

## **An Implementation of Hierarchical Clustering on Indian Liver Patient Dataset**

<sup>1</sup>Prof. M.S. Prasad Babu, <sup>2</sup>K. Swapna, <sup>3</sup>Tilakachuri Balakrishna, <sup>4</sup>Prof. N.B.Venkateswarulu,

<sup>1,2</sup>Dept. of CS & SE, Andhra University, Visakhapatnam, A.P, India.

<sup>3</sup>M.Tech-CST-AIR, Dept. of CS & SE, Andhra University, Visakhapatnam, A.P, India.

<sup>4</sup>Dept of C.S.E, AITAM, Tekkali, A.P, India.

**Abstract:** In modern medical applications data mining techniques are very popular and produce accurate results, diagnosing a liver disease is a complicated process that largely depends on the doctor's knowledge, experience, ability to evaluate the patient's current test results and analyse risk factors that might be causation of illness. Therefore, a need for system to assist physician in making accurate and fast decision has arisen. The main focus of the present paper is to analyse the performance of "Hierarchical clustering algorithm" for ILPD dataset. The results are compared with the normal values given medical books and shown that the hierarchical clustering technique was sufficiently effective to diagnose medical dataset especially, liver diseases and suggested that these results may be used for developing Liver Diagnosis Expert Systems.

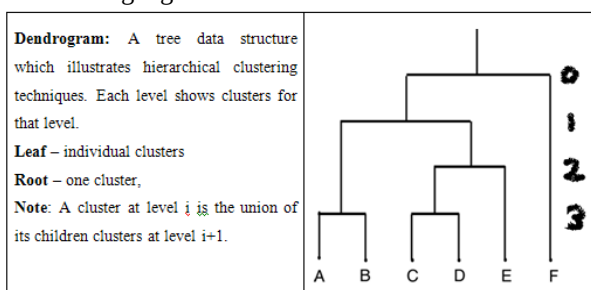
**Keywords:** Data mining, Clustering, Hierarchical Clustering

### **I. INTRODUCTION**

Data mining is an essential process of applying intelligent methods to extract data patterns. Data Mining is often defined as finding hidden information in Knowledgebase. And hence it is called exploratory data analysis, data driven discovery and deductive learning. The major techniques used in data mining are: Classification, Clustering, Association Rules, Regression, Summarization and Sequence Discovery. Clustering is a group of similar set of data objects.

Clustering analysis is a task of identifying characteristics found on the data. For exploratory data mining clustering plays an important role and it is a common technique for statistical data analysis used in many fields, which includes machine learning, Expert system, pattern recognition, image analysis, information retrieval, and bioinformatics. Clustering techniques are very popular in various medical applications for accurate disease diagnosis.

The most popular clustering methods used in data mining are: Hierarchical Clustering, Partitional Clustering, Density Based Clustering, Hierarchical Clustering, Grid based clustering. The Hierarchical method works by grouping data objects (records) into a tree of clusters. It uses distance (similarity) matrix as clustering criteria with a termination condition. There are mainly two approaches used in hierarchical clustering method. They are: Agglomerative Hierarchical Clustering and Divisive Hierarchical Clustering. A tree data structure may be used to illustrate hierarchical clustering algorithm.



*In Hierarchical Clustering – Agglomerative*, Data objects are represented in a bottom-up fashion with data objects are initially in its own cluster and then combines these tiny clusters into larger clusters, until all of the data objects are in a single cluster or until certain termination condition specified by the user is satisfied.

*Where as in Hierarchical Clustering – Divisive* data objects are represented in a top down fashion with all objects are in one cluster initially and then the cluster is subdivided into smaller pieces, until waiting each data object forms a own cluster or certain termination condition specified by the user is satisfied. Here distance between objects in two clusters may be Single link, Average link and complete link based on the distance between clusters is small, average and large respectively. In this paper Hierarchical Clustering is considered because, Tree representation of the cluster is more informative compared to all the remaining clustering algorithms.

## II. PROBLEM STUDY

### A. Existing system

In the existing system, several classification algorithms were applied on ILPD (Indian Liver Patient Dataset), such as Bayesian Classification, Decision Tree Classification, and Classification by Back Propagation, Support Vector Machines (SVM) and Classification based on Association rule mining [1]. The prerequisite attributes required to classify the input data are Age, Gender, TB, DB, TP, ALB, A/G Ratio, SGOT, SGPT and ALP. The system gives the output in the form of a class label. In Existing ILPD the class label '1' represents a patient with liver disease, where as class label '2' represents a patient with no liver disease [2].

### B. Proposed system

In the proposed system hierarchical clustering algorithms is implemented for ILPD set [3] by using WEKA tool for liver diagnosis. The input given to the clustering system is same as the classification system i.e., ILPD features. This algorithm produces desired number of clusters as output. Existing system deviates sometimes from its actual behavior due to the existence of outliers in the training set (ILPD) and predicts a non liver patient as liver patient and vice versa. But proposed algorithm overcome the above disadvantage and produces appropriate results.

## III. METHODOLOGY

### A. Features of Indian Liver Patient Dataset (ILPD)

The ILPD dataset contains 583 liver patient records with 10 attributes that are eight simple blood tests. In this dataset the liver function tests are total bilirubin, direct bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT and Alkphos. This dataset contains 416 liver patients' records and 167 non liver patients' records.

The attributes are simple blood tests used to measure the levels of enzymes, proteins and bilirubin levels in the blood that helps to detect the liver damage. Proteins are large molecules that are needed for the overall health. Enzymes here are protein cells that play important role to help important chemical reactions that occur in the body. Bilirubin helps the body to break down and digest fats.

ALT (SGPT), AST (SGOT), ALP and GGT are the enzymes made by the liver. The ALT, AST, ALP and GGT are the liver enzyme tests that measure the level of ALT, AST, ALP and GGT in the blood respectively. High levels of ALT and AST in the blood can be assign of liver damage. High levels of ALP and GGT can be sign of Liver or bile duct damage.

The description of ILPD Dataset Attributes and Normal values of attributes are represented below

| Attribute | Attribute Information                                              | Liver Function Test | Normal Values  |
|-----------|--------------------------------------------------------------------|---------------------|----------------|
| Gender    | Gender of the patient                                              | TB                  | 0.22-1.0 mg/dl |
| Age       | Age of the patient                                                 | DB                  | 0.0-0.2 mg/dl  |
| TB        | Total Bilirubin                                                    | TP                  | 5.5-8 gm/dl    |
| DB        | Direct Bilirubin                                                   | ALB                 | 3.5-5 gm/dl    |
| TP        | Total Proteins                                                     | A/G                 | >=1            |
| ALB       | Albumin                                                            | SGPT                | 5-45 IU/L      |
| A/G Ratio | Albumin and Globulin Ratio                                         | SGOT                | 5-40 U/L       |
| SGPT(ALT) | Serum Gluamic Pyruvic Transaminase(Alamine Aminotransferase)       | ALP                 | 110-310 U/L    |
| SGOT(AST) | Serum Gluamic Oxaloacetic Transaminase(Aspartate Aminotransferase) |                     |                |
| Alkphos   | Alkaline Phosphate                                                 |                     |                |
| Cluster   | Used to split in to two different clusters                         |                     |                |
| Field     |                                                                    |                     |                |

The sample ILPD dataset in comma separated values format given with attributes Instance number, Age, Gender, TB, DB, ALP, SGPT, SGOT, TP, ALB, A/GRATIO and cluster respectively shown in the following figure 2. The cluster field is used to split data into two clusters. Cluster field 1 means the patient with liver disease, Cluster field 0 means the patient with no liver disease. Each row in this data set belongs to a patient.

```
0, 65, Female, 0.7, 0.1, 187, 16, 18, 6.8, 3.3, 0.9, cluster0
1, 62, Male, 10.9, 5.5, 699, 64, 100, 7.5, 3.2, 0.74, cluster1
2, 62, Male, 7.3, 4.1, 490, 60, 68, 7.3, 3.0, 0.89, cluster1
3, 58, Male, 1.0, 0.4, 182, 14, 20, 6.8, 3.4, 1, cluster1
4, 72, Male, 3.9, 2, 195, 27, 59, 7.3, 2.4, 0.4, cluster1
5, 46, Male, 1.8, 0.7, 208, 19, 14, 7.6, 4.4, 1.3, cluster1
6, 26, Female, 0.9, 0.2, 154, 16, 12, 7.3, 5.1, cluster0
7, 29, Female, 0.9, 0.3, 202, 14, 11, 6.7, 3.6, 1.1, cluster0
8, 17, Male, 0.9, 0.3, 202, 22, 19, 7.4, 4.1, 1.2, cluster1
9, 55, Male, 0.7, 0.2, 290, 53, 58, 6.8, 3.4, 1, cluster1
10, 57, Male, 0.6, 0.1, 210, 51, 59, 5.9, 2.7, 0.8, cluster1
11, 72, Male, 2.7, 1.3, 260, 31, 56, 7.4, 3.0, 0.6, cluster1
12, 64, Male, 0.9, 0.3, 310, 61, 58, 7.3, 4.0, 0.9, cluster1
13, 74, Female, 1.1, 0.4, 214, 22, 30, 8.1, 4.1, 1, cluster0
14, 61, Male, 0.7, 0.2, 145, 53, 41, 5.8, 2.7, 0.87, cluster1
15, 25, Male, 0.6, 0.1, 183, 91, 53, 5.5, 2.3, 0.7, cluster1
16, 38, Male, 1.8, 0.8, 342, 168, 441, 7.6, 4.4, 1.3, cluster1
17, 33, Male, 1.6, 0.5, 165, 15, 23, 7.3, 3.5, 0.92, cluster1
18, 40, Female, 0.9, 0.3, 293, 232, 245, 6.8, 3.1, 0.8, cluster0
19, 40, Female, 0.9, 0.3, 293, 232, 245, 6.8, 3.1, 0.8, cluster0
20, 51, Male, 2.2, 1, 610, 17, 28, 7.3, 2.6, 0.55, cluster1
```

Figure 2 Sample ILPD Data set in arff format

### B. Hierarchical Clustering Algorithm with Example

The pseudo code for Hierarchical Clustering algorithm with mean link is given as below

Input:

1.  $D = \{t_1, t_2, t_3, \dots, t_n\}$  // Set of elements
2. A //Adjacency matrix showing distance between elements.

Output:

3. DE // Dendrogram represented as a set of ordered triples
- Agglomerative Hierarchical clustering algorithm with mean link
4.  $d=0$ ;
  5.  $k=n$ ;
  6.  $K = \{\{t_1\}, \dots, \{t_n\}\}$ ;
  7.  $DE = \{<d, k, K>\}$ ; // initially dendrogram contains each element in its own cluster.
  8.  $M = MST(A)$ ;
  9. Repeat
    - oldk=k;
    - $K_i, K_j$ =two mean clusters closest together in MST;
    - $K = K - \{K_i\} - \{K_j\} \cup \{K_i \cup K_j\}$ ;
    - $K = \text{oldk} - 1$ ;
    - $d = \text{dis}(K_i, K_j)$ ;
    - $DE = DE \cup \{<d, k, K>\}$ ; //New set of clusters added to dendrogram
    - $\text{dis}(K_i, K_j) = \infty$ ;

until  $k=1$ ;

Hierarchical algorithm with example

|   |   |   |   |   |   |  |
|---|---|---|---|---|---|--|
|   | A | B | C | D | E |  |
| A | 0 | 1 | 2 | 2 | 3 |  |
| B | 1 | 0 | 2 | 4 | 3 |  |
| C | 2 | 2 | 0 | 1 | 5 |  |
| D | 2 | 4 | 1 | 0 | 3 |  |
| E | 3 | 3 | 5 | 3 | 0 |  |

$<d, K, K_1> = \{0, 5, \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}\}\}$   
 $<d, k, K_1> = \{1, 3, K\}$ ;  $d=1$ ;  
 $K_1 = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}\} - \{A\} - \{B\} \cup \{A \cup B\}$ ;  
 $K_1 = \{\{C\}, \{D\}, \{E\}\} \cup \{A, B\}$ ;  
 $K_1 = \{\{A, B\}, \{C\}, \{D\}, \{E\}\}$ ;  
 $K_1 = \{\{A, B\}, \{C\}, \{D\}, \{E\}\} - \{C\} - \{D\} \cup \{C \cup D\}$ ;  
 $K_1 = \{\{A, B\}, \{C, D\}, \{E\}\}$

## IV. RESULTS

*Description:* Decision trees are generated using Weka Data mining open source software tool is used. It is used on AMD Processor with 512MB RAM. In this screen shot the user can enter the proposed hierarchical clustering parameters like debug, distance, distance IsBranchLength, linkType, desirednumClusters and print Network in the WEKA explorer shown in figure 3 from the above screen shot, the following results are obtained.

- Cluster tree and
- Cluster Assignments. Shown in fig 4 and fig 5.

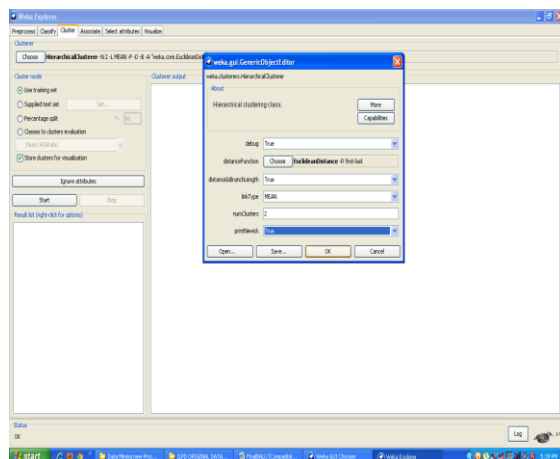


Figure 3 select the parameters for hierarchical clustering.

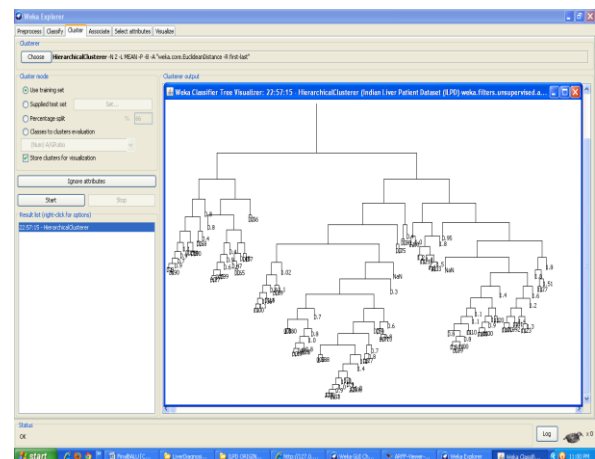


Figure 4 Screen to visualize the clustering tree.

## V. PERFORMANCE EVALUATION

### Comparison of Datasets:

The results generated using existing ILPD classified dataset alone and the results generated using Hierarchical clustering applied on ILPD dataset the fig 6 and 7 respectively. The patient record in the original ILPD dataset “26, Female, 0.9, 0.2, 154, 16, 12, 7, 3.5, 1” was misclassified as a liver patient but the person has no liver disease, which is correctly classified by the proposed hierarchical clustering algorithm. To compare the performance of the proposed hierarchical clustered ILPD dataset with normal values verses and hierarchical clustered ILPD dataset with existing classified ILPD dataset in all iterations are represented in the form of line graph shown in figure 8.

|                                                                 |                                                       |
|-----------------------------------------------------------------|-------------------------------------------------------|
| 0.65, Female, 0.7, 0.1, 187, 16, 18, 6.8, 3.3, 0.9, cluster0    | 65, Female, 0.7, 0.1, 187, 16, 18, 6.8, 3.3, 0.9, 1   |
| 1.62, Male, 10.9, 5.5, 699, 64, 100, 7.5, 3.2, 0.74, cluster1   | 62, Male, 10.9, 5.5, 699, 64, 100, 7.5, 3.2, 0.74, 1  |
| 2.62, Male, 7.3, 4.1, 490, 60, 68, 7.3, 3.3, 0.89, cluster1     | 62, Male, 7.3, 4.1, 490, 60, 68, 7.3, 3.3, 0.89, 1    |
| 3.58, Male, 1.0, 0.4, 182, 14, 20, 6.8, 3.4, 1.1, cluster1      | 58, Male, 1.0, 0.4, 182, 14, 20, 6.8, 3.4, 1.1, 1     |
| 4.72, Male, 3.9, 2, 195, 27, 59, 7.3, 2.4, 0.4, cluster1        | 72, Male, 3.9, 2, 195, 27, 59, 7.3, 2.4, 0.4, 1       |
| 5.46, Male, 1.8, 0.7, 208, 19, 14, 7.6, 4.4, 1.3, cluster1      | 46, Male, 1.8, 0.7, 208, 19, 14, 7.6, 4.4, 1.3, 1     |
| 6.26, Female, 0.9, 0.2, 154, 16, 12, 7.3, 3.5, 1, cluster2      | 26, Female, 0.9, 0.2, 154, 16, 12, 7.3, 3.5, 1, 1     |
| 7.29, Female, 0.9, 0.3, 202, 22, 19, 7.4, 4.1, 1.2, cluster0    | 29, Female, 0.9, 0.3, 202, 22, 19, 7.4, 4.1, 1.2, 1   |
| 8.17, Male, 0.9, 0.3, 202, 22, 19, 7.4, 4.1, 1.2, cluster1      | 17, Male, 0.9, 0.3, 202, 22, 19, 7.4, 4.1, 1.2, 1     |
| 9.55, Male, 0.7, 0.2, 290, 53, 58, 6.8, 3.4, 1.1, cluster1      | 55, Male, 0.7, 0.2, 290, 53, 58, 6.8, 3.4, 1.1, 1     |
| 10.57, Male, 0.6, 0.1, 210, 51, 59, 5.9, 2.7, 0.8, cluster1     | 57, Male, 0.6, 0.1, 210, 51, 59, 5.9, 2.7, 0.8, 1     |
| 11.72, Male, 2.7, 1.3, 260, 31, 56, 7.4, 3.0, 0.6, cluster1     | 72, Male, 2.7, 1.3, 260, 31, 56, 7.4, 3.0, 0.6, 1     |
| 12.64, Male, 0.9, 0.3, 310, 61, 58, 7.3, 4.0, 0.9, cluster1     | 64, Male, 0.9, 0.3, 310, 61, 58, 7.3, 4.0, 0.9, 1     |
| 13.74, Female, 1.1, 0.4, 214, 22, 30, 8.1, 4.1, 1.1, cluster0   | 74, Female, 1.1, 0.4, 214, 22, 30, 8.1, 4.1, 1.1, 1   |
| 14.61, Male, 0.7, 0.2, 145, 53, 41, 5.8, 2.7, 0.87, cluster1    | 61, Male, 0.7, 0.2, 145, 53, 41, 5.8, 2.7, 0.87, 1    |
| 15.25, Male, 0.6, 0.1, 183, 91, 53, 5.5, 2.3, 0.7, cluster1     | 25, Male, 0.6, 0.1, 183, 91, 53, 5.5, 2.3, 0.7, 1     |
| 16.38, Male, 1.8, 0.8, 342, 168, 441, 7.6, 4.1, 1.3, cluster1   | 38, Male, 1.8, 0.8, 342, 168, 441, 7.6, 4.1, 1.3, 1   |
| 17.33, Male, 1.5, 0.5, 165, 15, 23, 7.3, 3.5, 0.92, cluster1    | 33, Male, 1.5, 0.5, 165, 15, 23, 7.3, 3.5, 0.92, 1    |
| 18.40, Female, 0.9, 0.3, 293, 232, 245, 6.8, 3.1, 0.8, cluster0 | 40, Female, 0.9, 0.3, 293, 232, 245, 6.8, 3.1, 0.8, 1 |
| 19.40, Female, 0.9, 0.3, 293, 232, 245, 6.8, 3.1, 0.8, cluster0 | 40, Female, 0.9, 0.3, 293, 232, 245, 6.8, 3.1, 0.8, 1 |
| 20.51, Male, 2.2, 1.6, 10, 17, 28, 7.3, 2.6, 0.55, cluster1     | 51, Male, 2.2, 1.6, 10, 17, 28, 7.3, 2.6, 0.55, 1     |

Fig 6 Output Screen Hierarchical Clustering

Fig 7 Output Screen without Clustering

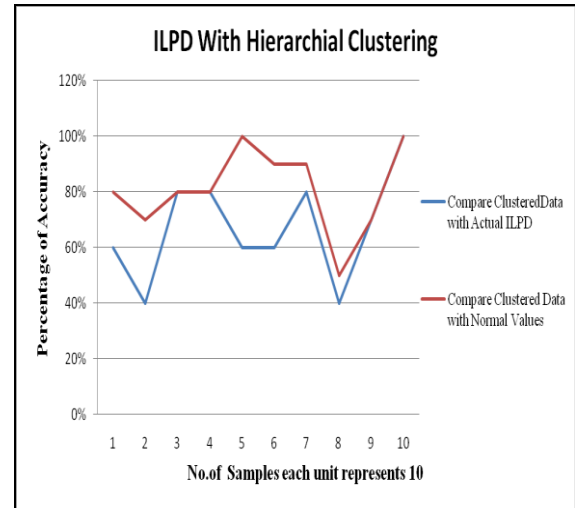


Figure: 8. Graph describing the performance of Clustered data with Normal Values

Based on the above graph, it is concluded that the performance of proposed hierarchical clustered data on ILPD dataset is high compared with existing ILPD data using classification. The obtained knowledge base for proposed Hierarchical clustering is shown in the following table 9.

| S.No | Age | Gender | Tb   | Db  | Alp | Sgpt | Sgot | Tp  | Alb | A/g ratio | Cluster |
|------|-----|--------|------|-----|-----|------|------|-----|-----|-----------|---------|
| 1    | 62  | Male   | 10.9 | 5.5 | 699 | 64   | 100  | 7.5 | 3.2 | 0.74      | C1      |
| 2    | 62  | Male   | 7.3  | 4.1 | 490 | 60   | 68   | 7   | 3.3 | 0.89      | C1      |
| 3    | 58  | Male   | 1    | 0.4 | 182 | 14   | 20   | 6.8 | 3.4 | 1         | C1      |
| 4    | 72  | Male   | 3.9  | 2   | 195 | 27   | 59   | 7.3 | 2.4 | 0.4       | C1      |
| 5    | 46  | Male   | 1.8  | 0.7 | 208 | 19   | 14   | 7.6 | 4.4 | 1.3       | C1      |
| 6    | 17  | Male   | 0.9  | 0.3 | 202 | 22   | 19   | 7.4 | 4.1 | 1.2       | C1      |
| 7    | 55  | Male   | 0.7  | 0.2 | 290 | 53   | 58   | 6.8 | 3.4 | 1         | C1      |
| ...  | ... | ...    | ...  | ... | ... | ...  | ...  | ... | ... | ...       | ...     |
| 443  | 38  | Male   | 1    | 0.3 | 216 | 21   | 24   | 7.3 | 4.4 | 1.5       | C1      |

Table 9 Knowledge Base for Proposed hierarchical Clustering

## VI. CONCLUSION

The performance evaluation is conducted with respect to the performance parameter: Accuracy and found that the Proposed Hierarchical Clustering Algorithm applied on ILPD Data set exhibits more accurate than existing ILPD dataset using classification. These results are used in developing the Liver Diagnosis Expert system for decision making in diagnosing the liver diseases by both patients and doctors. The details of the proposed expert system are included in this paper.

## VII. REFERENCES

- [1] “A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis”, Bendi Venkata Ramana, Prof. M.Surendra Prasad Babu, Prof. N. B. Venkateswarlu, International Journal of Database Management Systems (IJDBMS), Vol.3, No.2, May 2011.
- [2] ILPD Dataset. UCI repository of machine learning databases. Available from [http://archive.ics.uci.edu/ml/datasets/ILPD+\(Indian+Liver+Patient+Dataset\)](http://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)).
- [3] “Survey of Clustering Algorithms” Rui Xu, Student Member, IEEE and Donald Wunsch II, Fellow, IEEE, IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 16, NO. 3, MAY 2005.
- [4] “A Critical Comparative Study of Liver Patients from USA and INDIA: An Exploratory Analysis”, Bendi Venkata Ramana, Prof. M.Surendra Prasad Babu, Prof. N. B. Venkateswarlu, IJCSI International Journal of Computer Science Issues, vol. 9, Issue 3, No 2, May 2012.
- [5] “Development of Maize Expert System Using Ada-Boost Algorithm and Naïve Bayesian Classifier”, M.S.PrasadBabu, VenkateshAchanta, N.V.Ramana Murty, Swapna.K, International Journal of Computer Applications Technology and Research, Volume 1– Issue 3, 89-93, 2012.
- [6] “Survey of Clustering Algorithms” Rui Xu, Student Member, IEEE and Donald Wunsch II, Fellow, IEEE, IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 16, NO. 3, MAY 2005.

- [7] "A Survey of Clustering Techniques", Pradeep Rai, Shubha Singh, International Journal of Computer Applications (0975 – 8887), Volume 7– No.12, October 2010.
- [8] "Research of Knowledge based Expert System used in Maternity diagnosis", Lu.Binjie, In Proceedings of the International Conference on Computer Applications and System Modeling, Pages V-11405-V-1108, 2010.
- [9] "Experiments with a New Boosting Algorithm", Freund, Y. and Schapire, In ICML-96, pp.148-156.
- [10] "A Web Based Sweet Orange Crop Expert System using Rule Based System and Artificial Bee Colony Optimization Algorithm", Prof. M.S. Prasad Babu, Mrs. J. Anitha, K. Hari Krishna, International Journal of Engineering Science and Technology, vol.2(6),2010.
- [11] "An empirical study of the Naive Bayes classifier", Rish Irina, IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence.