

A Comparative Study on Classification Algorithm Using Plant Dataset

P.Dheivanai.

M.Phil. Scholar,

Department of Computer Science,

National College, Trichirappalli, Tamilnadu, India.

P.Sundari

Asst. Professor,

Department of Computer Science,

National College, Trichirappalli, Tamilnadu, India

ABSTRACT

Data mining is the process of analyze data from different perspectives and summarizing it into useful information. Clustering is a practical unsupervised data mining task that subdivides an input data set into a desired number of subgroups so that members will have high similarity and the member of different groups have large differences. In this paper it is considered as classification algorithms for examining some mushroom, Irish and Soybean datasets. Some dataset Classified algorithm like Naive Bayes (NB), Support Vector Machine (SVM), Multilayer Perceptron (MLP). This paper it aims to consider the accuracy, sensitivity and specificity percentage to provide a desired result. While considering these algorithms it gives a comprehensive study of three grouping algorithms elements and its limitations.

Keywords: Data mining, Naive Bayes (NB), Support vector Machine (SVM), Multilayer Perceptron (MLP)

I. INTRODUCTION

Data Mining is a knowledge mining process. It is an interdisciplinary subfield of computer science. It is the computational process of discovering patterns in bulky data sets involving method at the intersection of Artificial Intelligence, Machine Learning, Statistics and Database Systems. The enlargement of scientific databases brings a lot of challenges earlier than the research to take out practical information from them using traditional data base techniques. Hence effective mining methods are important to discover the implicit information from huge databases. Cluster analysis is one of the major

data mining algorithms, extensively used for a lot of practical application in various emerging areas like Bioinformatics. Clustering is an unverified method that subdivides an input data set into a desired number of subgroups so that the objects of the same subgroup will be similar or associated to one another and different from or unrelated to the objects in other groups. A high-quality clustering method will produce high quality clusters with high intra-cluster similarity and low inter-cluster similarity. The excellence of a clustering result depends on both the similarity measure used by the method and its execution and also by its ability to discover Some or all of the hidden patterns decision tree is a usually used clustering method that tries to find a user specified number of clusters parents, which are represent by their centroids, by minimizing the square error function. The clustering method aims at optimizing

the cost purpose to minimize the difference of the objects within each cluster, while maximizing the dissimilarity of different clusters.

II LITERATURE SURVEY

Agustin Blas et al.[8] explains the recital of the grouping Genetic algorithm in clustering, happening with proposed encoding, and diverse alteration of crossover and change operation and also initiated the local search contain with the island model for improve the performance of the problem. The real data sets like iris and wine were used and compared the results with the classical approaches such as DBSCAN and K-means, and obtaining the excellent results in projected group is based methodology the upcoming approach such as Genetic algorithm. The presentation of the algorithm was calculated by means of the dissimilar fitness function.

Tzung-Pei-Hong et al.[9] present the act of the Genetic algorithm based characteristic clustering process were enhanced based on the grouping Genetic algorithm. The chromosome depiction, Genetic operations, and fitness function explains in gathering Genetic algorithm for solving the clustering difficulty. The consequence of grouping Genetic algorithm based on clustering algorithm enhanced the convergence speed and fitness value of the clustering problem. In adding the algorithm can also deal with the difficulty of missing values. Another optimization algorithm is used to resolve the difficulty in attribute grouping.

Daniel Gomes Ferrari et al. [10] proposed a new approach to characterize the clustering problems based on the similarity among objects and the method for combine internal indices for ranking algorithms based on the performance of the problem. The experimental results indicated the viability of meta learning systems for an unlabeled approach to the clustering algorithm selection problem. This practice presents the best result from the reserve based set over the attribute based approach.

Kunnuri Lahari et al. [13] enhanced reduce the local minima using evolutionary and population based methods like Genetic algorithm and teaching learning based optimization. The data sets iris and wine are used, and the experimental results are compared with the Genetic algorithm and teaching learning based optimization based clustering with k-means algorithm. The performance of the evolutionary based clustering method compared with some existing clustering method.

Rahila H.Sheikh et al. [14] proclaimed a brief study of Genetic algorithm based clustering. It discussed on proportional examination of K-means and Genetic algorithm based on clustering. Arun Prabha et al.[15] with respect to the idea were improved the cluster quality from K-means clustering using a Genetic algorithm. Huge level clustering problems in data mining also address by this method. The best results are achieved by using this method.

Anusha et al.[16] depicted an enhanced K-means Genetic algorithm for optimal clustering. It overcomes the difficulty of restricted optima with appropriate dataset and also the algorithm fails in experimental time. It is incidental that the method shaped more than the 90% correctness for real life dataset. This also adopted a region information plan for optimizing multi objective troubles. This algorithm used k means Genetic algorithm to identify the compactness of the clusters. It is identified that the algorithm might create least index value for the maximum datasets.

III METHODOLOGY

3.1 OVERVIEW OF DATAMINING

Data Mining is the modernism of concealed information establish in huge quantities of data and can be viewed as a step in the information detection process. It defined as a set of computer-assisted techniques intended to robotically extract huge volumes of integrated data for novel, concealed or unexpected information, or interesting patterns. With little set of data, conventional algebraic study can be efficiently used. The easiest logical step in data mining is to give details of the data summarize its arithmetical attributes such as means and standard deviations, visually evaluation it using chart and graphs, and look for significant links amongst variables such as principles that frequently occur together (Edelstein 1998).

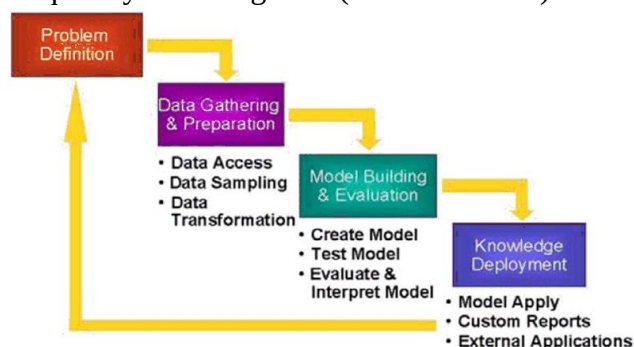


Figure 3.1 An overview of steps that compose KDD process

3.2 PROPOSED ALGORITHMS

3.2.1 SVM

SVM, a decently brand new sort of learning calculation, initially presented. Actually, SVM go for hyper plane incredible isolates the classes of information. SVMs has affirmed the ability not just too precisely isolate elements into right classes, additionally to recognize case whose build up arrangement is not upheld by information. In spite of the fact that SVM are nearly harsh characterize dispersion of preparing cases of every class. It is just reached out numerical figuring's. Two such expansions, the first is to stretch out SVM relapse examination, where the objective to deliver a direct capacity that can genuinely exact that objective capacity. An additional expansion is to figure out how to rank components as opposed to creating a characterization for individual components. Positioning can be decreased to looking at sets of case and delivering a +1 assess if the combine is in the right positioning request not withstanding -1 generally.

Algorithm for SVM:

Step1: Select candidate = {closest pair of opp class}

Step2: while there are violating points do

Step3: Find a violator

Step4: Candidate = Candidate U violator

Step5: if any $z < 0$ due to addition of c to s then

Step6: Candidate = candidate/ p

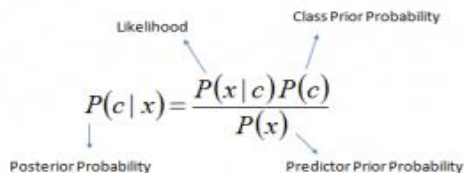
Step7: repeat till all such points are pruned

Step8: End of if

Step9: End of while

3.2.2 NAIVE BAYES

The method that is easy to build the classifier is the Naive Bayes. In this model that will assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There are more classifier algorithm but a family of algorithms based on a general rule: all naive Bayes classifiers believe that the value of a exacting characteristic is independent of the assessment of other feature, given the class variable. It is a categorization method based on Bayes' Theorem with a supposition of self-rule among predictors. Finally, a Naive Bayes classifier assumes that the attendance of a exacting characteristic in a class is unrelated to the occurrence of any other feature. Naive Bayes model brings flexibility to build and mainly useful for huge data sets. Along with ease, Naive Bayes is recognized to break even extremely complicated categorization methods. Bayes theorem provides a way of calculating posterior chance $P(c|x)$ from $P(c)$, $P(x)$ and $P(x|c)$. Look at the equation below:



$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Labels in the diagram: Likelihood points to $P(x|c)$; Class Prior Probability points to $P(c)$; Posterior Probability points to $P(c|x)$; Predictor Prior Probability points to $P(x)$.

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

Above,

$P(c|x)$ 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*.

3.2.3 MULTILAYER PERCEPTRON

A multilayer perceptron (MLP) is a partition of feedforward ANN. A MLP contains three layers of nodes: an input layer, a hidden layer and an output layer. Apart from the input nodes, each node is a neuron that uses a nonlinear launch utility. MLP occupy a supervised learning technique called back propagation for training. Its several layers and non-linear establishment differentiate MLP from a linear perceptron. It can discriminate data that is not linearly divisible.

If a multilayer perceptron has a linear creation occupation in all neurons, that is, a linear function that maps the subjective inputs to the output of each neuron, then linear algebra shows that any number of layers can be reduced to a two-layer input-output model. In MLPs a few neurons use a nonlinear foundation occupation that was developed to model the occurrence of action potentials, or firing, of biological neurons. The two ordinary foundation functions are both sigmoids, and are described by

$$y(v_i) = \tanh(v_i) \text{ and } y(v_i) = (1 + e^{-v_i})^{-1}$$

The first is a hyperbolic tangent that ranges from -1 to 1, while the other is the logistic function, which is similar in shape but ranges from 0 to 1. Here y_i is the output of the i th node (neuron) and v_i is the weighted sum of the input connections. Alternative activation functions have been proposed, including the rectifier and softplus functions. More specialized activation functions include radial basis functions.

3.3 WEKA TOOL

Weka is a familiar sight system in the history of the machine learning research communities and data mining, because it is the only toolkit that has gained such extensive acceptance and survived for an extensive period of time. It provides lots of diverse algorithms for data mining and machine learning. Weka is open source and freely available. It is also platform-independent.

Advantages

- As weka is fully implemented in java programming languages, it is platform independent & portable.
- It is freely available under GNU General Public License.
- Weka s/w contain very graphical user interface, so the system is very easy to access.
- There is very large collection of different data mining algorithms.

Disadvantages

Lack of possibilities to interface with other software

Performance is often sacrificed in favour of portability, design transparency, etc.

Memory limitation, because the data has to be loaded into main memory completely

IV. EXPERIMENT AND RESULTS

The proposed system has been implemented using WEKA tool. The results are obtained as follows after execution. Sample data set is presented. Data Mushroom, Irish and Soybeans dataset has been used with 119 items each for analysis. In this set of association rules are obtained by applying SVM, NB, MLP algorithm. By analyzing the data, and giving different support and confidence values, execution time, can obtain different number of rules. During analysis it found that SVM is much faster for huge number of transactions as compare to NB & MLP.

Table 4.1 showing comparison of various dataset in the proposed algorithm

Dataset	No of Record	Number of Items	Number of Items Per Record
Mushroom	8124	119	30
Soybean	683	36	36
Fishers Iris	150	10	12

4.2 UPLOAD DATASET IN WEKA TOOL

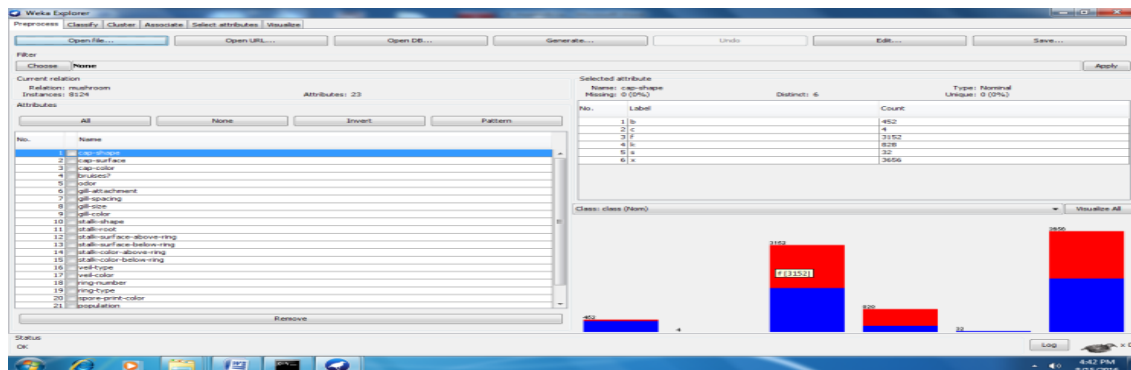


Figure 4.2 (a) .Mushroom Dataset

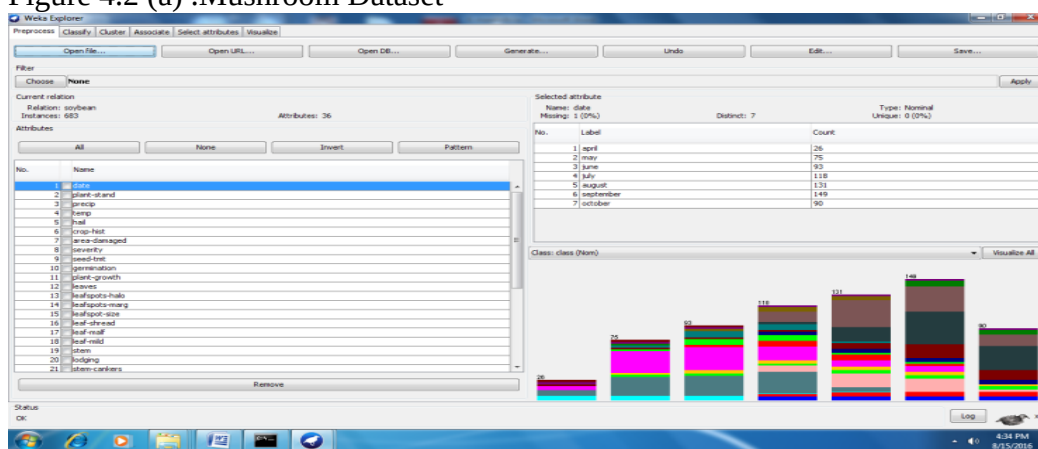


Figure 4.2 (b) .Soya bean Dataset

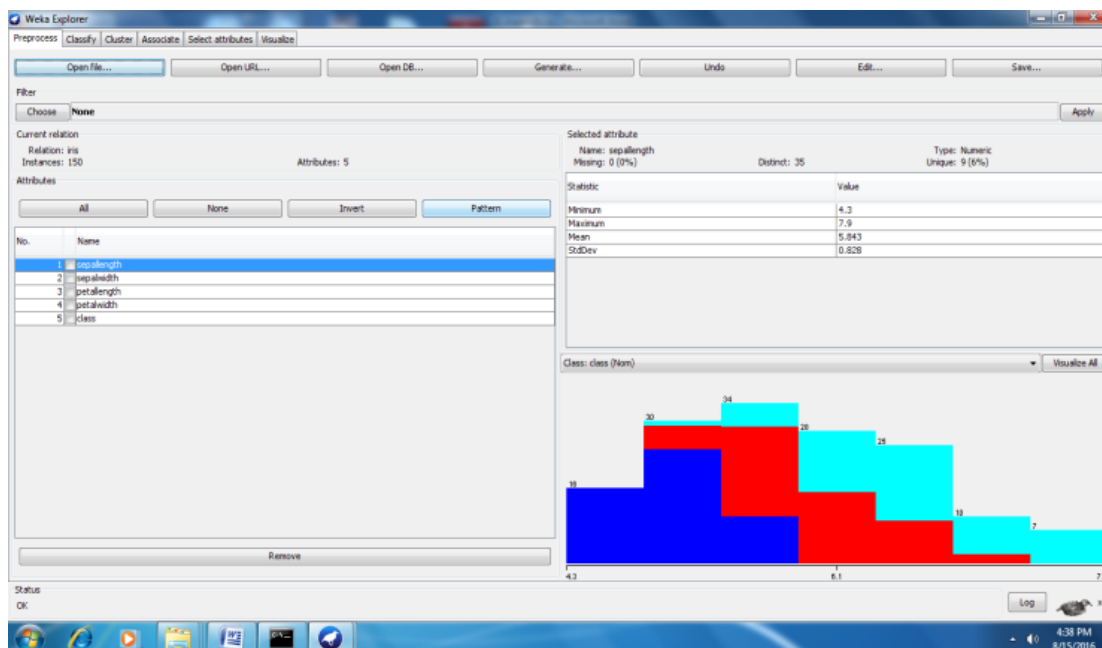


Figure 4.2 (c) Fishers Iris Dataset

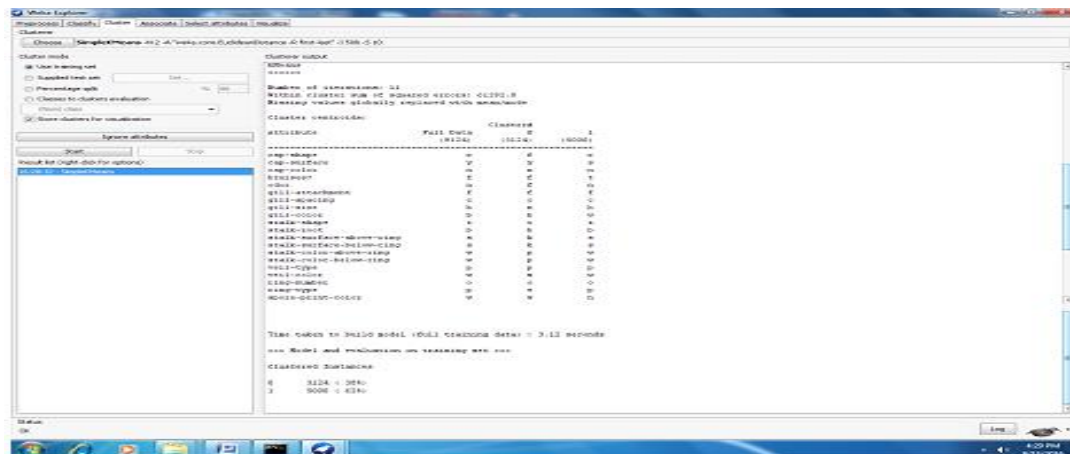


Figure 4.3 (a) Mushroom Dataset SVM



Figure 4.3 (b) Mushroom Dataset NB

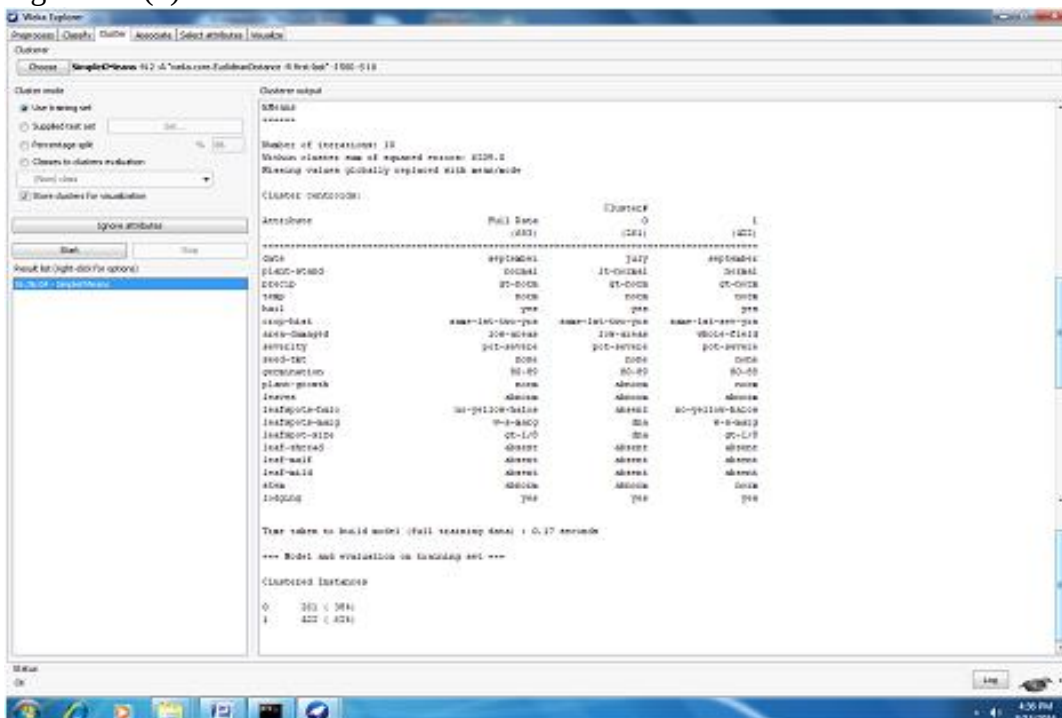
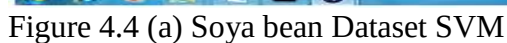
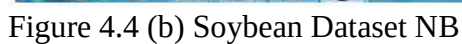
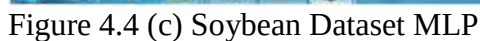


Figure 4.3 (c) Mushroom Dataset MLP



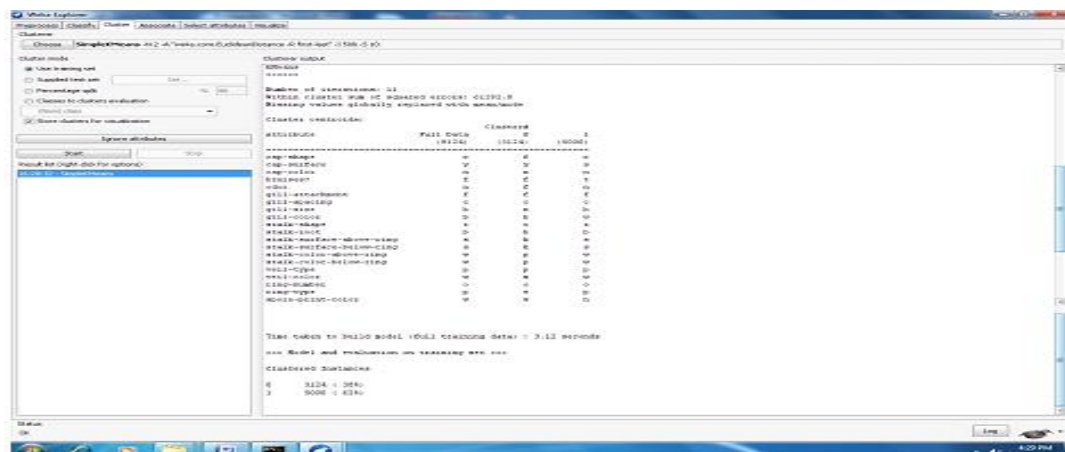


Figure 4.5 (a) Irish Dataset SVM



Figure 4.5 (b) Irish Dataset NB

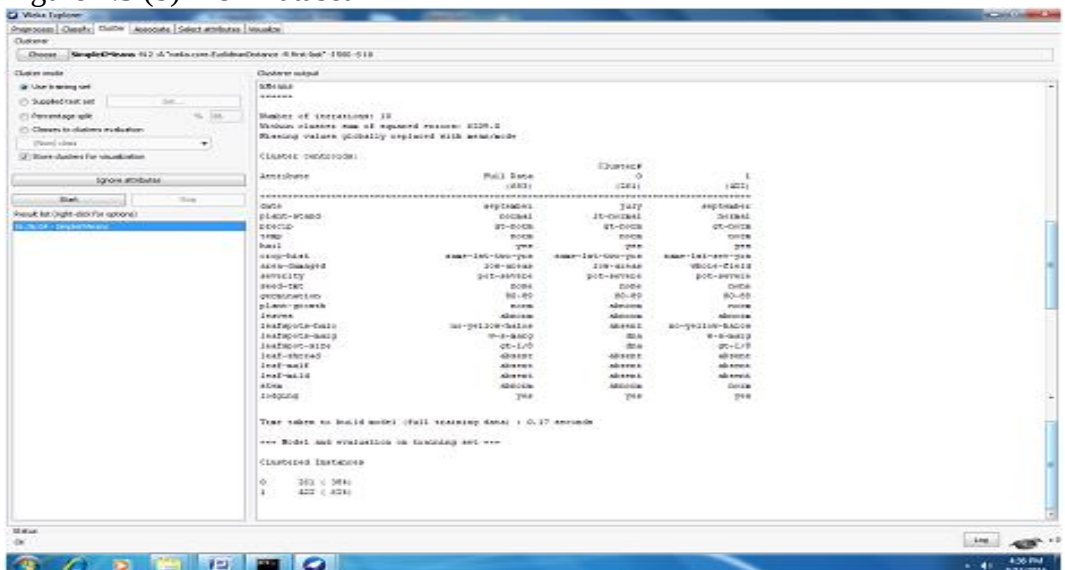


Figure 4.5(c) Irish Dataset MLP

In this study, three types of techniques are used to find out the support, confidence, and execution time accuracy of mushroom, Irish, soybeans data set. High accuracy achieved through SVM technique compared to NB, MLP Techniques.

Comparison of Various Dataset using Algorithms.

Indicating the Minimum Support, confidence, memory space, time taken of the 3 methods used in this study is given below:

Table 4.5(1) showing Minimum Support for all Dataset

Dataset	SVM	NB	MLP
Minimum Support			
Mushroom	0.21	0.45	2.45
Soybean	0.2	0.5	0.17
Iris	0.1	0.9	0.8

Table 4.5(2) showing Confidence for all Dataset

Dataset	SVM	NB	MLP
Confidence			
Mushroom	0.2	0.8	0.5
Soybean	0.5	0.6	0.4
Iris	0.3	0.5	0.8

Table 4.5(3f) showing Execution Time for all Dataset

Dataset	SVM Time Taken (in mil.secs.)	NB Time Taken (in mil.secs.)	MLP Time Taken (in mil.secs.)
Mushroom	0.5	1.45	2.60
Soybean	0.3	0.17	0.25
Fishers Iris	0.1	0.5	0.8

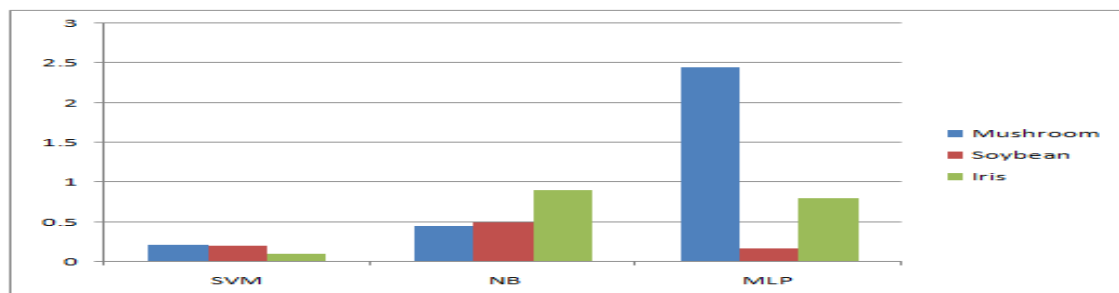


Chart 4.5(a) Graph representing Minimum Support

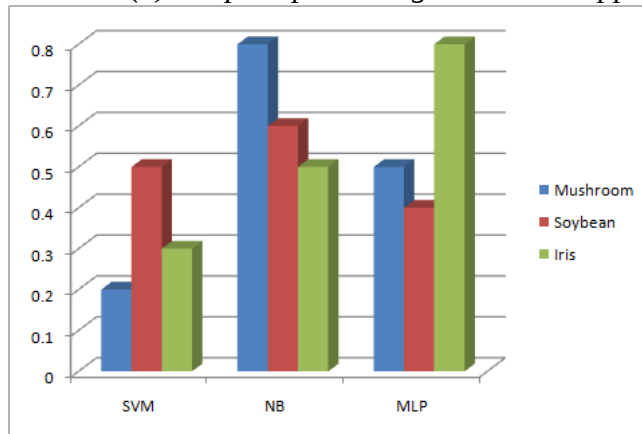


Figure 4.5(b) Graph representing Confidence

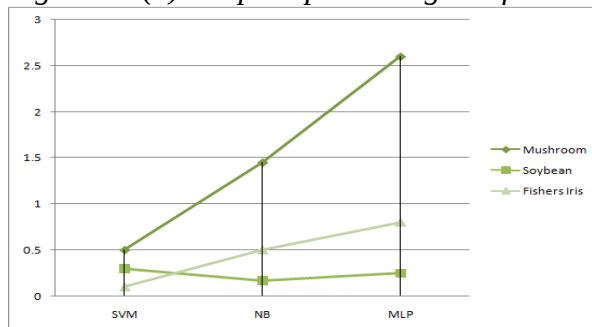


Chart 4.5(c) Graph representing Various Dataset measured in Seconds

V. CONCLUSION

In this comparative study SVM, NB, MLP techniques are used to find out the Minimum Support, Confidence and Execution Time second of Mushroom, Irish and Soybean data. In this paper different algorithm such as Support vector machine (SVM), Naive Bayes, MLP. Are experimentally evaluated using three dataset survey has been carried out. High accuracy achieved through Support Vector Machine technique. In future the Naive Bayes algorithm and MLP can be hybridized to obtain more effective results.

REFERENCES

- [1] Nikita Jain, Vishal Srivastava, "Data Mining techniques : A survey paper" , International Journal of Research in Engineering and Technology, pp. 116-119, 2013.
- [2] M.S.B PhridviRaj, C.V. GuruRao, " Data Mining – Past present and future data streams," Elsevier, pp. 256-264, 2013.
- [3] K.Kameshwaran, K. Malarvizhi, "Survey on Clustering Techniques in Data Mining," International Journal of Computer Science and Information Technologies, pp.2272-2276, 2014.
- [4] Gunjan Verma, Vineeta Verma, "Role and Application of Genetic Algorithm in Data Mining," International Journal of Computer Application, pp. 5-8, 2012.

- [5] Sharaf Ansari, Sailendra Chetlur, Srikanth Prabhu, N. Gopalakrishna Kini, Govardhan Hegde, Yusuf Hyder, "An Overview of Clustering Analysis Techniques used in Data Mining," International Journal of Emerging Technology
- [6] Aastha Joshi, Rajneet Kaur, "A Review: Comparative Study of Various Clustering Techniques in Data Mining," International Journal of Advanced Research in Computer Science and Software Engineering, pp.55-57, 2013.
- [7] Manoj Kumar, Mohammad Husian, Naveen Upreti, Deepti Gupta, Genetic Algorithm "Review and Application," International Journal of Information Technology and Knowledge Management, pp.451-454, 2010.
- [8] L.E. Agustín-Blas, S. Salcedo-Sanz, S. Jimenez-Fernandez, L. Carro-Calvo, J. Del Ser, J.A. Portilla-Figueras K. Elissa, "A new grouping genetic algorithm for clustering problems," Elsevier, pp.9695-9703, 2012.
- [9] Honga Tzung-Pei, Chun-Hao Chenc, Feng-Shih Lin, "Using group genetic algorithm to improve performance of attribute clustering," Elsevier, pp.1-8, 2015.
- [10] Danial Gomes Ferrari, Leandro Numes de Castro, "Clustering algorithm selection by meta-learning systems: A new distance based problems characterization and ranking combination methods," Elsevier, pp.181-194, 2015.
- [11] Rajashree Dash and Rasmita Dash, "Comparative analysis of K-means and Genetic algorithm based data clustering," International Journal of Advanced Computer and Mathematical Sciences, pp.257-265, 2012.
- [12] Edwin Aldana-Bobadilla, Angel Kuri-Morales, "A Clustering based method on the maximum entropy principle," Entropy Article, pp. 151-180, 2015.
- [13] Kannuri Lahari, M. Ramakrishna Murty, and Suresh C. Satapathy, "Prediction based clustering using genetic algorithm and Learning Based Optimization Performance Analysis," Advances in Intelligent Systems and Computing," pp. 338, 2015.
- [14] Rahila H. Sheikh, M. M. Raghuwanshi, Anil N. Jaiswal, "Genetic algorithm based clustering: A Survey," IEEE, pp.314-319, 2008.
- [15] K. Arun Prabha, R. Saranya, "Refinement of K-means clustering using Genetic algorithm," Journal of Computer Application, pp. 256-261, 2011.
- [16] M. Anusha and J. G. R. Sathiaselvan, "An Enhanced K-means Genetic Algorithms for Optimal Clustering," IEEE, pp.580-584, 2014. [17]