



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Performance Comparison of Machine Learning Algorithms on Integration of Clustering and Classification Techniques

Samir Kumar Sarangi and Dr. Vivek Jaglan

Department of Computer Science & Engineering,
Amity School of Engineering and Technology,
Amity University Haryana, Gurgaon (Manesar) – 122 413
Haryana, India

Abstract: *The abundance of data in business, research, industry, science and in many fields makes it very difficult to handle them. It is complicated to explore any valuable information, needed to take any important decision, but problem is how to discover this precious information. The effective solution may be data mining, which is a very popular topic at present research. Two main techniques of data mining are clustering and classification, which are basically studied as individual approach till now. In this paper we integrated both (clustering and classification) techniques. After combine application of most frequently used clustering (k-means) algorithm with classification (J48, Multilayer Perceptron, BayesNet, NavieBayes) algorithms, the results were compared and the WEKA data mining tool was used.*

Keywords: *data mining; machine learning; clustering; classification; KMeans; J48; MLP; WEKA*

I. INTRODUCTION

Knowledge discovery in databases (KDD) is known as a synonym of Data mining. The KDD method is applied on large amount of data from stored database/data warehouse/any data repository for extracting patterns, relationships, changes, anomalies and hidden or core information using algorithms and techniques [1]. There are many data mining techniques available like classification, clustering, pattern recognition, and association [2]. The data mining tools gather the data, while the machine learning algorithms are used for taking decisions based on the data collected. The two main techniques of Data mining with wide applicability are Clustering and Classification. In many cases the concept of classification is confused by means of clustering, but there is difference between these two methods. According to the perspective of machine learning, clustering method is unsupervised learning and tries to group sets of objects having relationship between them [3], whereas classification method is supervised and assign objects to sets of predefined classes [4]. Given our goal of clustering a large data set, in this study we used k-means [5] algorithm and for classification, we used J48 [6] [7], Multilayer Perceptron (MLP) [8], BayesNet and NavieBayes classifiers. In our research Weka data mining tool [9] [10] was used for performing classification and clustering techniques. The data set used in this research is of Diabetes Diagnosis, which consists of 8 attributes and 768 instances.

II. PROBLEM STATEMENT

The problem in particular is a comparative study of performance of integrated clustering and classification techniques i.e. simple K-Means clustering algorithm integrated with different classifiers such as J48, MLP, BayesNet and NavieBayes classifier by using various parameters of Diabetes Diagnosis dataset containing 8 attributes, 768 instances and one class attribute.

III. PROPOSED METHOD

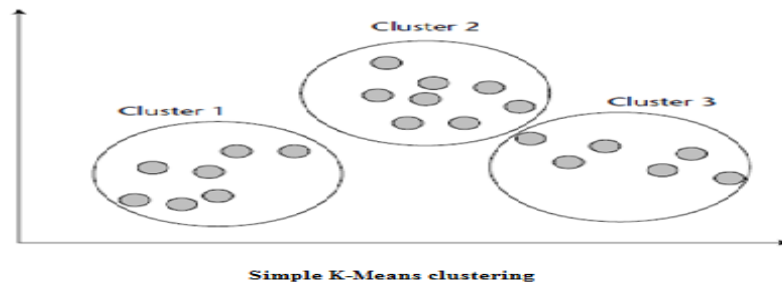
A. CLUSTERING

Clustering is an unsupervised method of data mining. In clustering user needs to define their own classes according to class variables, here no predefined classes are present. In weka number of clustering algorithms are present like cobweb, DBSCAN, FarthestFirst, SimpleKMeans etc. K-Means is the simplest technique and gives more accurate result than others [13].

KMeans algorithm:

- 1) Select number of clusters to be divide
- 2) Select initial value of centroid for each cluster randomly from the set of instances

- 3) Object clustering
 - I. Measure the Euclidean distance (Manhattan distance medianvalue) of each object from the centroid
 - II. place the object in the nearest cluster, whose centroid distance is minimum
- 4) after placing all objects calculate the MEAN value
- 5) If changes found in centroid value and newly calculated mean value
 - I. then make the MEAN value new centroid
 - II. count the repetitions
 - III. go to step 3
- 6) Else stop this process

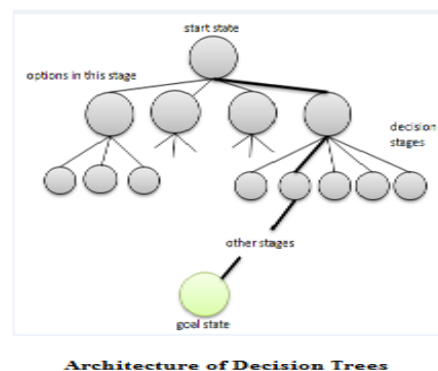
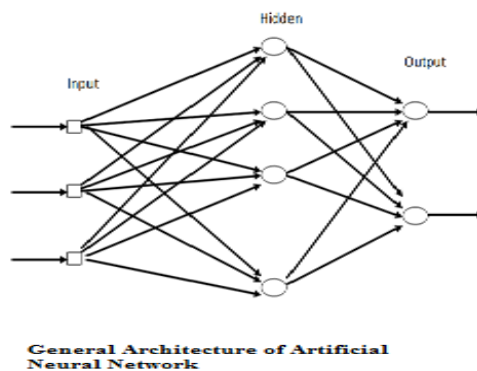


B. CLASSIFICATION

Data classification is a very important task in data mining. For classifying a data mining problem, numbers of classification algorithms are used like Bayes, Functions, misc, Rules, Trees etc. The aim of classification is to calculate the values of each variable and assign those variables to matched predefined classes [7].

I) Multi Layer Perceptron (MLP) algorithm:

MLP uses the concept of neural network. Neural network uses knowledge, perception and control their works to perform in a successive way based on human brain. Brain has the capability of computation, reorganization, processing, controlling actions, taking a correct decision. “Human brain is a collection of more than 10 billion interconnected neurons. Each neuron is a cell that uses biochemical reactions to receive, process, and transmit information”[11]. Neurons are connected to each other, one’s output simulate another neuron connected to that. MLP is a supervised machine learning rule. This MLP is a feed-forward neural net, where a set of input patterns and one or more than one outputs are present. In between the input and output layer there may be number of hidden layers. Here each neuron of every layer is connecting to neurons of its neighbouring layer. The training is provided at input layer and processed by hidden and output layer. In Data Mining the concept of Artificial neural network is very commonly applied in many fields like image processing, biology, controlling system, pathology, pattern recognition, numerical analysis etc [12].



II) J48 algorithm:

J48 are one type of decision tree. It is an optimized version of C4.5 algorithm. When any specific data item is classified, it will be divided in different levels starting from root node to the leaf or terminal node in a hierarchical manner. This process will continue until it reaches over the terminal node which further cannot be

subdivided. This Decision Tree are used in decision analysis, in this tree every non-leaf node represents a test or decision on the data item. Depending upon output at each level it will choose certain branch. For example a question has multiple answers, and each answer can further be divided, it will subdivide up to the last level. Decision Tree is a very powerful technique which is used for real world problems by classified the problem into tree structure and applies the control rules over the internal nodes.

III) *BayesNet*

By using the bayes theorem BayesNet can be developed. To structure a Baysian network first conditional probability of every node must be calculated. Acyclic graphs are used to represent the network. Before building the network, it is assumed that there are no missing values and all attribute values are nominal. Different types of estimators (BayesNetEstimator, BMAEstimator, MultiNominalBMAEstimator, SimpleEstimators) and algorithms (GeneticSearch, Hillclimber, K2, LAGDHillclimber, RepeatedHillclimber, SimulatedAnnealing, TabuSearch, TAN) were used to estimate the probability. The output was visualized by using graph.

IV) *NaiveBayes*

Navie bayes classification algorithm is based on Baysian classification theory. It is mainly used when the dimension of the input is very high. It is simple, robust and easy to perform. For taking a decision, NavieBayes calculate the prior probability (which is from previous experience) of the object and likelihood of the given object. Final classification is done by calculating the posterior probability of the object by multiplying the prior probability and likelihood. Based on the posterior probability, it takes the decision. The performance of NavieBayes depends upon the reality of data set.

C. *TOOL*

Weka is a widely used data mining tool. It analyzes the data sets by using various machine learning algorithms [14]. Weka tool is developed by using java language and having different data mining methods such as classification, clustering, association and visualization. This tool used four applications on dataset.

- Weka application Explorer
- Weka application Experiment
- Weka application Knowledge folw
- Weka application Simple CLI

The weka application explorer having following features:

- Preprocess: In this step the dataset was loaded in the weka tool and data filter and data transformation was applied on data.
- Classification: It was used for data classification. Many classifiers are available.
- Clustering: Clustering tab was used for clustering the dataset. Different clustering algorithms are available.
- Association: Association rules were used for making relationship in data.
- Select attributes: Used for selecting necessary attributes.
- Visualization: It was used for representing the mined knowledge to the user using an interactive 2D format.

D. *Data Set:*

Weka uses data set (Attribute-relationship) file of “.arff” format. This data set consists of attribute names, types, values and the data. In our research we used “Diabetes Diagnosis” data set having eight attributes, one class attribute and 768 instances.

@attribute Pregnancies numeric	@attribute BMI numeric
@attribute 'PG Concentration' numeric	@attribute 'DP Function' numeric
@attribute 'Diastolic BP' numeric	@attribute Age numeric
@attribute 'Tri Fold Thick' numeric	@attribute Diagnosis {Sick, Healthy}
@attribute 'Serum Ins' numeric	@data

IV. *RESULT & DISCUSSION*

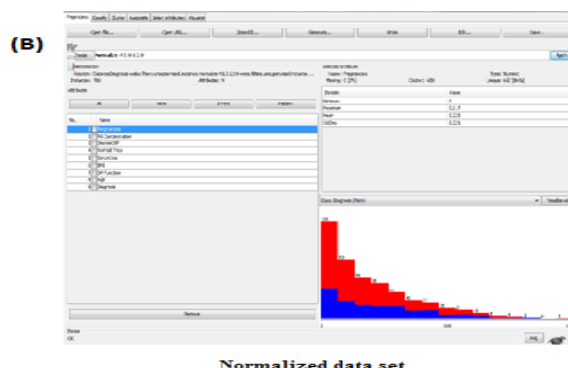
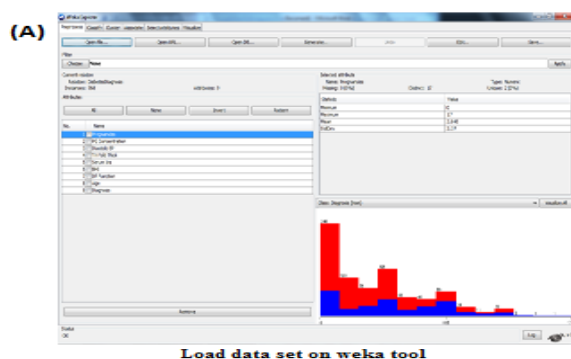
In our research the following parameters were used for evaluating the performance of the above integrated techniques:

Time: It is the total of time needed to construct the model

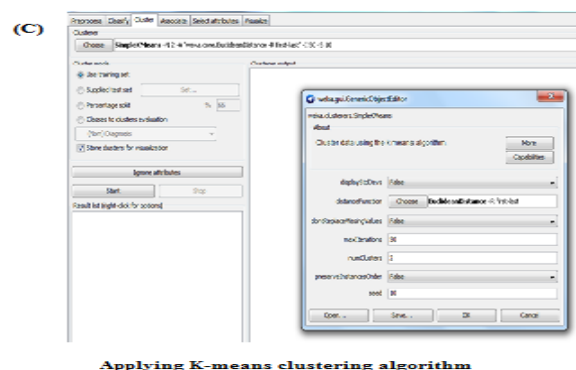
Mean Absolute Error (MAE): It measures of the average of absolute magnitude of individual errors.

Root Mean-Squared Error (RMSE): It is the difference between the predicted values and actual values, which was mainly used for measuring the accuracy.

- A. In the process tab, Open file is to load the data set on to the weka tool
- B. The data set needs to be normalized. Normalization is a process for removing missing values form the data set if any i.e. suppose any null field is present, that will be removed by adding zero instead of null. Normalization process is to transform the data set into a format which is suitable for processing.



- C. After normalization, the integrated technique was applied. First the k-means clustering algorithm was applied, which divided the dataset into number of clusters.



Euclidean distance is a simple distance measure algorithm which calculates the distance of each data values from centroid.

Maximum iteration, it is the value that the maximum number of clustering cycle iterates.

Number of clusters is what user needs to choose for dividing the data set.

No. Of clusters	No. Of iterations	Sum of squared errors	Time(seconds)
2	5	411.60980	0.03
4	12	168.11393	0.09
6	16	150.82860	0.14

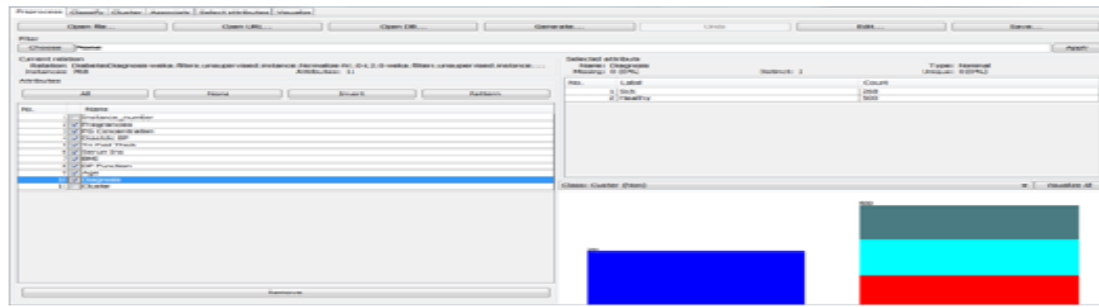
The sum of squared error of cluster 4 is near about cluster 6, but much better than cluster 2. So it is better to choose 4 numbers of clusters in k-means algorithm.

After clustering, the result needs to be saved in .arff format for applying the classification algorithm to make integration.

- D. First open the preprocess tab, load the saved clustering result on weka and apply classification algorithm.

- 1) Loading "clustered.arff" file:

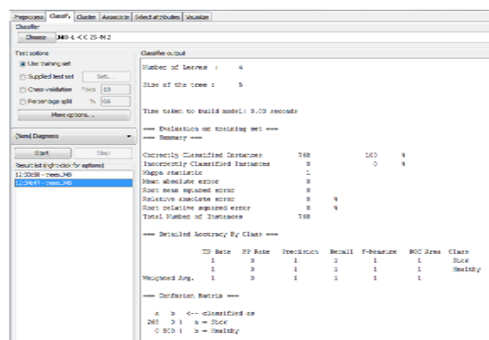
(D.1)



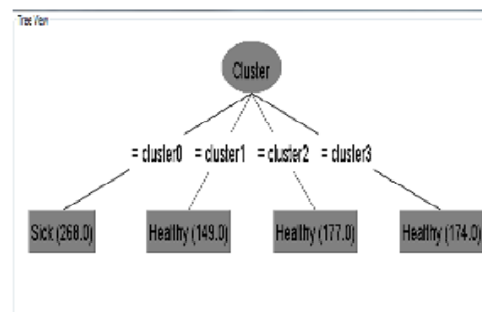
Loading Clustered.arff file

2) Applying J48 classification algorithm:

(D.2)



Output of J48 algorithm



Representation of J48 tree

Result:-

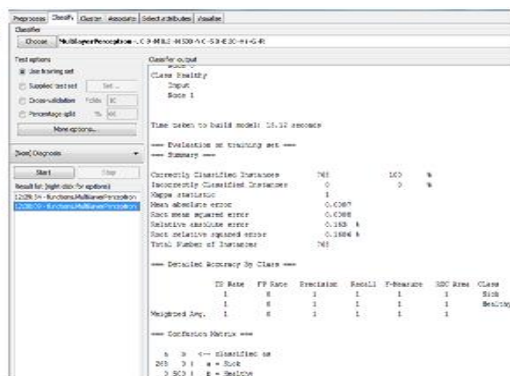
Time taken= 0.03 seconds

MAE=0.0

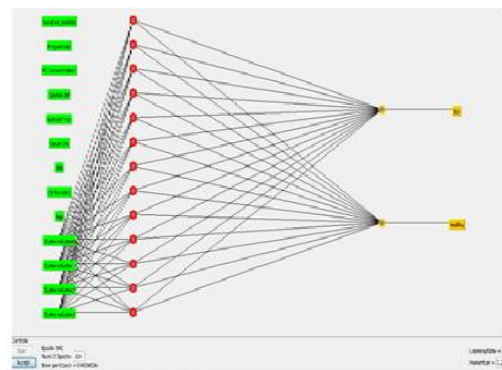
RMSE=0.0

3) Applying MLP classification algorithm:

(D.3)



Output of MLP algorithm



Representation of MLP net

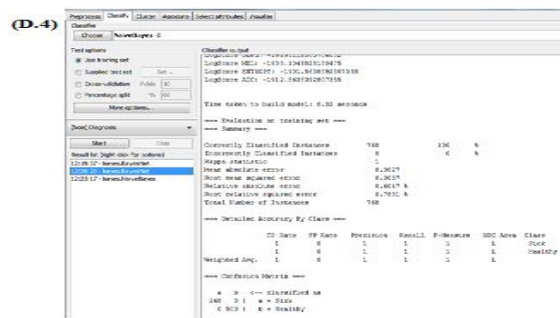
Result:-

Time taken= 15.12 seconds

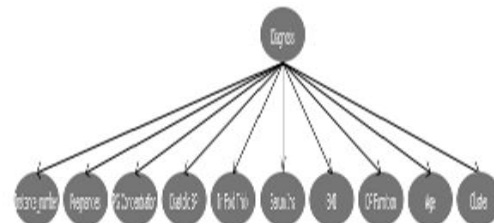
MAE=0.007

RMSE=0.008

4) Applying Bayes Net (BN) classification algorithm:



Output of Bayes Net algorithm



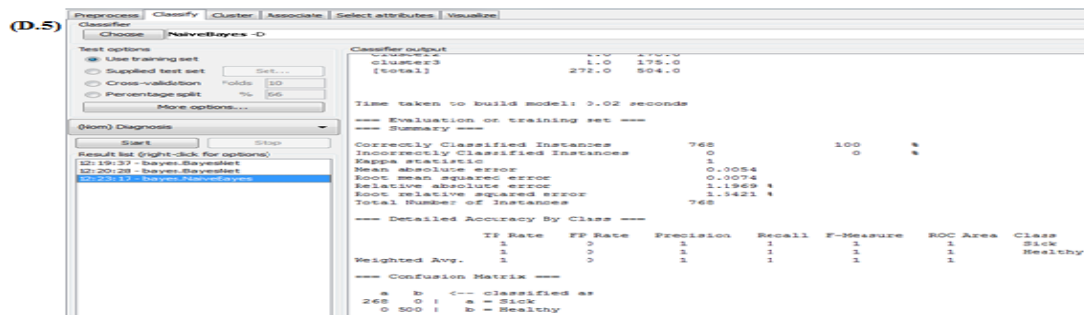
Representation of Bayes net

Result:-

Time taken= 0.02 seconds

MAE=0.0027

RMSE=0.0037

5) Applying Navie Bayes (NB) classification algorithm:

Output of Navie Bayes algorithm

Result:-

Time taken= 0.02 seconds

MAE=0.0054

RMSE=0.0074

V. OBSERVATION AND ANALYSIS:

	K-means+ J48	K-means+ MLP	K-means+ BN	K-means+ NB
Time Taken	0.03	15.12	0.02	0.02
MAE	0.0	0.007	0.0027	0.0054
RMSE	0.0	0.008	0.0037	0.0074

From this above table it may be observed that K-means+BN and K-means+NB both take less time than others, However, K-means+j48 has zero MAE, zero RMSE and time taken by K-means+J48 is slightly greater than K-means+BN and K-means+NB . When K-means+MLP compared with other algorithms, it took more time and showed more errors. According to test and outcome analysis in our research, it was found that in the integrated approach of clustering and classification, the performance of K-means+J48 is better than other algorithms.

VI. CONCLUSION AND FUTURE WORK:

In this paper four different classifiers are integrated with the simple k-means clustering algorithm and this integration technique were applied on “Diabetes Diagnosis” data set. From our observation and analysis it was concluded that the integration of K-means (clustering) + J48 (classification) have zero MAE and RMSE error and it also takes less time to build the model. So the performance of K-means+J48 is better than other algorithms. There are large numbers of classifiers present and many other data mining tools are present. So the future work will be based on other classifiers that can be applied on the data set and also to apply other data mining tools on the data set such that the best techniques can be identified.

REFERENCES

- [1] J. Han and M. Kamber, “Data Mining: Concepts and Techniques,” Morgan Kaufmann, 2000.
- [2] Ritu Chauhan, Harleen Kaur, M.Afshar Alam, “Data Clustering Method for Discovering Clusters in Spatial Cancer Databases”, International Journal of Computer Applications (0975 – 8887) Volume 10– No.6, November 2010.
- [3] Jain, A.K., Murty M.N., and Flynn P.J. (1999):” Data Clustering: A Review”.
- [4] Varun Kumar and Nisha Rathee , ITM University, “Knowledge discovery from database Using an integration of clustering and classification”, International Journal of Advanced Computer Science and Applications, Vol. 2, No.3, March 2011.
- [5] Narendra Sharma, Aman Bajpai, Mr. Ratnesh Litoriya, Department of computer science, Jaypee University of Engg. & Technology, “Comparison the various clustering algorithms of weka tools”, International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, Volume 2, Issue 5, May 2012.
- [6] Yogendra Kumar and Upendra, Samrat Ashok Technical Institute, “An Efficient Intrusion Detection Based on Decision Tree Classifier Using Feature Reduction”, International Journal and Research Publication, Volume 2, January 2012.
- [7] Yugal kumar, Hindu College of Engineering and G. Sahoo, Birla Institute of Technology, “Analysis of Bayes, Neural Network and Tree Classifier of Classification Technique in Data Mining using WEKA”.
- [8] Aaditya Desai and Dr. Sunil Rai, NMIMS University, TCET, Mumbai, “Analysis of Machine Learning Algorithms using WEKA”, International Conference & Workshop on Recent Trends in Technology, (TCET) 2012 Proceedings published in International Journal of Computer Applications® (IJCA).
- [9] Weka 3- Data Mining with open source machine learning software available, <http://www.cs.waikato.ac.nz/ml/weka/>
- [10] Holmes, G.,Donkin,A.,Witten,I.H.: “WEKA a machine learning workbench”. In: Proceeding second Australia and New Zealand Conference on Intelligent Information System, Brisbane, Australia, pp.357-361 (1994).
- [11] Ajith Abraham, “Artificial Neural Networks”, Oklahoma State University, Stillwater, OK, USA.
- [12] Hassoun M.H, (1999), “Fundamentals of Artificial Neural Networks”, Cambridge, MA: MIT press.
- [13] Narendra Sharma, Aman Bajpai, Mr. Ratnesh Litoriya, “Comparison the various clustering algorithms of weka tools”, International Journal of Emerging Technology and Advanced Engineering, (ISSN 2250-2459, Volume 2, Issue 5, May 2012)
- [14] Weka: Data Mining Software in Java <http://www.cs.waikato.ac.nz/ml/weka/>