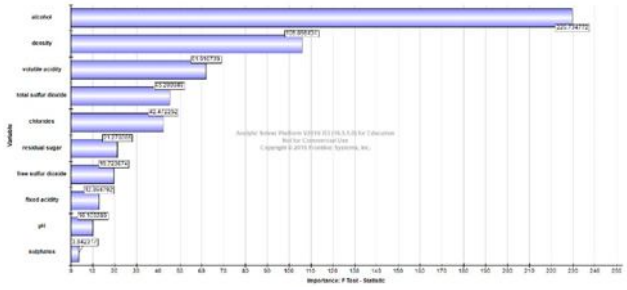

Figure 4. Feature of Wine Quality selection graph

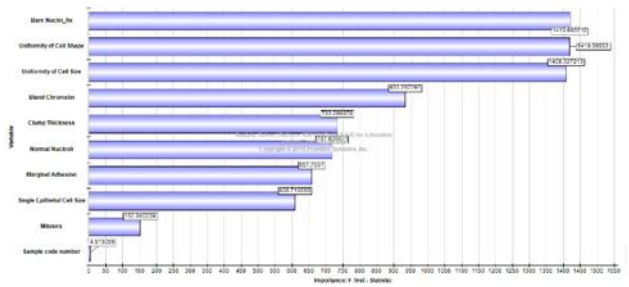

Figure 5. Feature of Breast Cancer selection graph

5. Splitting the dataset into the Training set and Test set

Data partitioning is done based on the build water quality model, that splits the data into 80-20% for training and test set.

6. Fitting classifier to training set

Fitting the data with higher accuracy on the training set. So, target function from the training data generalizes to new data.

7. Predicting the Test set results

Once the model is built, we make predictions on the test set. it extracts the test set feature vector and use the model to make prediction, whether the quality is benign or malign.

8. Visualizing the Training set results

Based on the above predicted feature vector, we make a grid set using the sequence of features providing them column names, and plotting out the results.

IV. CLASSIFICATION

Classification is a data mining algorithm hat creates a step by step guide for how to determine the output of a new data instance. The tree it creates is exactly that: a tree whereby each node in the tree represents a spot where a decision must be made based on the input, and you move to the next node and the next until you reach a leaf that tells you the predicted output. Classification is a data mining (machine learning) technique used to predict group membership for data instances.

1. Naive Bayes

The Bayesian Classification speaks to a directed learning strategy and also a Statistical technique for classification. It uses the knowledge of prior events to predict future events. It is the fastest on comparing to other classification algorithms. It gives functional learning algorithms and prior information and observed data can be combined. Bayesian Classification gives a valuable point of view to understand and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data. Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). The below eauation shows that:

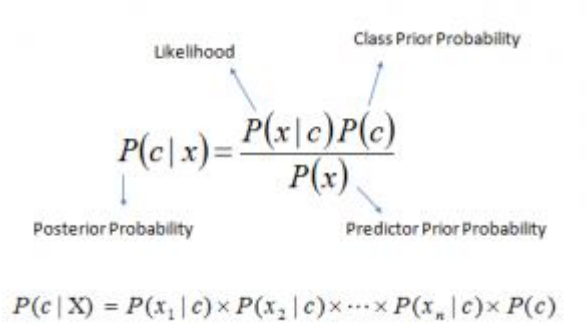

Figure 6.

- P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

2. Neural Network

Neural networks is based on a large collection of neural units (artificial neurons), loosely mimicking the way a biological brain solves problems with large clusters of biological neurons connected by axons. We have used Boosting Neural networks. Boosting is a general method for improving the performance of learning algorithms. A recently proposed boosting algorithm is AdaBoost. It has been applied with great success to several benchmark machine learning problems using mainly decision trees as base classifiers.

3. K-Nearest Neighbour

K-NEAREST NEIGHBOR (KNN) is a non-parametric LAZY learning algorithm. KNN is one of those algorithms that are very simple to understand but works incredibly well in practice. It is surprisingly versatile and its applications range from vision to proteins to computational geometry to graphs. Using of KNN make things very simple. It is one of the top 10 data mining algorithms.

4. Random Tree

Random trees is a collection (ensemble) of tree predictors that is called forest further. The clasification works as follows: the random trees classifier takes the input feature vector, classifies it with every tree in the forest, and outputs the class label that received the majority of "votes". In case of a regression, the classifier response is the average of the responses over all the trees in the forest.

5. Discriminant Analysis

Discriminant function analysis is a statistical analysis to predict a categorical dependent variable (called a grouping variable) by one or more continuous or binary independent variables (called predictor variables). Discriminant function analysis is useful in determining whether a set of variables is effective in predicting category membership. Discriminant analysis is used when groups are known a priori (unlike in cluster analysis). Each case must have a score on one or more quantitative predictor measures, and a score on a group measure. In simple terms, discriminant function analysis is classification - the act of distributing things into groups, classes or categories of the same type.

V. RESULTS AND DISCUSSIONS

Table 1. Performance compression of different classification techniques (in %)

Dataset	Naïve Bayes	Neural Network	KNN	Discriminant Analysis	Random Tree
Breast Cancer	96.785	95.995	93.571	95.708	95
Heart Disease	70	34.5	35	38.5	22.5
Wine Quality	97.96	44.88	53.497	53.778	55.88

In Preprocessing, we treated missing data that are represented by null values by replacing them with median or mode, that provide us with data cleansing, then we have performed feature selection, where we have selected the most important features by taking all the attributes as input and classifier as output variable. here all data is classification, so we have categorical output variable type. then we got the feature importance plot based on F Test- Statistic. from that we have deselected out the most important features based on the ranking, and done standard data partitioning, where we pickup rows randomly with partition percentage of training set as 60% and validation set as 40%. the output we have as the data partitioned.

Now, we will apply classification techniques, we have used discriminant analysis, K-Nearest neighbour, Classification Tree- Random Tree, Naive Bayes, and neural network boosting algorithm.

In each classification, we have data range, no. of columns, rows in training set and validation set.

Then we normalize the input data, wherever required. and we got the result as scoring training data and scoring validation data, where we got the detailed report and summary report. the outcome is confusion matrix, Error rate, and performance.

The results measuring the accuracy performance of each dataset with different classification techniques.

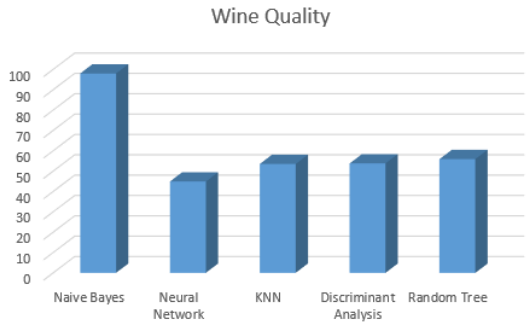

Figure 7. performance graph on wine quality data set

q

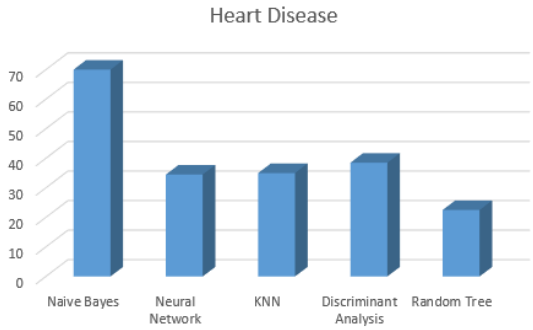

Figure 8. performance graph on heart disease data set

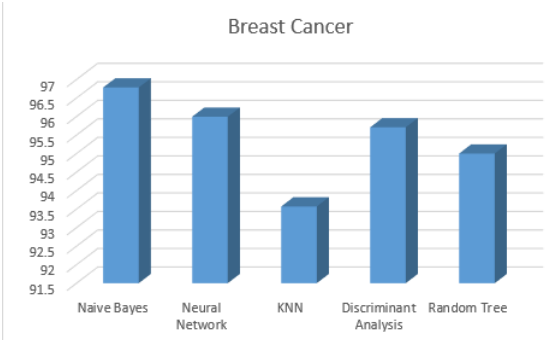

Figure 9. performance graph on breast cancer data set

VI. CONCLUSION

XLMiner is a simple tool for using the different techniques of data mining. We have used tool for classification of data sets and compared the performance of different classification algorithms. According to results we got the naïve bayes as highly accurate for classification techniques on small and mid-range of datasets and also neural networks perform well with 2 or more number of nets in it.

REFERENCES

[1] Polaka and a. Sukov, Using the xlminer tool for data mining in customer Relationship management, Scientific proceedings of riga technical university (2004)

[2] Manish varma d, Data mining classification techniques applied to Analyze the impact of ambient conditions on aero, Engine performance - a case study using xlminer (2010)

[3] Department of computer science engineering Manipal institute of technology, manipal

[4] Volume 32, Issue 6, August 2006, Pages 733–742 Analyzing association rule mining and clustering on sales day Data with xlminer and weka, International journal of database theory and application Vol. 3, no. 1, march, 2010, A M. Khattak, a. M. Khan, sungyoung lee*, and Young-koo lee.

[5] C.v.subbulakshmil, s.n.deepa2, n.malathi3, Comparative analysis of xlminer and weka for pattern classification (2012)

[6] K.srinivas b.kavihta rani dr. A.govrdhan, Applications of data mining techniques in Healthcare and prediction of heart attacks K.srinivas et al. / (ijcse) international journal on computer science and engineering Vol. 02, no. 02, 2010, 250-255

[7] G´erard biau, Analysis of a random forests model Journal of machine learning research 13 (2012) 1063-1095

[8] Random forests and decision trees Jehad ali1, rehanullah khan2, nasir ahmad3, imran maqsood4, Ijcsi international journal of computer science issues, vol. 9, issue 5, no 3, september 2012

[9] A survey on decision tree algorithm For classification 2014 ijedr | volume 2, issue 1 | issn: 2321-9939, 1mr. Brijain r patel, 2mr. Kushik k rana