Comparative Analysis of classification Algorithms using WEKA tool

Shivangi Gupta[1]
Shivg5994@gmail.com

Neeta Verma[2]
neeta140@gmail.com

[1,2]Inderprastha Engineering College Ghaziabad U.P INDIA

Abstract— To group the similar objects together and separate the dissimilar ones are done by using the data mining techniques.. Clustering is one of the unsupervised technique of Data Mining. Each object in the data set is assigned a class label in the clustering process using a distance measure. This paper explains the analysis of classification and clustering using some terms like Kappa Statistics ,Mean Absolute Error , Confusion Matrix ,Classification Accuracy correctly classified, incorrectly classified ,root mean square error for different algorithms of classification and clustering. This paper considers the most extensively used tools ,WEKA tool for this analysis purpose. The training and testing is performed for this analysis .There exist several validation indexes for training testing the performance and accuracy which have also been discussed here.

Index Terms— Clustering ,Classification, data mining, knowledge discovery, WEKA tool, mean absolute error, confusion matrix, kappa statistics.

1. Introduction

Data Mining (DM) discovers hidden relationships in data, in fact it is part of a wider process called "knowledge discovery". Knowledge discovery describes the phases which should be done to ensure reaching meaningful results through research. The objective of DM process is to obtain information out of a dataset and convert it into a comprehensible outline. An understanding of algorithms is combined with detailed knowledge of the datasets. Data mining must afford very complex and different situations to reach quality solutions. Therefore, data mining is a research field where many advances are being done to accommodate and solve emerging problems [7] Data sets in weka are validation, training and test set. In this paper weka tool is used to analyze the performance of classification and clustering algorithms. The performance of classification is analyzed using classified instances, error rate, and kappa statistics. Weka is a machine learning tool which complements data mining. Four machine learners, Neural Network, Support Vector Machine, Logistic, and 3-Nearest Neighbor, were the top performers with over 98% accuracy rates[1].Data mining has been obtained a great attention in the knowledge and information industry due to the vast availability of large amounts of data and the forthcoming need for converting such data into meaningful information and knowledge[8].

2. Clustering

Clustering is an unsupervised technique of Data Mining. It means grouping similar objects together and separating the dissimilar ones[2].To group a set of objects in such a way that similar objects are in the same group ,Cluster analysis one of the useful technique .This technique can be used in many fields, including machine learning, pattern recognition ,image analysis, information retrieval, bioinformatics, data compression. The computational complexity of the k-Means algorithm with LC. arff dataset is better than that of FarthestFirst algorithm for both of the dataset [3]. The k-Means algorithm is efficient than hierarchical clustering builds models based on distance connectivity .The k-means algorithm represents each cluster by a single mean vector is a centroid model .Clusters are modeled using statistical distributions, such as multivariate normal distributions used by the Expectation-maximization algorithm under the category of distributed model.The clusters are connected as dense regions in the data space. The DBSCAN and OPTICS are under the category of density model.

2.1 Hierarchical Clustering

Hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types.
Agglomerative: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.The greedy manner is used to merges and splits the clusters .In the general case, the complexity of agglomerative clustering is, which makes them too slow for large data sets. Divisive clustering with an exhaustive search is, which is even worse.

2.2 K- Means clustering

k-means is one of the simple unsupervised learning algorithms that solve the well known clustering problem..The main idea is to define k centers, one for each cluster. This algorithm aims at minimizing an objective function know as squared error function given by:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \quad (1)$$

where,

'$||x_i - v_j||$' is the Euclidean distance between x_i and v_j.

'c_i' is the number of data points in i^{th} cluster.

'c' is the number of cluster centers.

3. Classification

Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome. Next the algorithm is given a data set not seen before, called prediction set, which contains the same set of attributes, except for the prediction attribute – not yet known. The algorithm analyses the input and produces a prediction. The prediction accuracy defines how "good" the algorithm is.
Classification may refer specifically to:

- *Statistical classification*, identifying to which of a set of categories a new observation belongs, on the basis of a training set of data.
- *Mathematical classification*, a collection of sets which can be unambiguously defined by a property that all its members share Classification theorems in mathematics.
- *Attribute-value system*, a basic knowledge representation framework.

3.1 Naive Bayes Classifier

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. For some types of probability models, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods.
In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.
In Naïve Bayes classifier attributes are conditionally independent . This greatly reduces the computation cost. It counts only the class distribution.
There are *m* classes C1, C2… Cm. With tuples **X** = (x1, x2…xn), The Classification of such classes is derived using the maximum posteriori, i.e., the maximal P (Ci|**X**). This canbe derived from Baye's theorem . P(X) delete constant for all classes, only needs to be maximized. The goal of this classification is to correctly predict the value of a designated discrete class variable given a vector of attribute using 10 fold cross validation . Naïve Bayes classifier is applied to trained and test set and the performance is evaluated individually with kappa statistics, error rate.

4. Data Used

In this paper the dataset for 1984 United States Congressional Voting Records is used for analysis. This data set includes votes for each of the U.S. House of Representatives Congressmen on the 16 key votes identified by the CQA. The CQA lists nine different types of votes: voted for, paired for, and announced for (these three simplified to yea), voted against, paired against, and announced against (these three simplified to nay), voted present, voted present to avoid conflict of interest, and did not vote or otherwise make a position known (these three simplified to an unknown disposition)[11].

5.WEKA Tool

Waikato Environment for Knowledge Analysis (Weka) is a popular suite of machine learning software written in Java. The original non-Java version of Weka was a Tcl/Tk front-end to (mostly third-party) modeling algorithms implemented in other programming languages, plus data preprocessing utilities in C, and a Makefile-based system for running machine learning experiments. This original version was primarily designed as a tool for analyzing data from agricultural domains, but the more recent fully Java-based version (Weka 3), for which development started in 1997, is now used in many different application areas, in particular for educational purposes and research. Advantages of Weka include: WEKA has been a resounding success which we believe has significantly advanced the application of machine learning techniques in today's world.[4]
The Data Mining is the process of extracting information from large data sets through different techniques .Data Mining, popularly called as knowledge discovery is a work bench that contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions. In this paper, we have used WEKA, a Data Mining tool for classification techniques. Weka provides the required data mining functions and methodologies. The data format for WEKA is MS Excel and ARFF formats respectively. Weka a machine learning workbench implements algorithms for data preprocessing, classification, regression, clustering and association rules . Implementation in weka is classified as:
1. Implementation scheme for classification;
2. Implementation schemes for numeric prediction;
3. Implemented meta-schemes.
Learning methods in weka are called classifiers which contain tunable parameters that can be accessed through a

property sheet or object editor.

Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform. A comprehensive collection of data preprocessing and modeling techniques .Ease of use due to its graphical user interfaces.

Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Weka is a data mining tools.[5]
WEKA implements algorithms for data preprocessing, classification, regression, clustering and association rules; It also includes visualization tools. The new machine learning schemes can also be developed with this package[9]
Weka is a machine learning tool which complements data mining. An understanding of algorithms is combined with detailed knowledge of the datasets. Data sets in weka are validation, training and test set.[10]

6. Performance Evaluations

6.1 Mean Absolute Error

The mean absolute error (MAE) is a quantity used to measure predictions of the eventual outcomes. The mean absolute error is an average of the absolute errors.
The mean absolute error is given by

$$\mathrm{MAE} = \frac{\mathrm{SAE}}{N} = \frac{\sum_{i=1}^{N} |x_i - \hat{x}_i|}{N} \quad (1)$$

Where:

$\{x_i\}$ is the actual observations time series
$\{\hat{x}_i\}$ is the estimated or forecasted time series

	Correctly classified	Incorrectly classified	Mean absolute error	Root mean error	Kappa statistic
Training	90.3448%	9.6552%	.0975	.2944	.7999
Testing	91.2162%	8.7838%	.0912	.2858	.8232

SAE is the sum of the absolute errors (or deviations)
N is the number of non-missing data points

6.2 Root mean squared error

Root mean squared error is the square root of the mean of the squares of the values. It squares the errors before they are averaged [18] and RMSE gives a relatively high weight to large errors.
The RMSE Ei of an individual program *i* is evaluated by the equation:

The RMSE of a model prediction with respect to the estimated variable X_{model} is defined as the square root of the mean squared error:

$$RMSE = \sqrt{\frac{\sum (X_{obs,i} - X_{mo\ del,i})^2}{n}} \quad (2)$$

where X_{obs} is observed values and X_{model} is modelled values at time/place *i*.

	TP Rate	FP Rate	Precision	F-Measure
Democrat	0.891	0.077	0.984	0.919
Republican	0.923	0.109	0.842	0.881
Weighted Average	0.903	0.089	0.907	0.904

Training a data set generally mininmizes the error rate for test set. Error rate for training set is comparatively higher than that of the test set. From the above diagram. If any two algorithm has the same mean absolute error rate then root mean squared error rate is taken into consideration for choosing the best classification algorithm.

Testing set has low error rate than the training data set. It is clear from the above diagram for the animal kingdom test set that Naive Bayes classifier has the lowest mean absolute error rate.

(1)

6.3 Confusion Matrix Classification Accuracy

Classification accuracy is the degree of correctness in classification. The degree of correctness can be evaluated using various classifiers for individual instances in the data set. The Larger the training set and the higher the classifier accuracy is ; the smaller the test set and the lesser the classifier accuracy is. Similarly larger test set provides a good assessment on classifier accuracy . In this paper United States Congressional Voting Records training set is higher than the test set which gives higher accuracy rate. Training set contains 60% of the whole data set an the remaining is used as test set for classification .

7. Results

Table 1.1 Anaylysis of correctly and incoerrectly data

The table 1.1 shows the anaylysis of correctly and incoerrectly data results. The mean absolute ,root mean square, Kappa statistics are calculated for training and testing data.

Table 1.2 Classification results of training data

Table 1.2 shows the classification of training data for different

classess on TP rate, FP rate, Precision, F-measure.

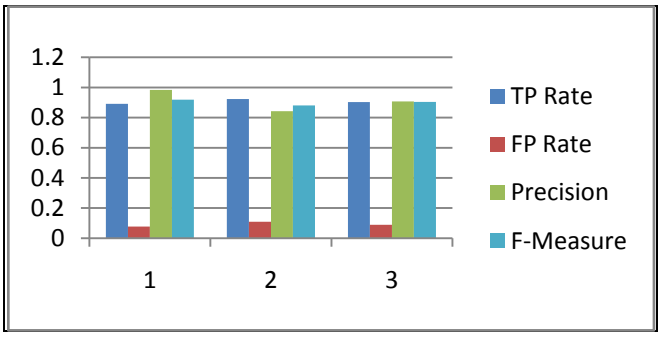

Figure 1 Analysis of Democrat, Republican and weighted average classess for training data

The analysis of Democrat, Republican and weighted average on the basis of TP rate, FP rate ,precision and F-Measure parameters are done.

	TP Rate	FP Rate	Precision	F-Measure
Democrat	0.872	0.032	0.974	0.92
Republican	0.968	0.128	0.845	0.902
Weighted Average	0.912	0.072	0.92	0.913

Table 1.3 Classification results of testing data

Table 1.3 shows the classification of testing data for different classes on TP rate, FP rate, Precision, F-measure.

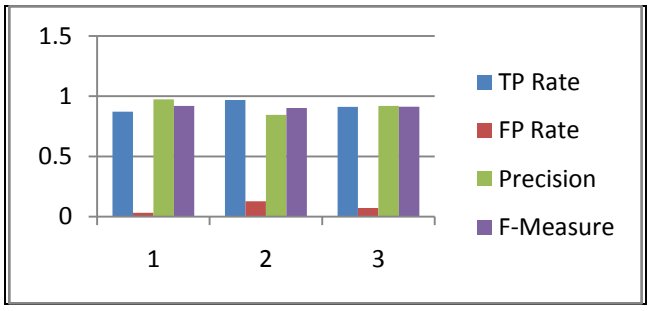

Figure 2 Analysis of Democrat, Republican and weighted average classess for testing data

The analysis of Democrat, Republican and weighted average on the basis of TP rate, FP rate ,precision and F-Measure parameters are done.

8. Conclusion

we evaluate the performance using naïve Bayes classifier for training and test data. The results are analysed using various accuracy measures like TP rate, FP rate, Precision, , F-measure .Mean absolute error and root mean error is less is case of training data. we have met our objective which is to evaluate and investigate classification algorithm based on Weka. These results suggest that among the machine learning algorithm , Naïve Bayes classifier is one of the algorithm for classification.

REFERENCES

[1] Chen-Huei Chou, Using Tic-Tac-Toe for Learning Data Mining Classifications and Evaluations , *International Journal of Information and Education Technology, Vol. 3, No. 4, August 2013*

[2] Parul Agarwal ,M. Afshar Alam ,Ranjit Biswas , Issues, Challenges and Tools of Clustering Algorithms, *IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 2, May 2011* .

[3] A. Dharmarajan1 and T. Velmurugan2, Lung Cancer Data Analysis by k-means and FarthestFirst Clustering Algorithms, ISSN (Print) : 0974-6846ISSN (Online) : 0974-5645 *Indian Journal of Science and Technology, Vol 8(15),DOI: 10.17485/ijst/2015/v8i15/73329, July 2015*

[4] Ian H.Witten, WEKA—Experiences with a Java Open-Source Project, *Journal of Machine Learning Research 11 (2010) 2533-2541*

[5] Md. Nurul Amin, Md. Ahsan Habib, Comparison of Different Classification Techniques Using WEKA for Hematological Data, *American Journal of Engineering Research (AJER) e-ISSN : 2320-0847 p-ISSN : 2320-0936 Volume-4, Issue-3, pp-55-61 www.ajer.org*

[6] Mark Hall Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer Peter Reutemann, Ian H. Witten, The WEKA Data Mining Software: An Update, *SIGKDD Explorations Volume 11, Issue1*

[7] José del Campo-Ávila, Ricardo Conejo, Francisco Triguero, Rafael Morales-Bueno,Mining Web-based Educational Systems to Predict Student Learning Achievements, *International Journal of Artificial Intelligence and Interactive Multimedia, Vol. 3, Nº 2.*

[8] Rohit Arora, Suman, Comparative Analysis of Classification Algorithms on Different Datasets using WEKA, *International Journal of Computer Applications (0975 – 8887) Volume 54–No.13, September 2012*

[9] Sonam Narwal Mr. Kamaldeep Mintwal, Comparison the Various Clustering and Classification Algorithms of WEKA Tools, *International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 12, December 2013 ISSN: 2277 128X*

[10] E. Bhuvaneswari V. R. Sarma Dhulipala2, "The Study and Analysis of Classification Algorithm for Animal Kingdom Dataset. Information Engineering Volume 2 Issue 1, March 2013

[11] http://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/vote.arff