IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 10, Issue 6 (May. - Jun. 2013), PP 01-06
www.iosrjournals.org

Performance Evaluation of Different Data Mining Classification Algorithm and Predictive Analysis

Syeda Farha Shazmeen[1], Mirza Mustafa Ali Baig[2], M.Reena Pawar[3]
[1, 2, 3] *Department of Information Technology, Balaji Institute of Technology and Science, Warangal, A.P, India,*

Abstract: *Data mining is the knowledge discovery process by analyzing the large volumes of data from various perspectives and summarizing it into useful information; data mining has become an essential component in various fields of human life. It is used to identify hidden patterns in a large data set. Classification techniques are supervised learning techniques that classify data item into predefined class label. It is one of the most useful techniques in data mining to build classification models from an input data set; these techniques commonly build models that are used to predict future data trends. In this paper we have worked with different data mining applications and various classification algorithms, these algorithms have been applied on different dataset to find out the efficiency of the algorithm and improve the performance by applying data preprocessing techniques and feature selection and also prediction of new class labels.*
Keywords: *Classification, Mining Techniques, Algorithms.*

I. Introduction

Data miming consists of a set of techniques that can be used to extract relevant and interesting knowledge from data. Data mining has several tasks such as association rule mining, classification and prediction, and clustering. Classification [1] is one of the most useful techniques in data mining to build classification models from an input data set. The used classification techniques commonly build models that are used to predict future data trends [2, 3].
Model construction: describing a set of predetermined classes

1. Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute
2. The set of tuples used for model construction: training set.
3. The model is represented as classification rules, decision trees, or mathematical formulae.
4. Accuracy rate is the percentage of test set samples that are correctly classified by the model.
5. Test set is independent of training set, otherwise over-fitting will occur.

The goal of the predictive models is to construct a model by using the results of the known data and is to predict the results of unknown data sets by using the constructed model [13].
In the classification and regression models the following techniques are mainly used
1. Decision Trees;
2. Artificial Neural Networks
3. Support Vector Machine
4. K-Nearest Neighbor
5. Navie-Bayes.

Before classification, data preprocessing techniques [4, 5] like data cleaning, data selection is applied this techniques improve the efficiency of the algorithm to classify the data correctly.

II. Data Sets used in the application:

IRIS Datasets: -we make use of a large database 'Fisher's Iris Dataset' containing 5 attributes and 150 instances. It consists of 50 samples from each of three species of Iris flowers (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample; they are the length and the width of sepal and petal, in centimeters. Based on the combination of the four features, Fisher developed a linear discriminant model to distinguish the species from each other classification method to identify the class of Iris flower as Iris-setosa, Iris-versicolor or Iris-virginica using data mining classification algorithm.

Liver Disorder: -The observations in the dataset consist of 7 variables and 345 observed instances. The first 5 variables are measurements taken by blood tests that are thought to be sensitive to liver disorders and might arise from excessive alcohol consumption. The sixth variable is a sort of selector variable. The subjects are single male individuals. The seventh variable is a selector on the dataset, being used to split it into two sets,

indicating the class identity. Among all the observations, there are 145 people belonging to the liver-disorder group (corresponding to selector number 2) and 200 people belonging to the liver-normal group.

E-coli: - (Escherichia coli commonly abbreviated E. coli) is a Gram-negative, rod-shaped bacterium that is commonly found in the lower intestine of warm-blooded organisms (endotherms). This dataset is consisting of 8 attributes and 336 instances. The class distribution of 8 attributes are classified into 8 classes

cp (cytoplasm)	143
im (inner membrane without signal sequence)	77
pp (perisplasm)	52
imU (inner membrane, uncleavable signal sequence)	35
om (outer membrane)	20
omL (outer membrane lipoprotein)	5
imL (inner membrane lipoprotein)	2
imS (inner membrane, cleavable signal sequence)	2

III. Classification algorithm:

Differentiation classification algorithms [6, 7] have been used for the performance evaluation, below are listed.

1: -j48 (C4.5): J48 is an implementation of C4.5 [8] that builds decision trees from a set of training data in the same way as ID3, using the concept of Information Entropy. The training data is a set S = s1, s2... of already classified samples. Each sample si = x1, x2... is a vector where x1, x2… represent attributes or features of the sample. Decision tree are efficient to use and display good accuracy for large amount of data. At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other.

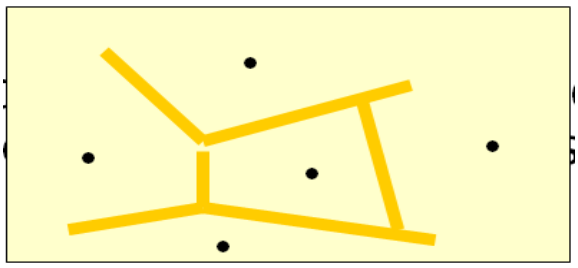
2: -**Naive Bayes: -**a naive Bayes classifier assumes that the presence or absence of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. Bayesian belief networks are graphical models, which unlikely naive Bayesian classifier; allow the representation of dependencies among subsets of attributes[10]. Bayesian belief networks can also be used for classification. A simplified assumption: attributes are conditionally independent:

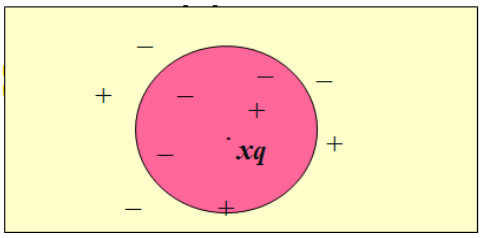
$$P(C_j|V) \propto P(C_j)\prod_{i=1}^{n} P(v_i|C_j)$$

Where V are the data samples, v_i is the value of attribute i on the sample and $C_j$is the j-th class. Greatly reduces the computation cost, only count the class distribution.

3: - k-nearest neighborhood:-

The k-NN algorithm for continuous-valued target functions Calculate the mean values of the *k* nearest neighbors Distance-weighted nearest neighbor algorithm Weight the contribution of each of the k neighbors according to their distance to the query point x_q g giving greater weight to closer neighbors Similarly, for real-valued target functions. Robust to noisy data by averaging k-nearest neighbors.



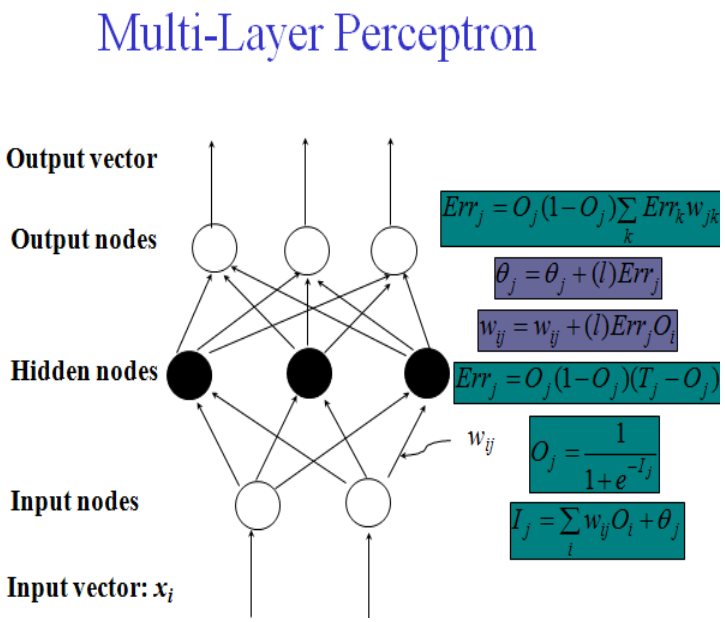


The Euclidean distance between two points, X=(x_1, x2,.....x_n) and Y=(y_1, y2,.....y_n) is

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

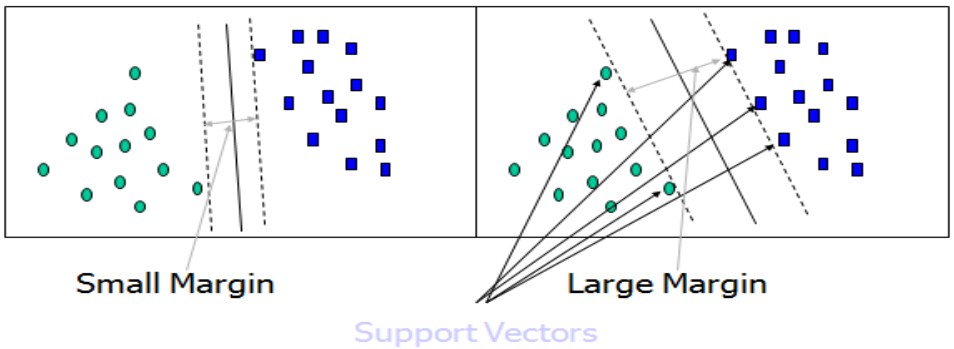
4: -Neural Network:-

Neural networks have emerged as an important tool for classification. The recent vast research activities in neural classification have established that neural networks are a promising alternative to various conventional classification methods. The advantage of neural networks lies in the following theoretical aspects. First, neural networks [9] are data driven self-adaptive methods in that they can adjust themselves to the data without any explicit specification of functional or distributional form for the underlying model.



5: -Support Vector Machine: -

A new classification method for both linear and non linear data.It uses a nonlinear mapping to transform the original training data into a higher dimension. With the new dimension, it searches for the linear optimal separating hyper plane (i.e., "decision boundary"). With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyper plane SVM finds this hyper plane using support vectors ("essential" training tuples) and margins (defined by the support vectors).



IV. Implementation

The Following tools and technologies used for the Experimentation is Rapid Miner [14] is an open source-learning environment for data mining and machine learning. This environment can be used to extract meaning from a dataset. There are hundreds of machine learning operators to choose from, helpful pre and post processing operators, descriptive graphic visualizations, and many other features. It is available as a stand-alone application for data analysis and as a data-mining engine for the integration into own products.

Weka[15] is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well suited for developing new machine learning schemes. Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query.

JFreeChart [16] is an open-source framework for the programming language Java; it is an open source library

available for Java that allows users to easily generate graphs and charts. It is particularly effective for when a user needs to regenerate graphs that change on a frequent basis. JFreeChart supports pie charts (2D and 3D), bar charts, line charts, scatter plots, time series charts, and high-low-open- close charts.

V. Experiment and Results:

The different algorithms are applied to different dataset and the performance is evaluated [11, 12]. The information gain measure is used to select the test attribute at each node in the tree. Such a measure is referred to as an attribute selection measure or a measure of the goodness of split. The attribute with the highest information gain is chosen as the test attribute for the current node. This attribute minimizes the information needed to classify the samples in the resulting partitions and reflects the least randomness or "impurity" in these partitions.

Information gain

If a set S of records is partitioned into classes C1, C2, C3. . . Cion the basis of the categorical attribute, then the information needed to identify the class of an element of S is denoted by:

$I(S) = -(p_1(1 \log_1 2 \ [p1 +]\ ) p_1(2 \log_1 2 p2) + p_1(3 \log_1 2 p3)$ .....

$p_1(i \log_1 2 pi))$

Where pis the probability distribution of the partition Ci.Thus Entropy can be written like this

$$entropy\ E(A) = \sum_{i=1}^{n} \frac{|S_i|}{|S|} I(S_i)$$

Thus the information gain in performing a branching with attribute A can be calculated with this equation.
Gain (A) =I (S)-E (A)
After the preprocessing of data, it comes computing the information gain necessary to construct the decision tree. After these information gains are computed, the decision tree is constructed.

Gain (sepal length) =0.44

Gain (sepal width) =0.0

Gain (petal length) =1

Gain (petal width) =1

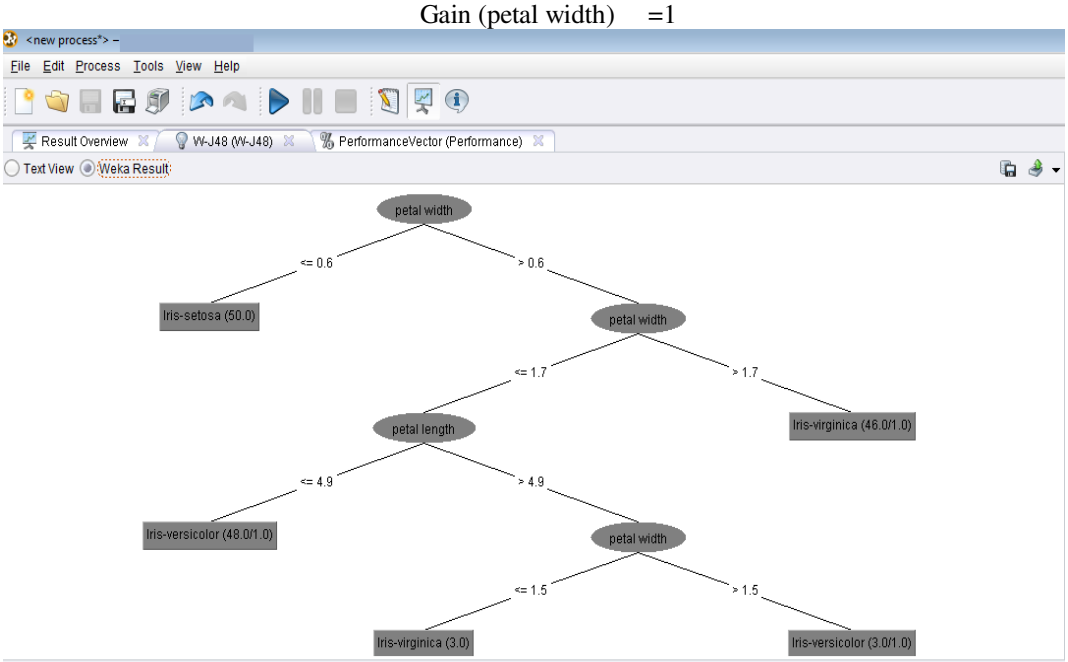


Fig: 1: -J-48 Decision tree for iris dataset

The feature selection can apply to remove the zero weight attributes; by removing those attributes the performance can be improved. The most useful part of this is attribute selection.

1. Select relevant attributes
2. Remove redundant and/or irrelevant attributes.

The screen shots and results are displayed below.

Datasets	Classification	Performance
Iris	J48	95.33%
	Neural net	97.33%
E-coil	Decision tree	80.06%
	Neural net	83.01%
Liver Disorder	Neural net	67.59%
	J48	69.58%

Table1: Classification results by using different Algorithms

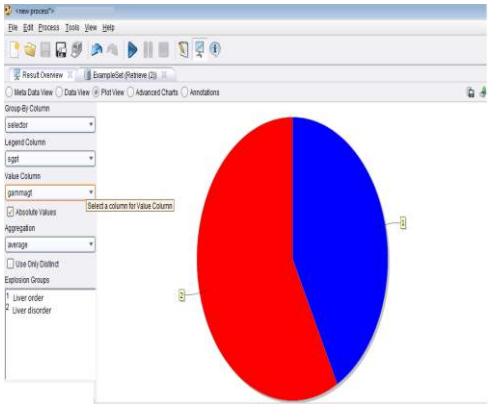


Fig: 2: -Classification result Pie chart for liver Disorder dataset.

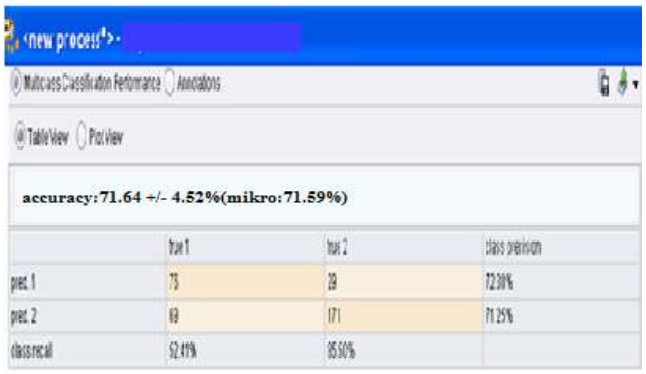


Fig: 3:-Improved classification result for liver disorder Using neural net with removal of zero weight

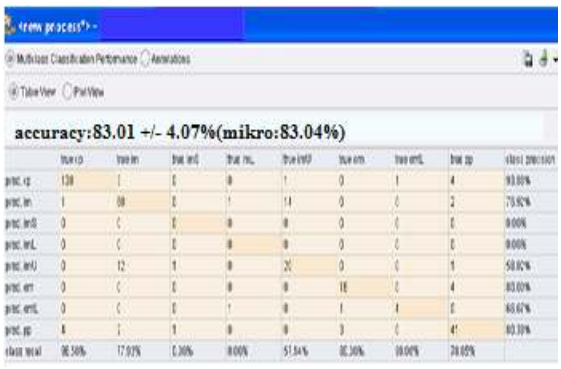


Fig: 4: -Classification result for e-coli dataset using neural net

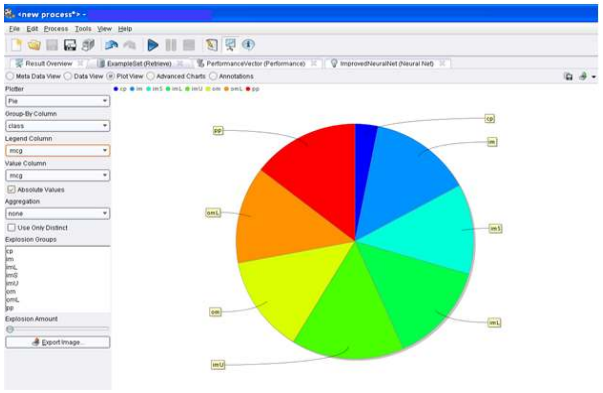


Fig: 5: -pie chart of e-coil dataset

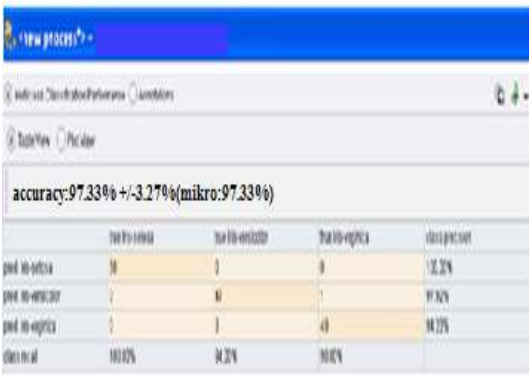


Fig: 6: -Classification result for iris dataset using neural net

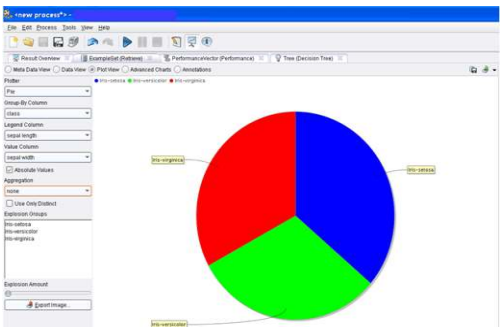


Fig: 7: -Pie chart for iris dataset

Predictive analysis can be used to predict the new class label. By using the different algorithms but Neural Network and K-Nearest Neighborhood was the efficient.

Datasets	Classification	Performance Of Prediction
Iris	Naive Bayes	Incorrect
	Neural net	Correct
E-coil	SVM	Incorrect
	KNN	Correct
Liver Disorder	Decision tree	Incorrect
	KNN	Correct

Table2. Displays the Predictions of new class label

VI. Conclusion:

Classification is one of the most useful techniques in data mining to build classification models from an input data set. Choosing the right data mining technique is another important fact to overcome the problems. The results of different classification algorithm are compared and to find which algorithm generates the effective results and graphically displays the results. Compare the performance of the different classification algorithms when applied on different datasets.

References:

[1]. SERHAT ÖZEKES and A.YILMAZ ÇAMURCU:" CLASSIFICATION AND PREDICTION IN A DATA MINING APPLICATION "Journal of Marmara for Pure and Applied Sciences, 18 (2002) 159-174 Marmara University, Printed in Turkey.

[2]. Kaushik H and Raviya Biren Gajjar ,"Performance Evaluation of Different Data Mining Classification Algorithm Using WEKA",Indian Journal of Research(PARIPEX) Volume : 2 | Issue : 1 | January 2013 ISSN - 2250-1991.

[3]. WaiHoAu,KeithC.C.Chan;XinYao.ANovelEvolutionaryDataMiningAlgorithmwithApplicationstoChurn Prediction. IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, Vol. 7, No. 6, Dec 2003, PP: 532- 545

[4]. Reza Allahyari Soeini and Keyvan Vahidy Rodpysh: "Evaluations of Data Mining Methods in Order to Provide the Optimum Method for Customer Churn Prediction: Case Study Insurance Industry."2012 International Conference on Information and Computer Applications (ICICA 2012)IPCSI vol. 24 (2012) © (2012) IACSIT Press, Singapore.

[5]. Surjeet Kumar Yadav and Saurabh Pal:" Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification"World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 2, No. 2, 51-56, 2012.

[6]. Zhong, N.; Zhou, L.: "Methodologies for Knowledge Discovery and Data Mining", The Third Pacific-Asia Conference, Pakdd-99, Beijing, China, April 26-28, 1999; Proceedings, Springer Verlag, (**1999**).

[7]. Raj Kumar, Dr. Rajesh Verma:" Classification Algorithms for Data Mining: A Survey"International Journal of Innovations in Engineering and Technology (IJIET)

[8]. Fayyad, U.: "Mining Databases: Towards Algorithms for Knowledge Discovery", IEEE Bulletin of the Technical Committee on Data Engineering, 21 (1) **(1998)** 41-48.

[9]. Ling Liu:" From Data Privacy to Location Privacy: Models and Algorithms"September 23-28, 2007, Vienna, Austria.

[10]. Qi Li and Donald W. Tufts:,"Principal Feature Classification " IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 8, NO. 1, JANUARY 1997.

[11]. Deen, Ahmad **,** classification based on association-rule mining techniques: a general survey and empirical comparative evaluation **,** Ubiquitous Computing and Communication Journal vol 5 no 3

[12]. Sanjay D. Sawaitul, Prof. K. P. Wagh, Dr. P. N. Chatur," Classification and Prediction of Future Weather by using Back Propagation Algorithm-An Approach",International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 1, January 2012)

[13]. Qasem A. Al-Radaideh & Eman Al Nagi," Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance",(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 2, 2012.

[14]. RapidMiner is an open source-learning environment for data mining and machine learning. http://rapid-i.com/content/view/181/190/

[15]. Weka is a collection of machine learning algorithms for data mining tasks http://www.cs.waikato.ac.nz/ml/weka/downloading.html

[16] JFreeChart, JFreeChart is an open source library available for Java that allows users to easily generate graphs and charts-http://www.jfree.org/index.html