Evaluation of Clustering Capability Using Weka Tool

S.Gnanapriya
Department of Information Technology
Easwari Engineering College, Chennai, Tamil Nadu, India

R. Adline Freeda
Department of Information Technology
KCG College of Technology, Chennai, Tamil Nadu, India

M.Sowmiya
Department of Information Technology
Easwari Engineering College, Chennai, Tamil Nadu, India

Abstract- Intelligent soft computing approaches are used to perform gender based clustering of liver patient based on ILPD (Indian Liver Patient Dataset) Data Set with 583 instances consisting of nine attributes. The classification is done using simple K-means, MakeDensityBasedClusterer, hierarchicalclusterer, filteredclusterer, farthestfirst on gender attribute as the classes to cluster evaluation with initial assignment of instances (seed) to cluster as 100. Experimental results show that Hierarchical clusterer is best compare to other algorithms taken for study, neglecting the time factor.

Keywords – simple K-means, Make Density Based Clusterer, hierarchical clusterer, filtered clusterer, farthest first, Weka tool, Mean absolute error.

I. Introduction

With change in traditional food habit and attraction towards international culture the number of liver patient is on increasing side. The liver disease is broadly classified into two types, Nonalcoholic fatty liver disease (NAFLD) and Nonalcoholic steatohepatitis (NASH). NAFLD is isolated fatty liver that accumulates fat in liver. Later stage of NAFLD is NASH causes inflammation and damage to liver cells which leads to liver cirrhosis. The causes for fatty liver are Obesity, diabetes, Manifestation of metabolic syndrome, and symptoms are gastrointestinal bleeding, encephalopathy, fluid accumulation and liver cancer. The fatty liver and NASH can be differentiated using liver biopsy. This disease can be treated by diet, exercise, weight loss and bariatric surgery which reduces food intake and decreases absorption. The Indian Liver Patient dataset is clustered into two, where cluster0 corresponds to Male and cluster1 corresponds to female.

II. Proposed WORK

2.1 Clustering Algorithms–

The clustering efficiency of the Simple K-Means clustering,MakeDensityBasedClusterer,hierarchicalclusterer, filteredclusterer, farthestfirst is evaluated using WEKA tool on ILPD with 583 instances upon the simplest attribute gender. The data set has 441 records corresponding to Male and 142 records corresponding to Female. Result of these clustering algorithms are used to determine the efficiency of the clustering algorithm and calculated mean absolute error is used to determine the clustering accuracy.

Mean absolute error (MAE) is used to measure how close forecasts or predictions are to the eventual outcomes. The mean absolute error is given by

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|e_i\right|, \text{ Where } \left|e_i\right| = \left|f_i - y_i\right|$$

n- number of instances

y_i - True value

f_i -Predicted value

n=583,
Number of instances corresponding to Male = 441, writing in percentage 75.6%
Number of instances corresponding to Female = 142, writing in percentage 24.4%

Class attribute is gender, therefore classes to cluster=2,

2.1.1. K-Means Clustering algorithm –

K-Means clustering intends to partition n objects into k clusters in which each object belongs to the cluster with the nearest mean. This method produces exactly k different clusters of greatest possible distinction. The best number of clusters k leading to the greatest separation (distance) is not known as a priori and must be computed from the data. The objective of K-Means clustering is to minimize total intra-cluster variance, or, the squared error function:

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

Where J-objective function

k- Number of clusters

n-Number of instances

$\left|x_i^j - c_j\right|^2$ - Euclidean distance function.

Algorithm

i. Clusters the data into k groups where k is predefined.
ii. Select k points at random as cluster centers.
iii. Assign objects to their closest cluster center according to the Euclidean distance function.
iv. Calculate the centroid or mean of all objects in each cluster.
v. Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds

2.1.2. Make Density Based Clusterer

According to Density based clustering algorithm points are classified as *core points*, *reachable points* and *outliers*, as follows:
A point *p* is a core point if at least min Pts points are within distance *ε* (*ε* is the maximum radius of the neighborhood from *p*) of it (including *p*). Those points are said to be *directly reachable* from *p*. By definition, no points are *directly reachable* from a non-core point.
A point *q* is reachable from *p* if there is a path
p_1, ..., p_n with $p_1 = p$ and $p_n = q$, where each p_{i+1} is directly reachable from p_i . All points not reachable from any other point are outliers.

Algorithm

1. Let X = {x_1, x_2, x_3, ..., x_n} be the set of data points. DBSCAN requires two parameters: ε (eps) and the minimum number of points required to form a cluster (minPts).
2. Start with an arbitrary starting point that has not been visited.

3. Extract the neighborhood of this point using ε (All points which are within the ε distance are neighborhood).
4. If there are sufficient neighborhood around this point then clustering process starts and point is marked as visited else this point is labeled as noise (Later this point can become the part of the cluster).
5. If a point is found to be a part of the cluster then its ε neighbourhood is also the part of the cluster and the above procedure from step 2 is repeated for all ε neighbourhood points. This is repeated until all points in the cluster are determined.
6. A new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise.
7. This process continues until all points are marked as visited.

2.1.3. Hierarchical clusterer

This approach 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.

In general, the merges and splits are determined in a greedy manner.

Algorithm

Given a set of N items to be clustered,

1. Start by assigning each item to its own cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters equal the distances (similarities) between the items they contain.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster.
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

2.1.4. Filtered Clusterer

Filter is a special subset of a partially ordered set. Let X be a topological space and x a point of X. A filter base B on X is said to cluster at x if and only if each element of B has nonempty intersection with each neighborhood of x. If a filter base B clusters at x and is finer than a filter base C, then C clusters at x too.

Algorithm

i. Every limit of a filter base is also a cluster point of the base.
ii. A filter base B that has x as a cluster point may not converge to x. But there is a finer filter base that does
iii. For a filter base B, the set ∩{cl(B0) : B0∈B} is the set of all cluster points of B
iv. The limit inferior of B is the infimum of the set of all cluster points of B.
v. The limit superior of B is the supremum of the set of all cluster points of B.
vi. B is a convergent filter base if and only if its limit inferior and limit superior agree; in this case, the value on which they agree is the limit of the filter base.

2.1.5. Farthest first Clusterer

This also chooses centroids; assign the objects in cluster but with maximum distance. The initial seeds are value which is at largest distance to the mean of values. Here cluster assignment is different and at initial cluster, get link with high Session Count, like at cluster-0 more than in cluster-1 and so on. Farthestfirst is a variant of k-Means. This places the cluster center at the point further from the points that are farther are clustered together first. This feature of farthest first clustering algorithm speeds up the clustering process in many situations like less reassignment and adjustment is needed.

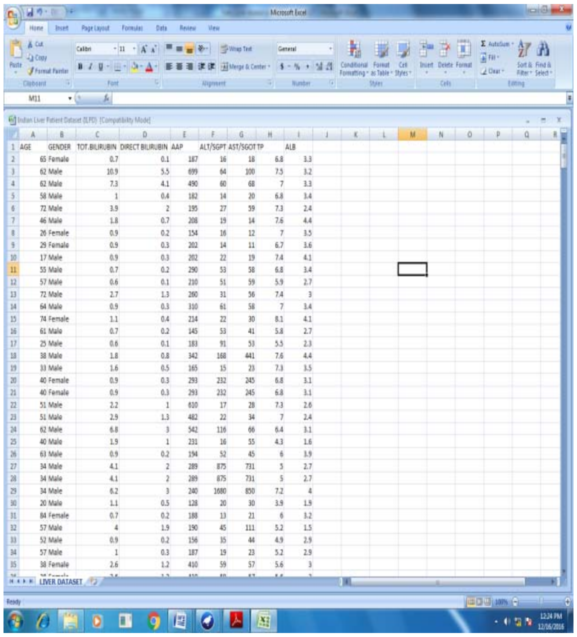
Algorithm

The steps of the algorithm are as follows:

i. Choose a random data as the center point first.
ii. Finding the data that is the farthest point from the first point.
iii. Finding a third point which is the farthest point from two existing points.
iv. Henceforth i=3,4,…,n

Find the data that has not been selected. It is the furthest point from {1,2,…,i-1} and mark it as point i. Use d(x,S) = minyϵS d(x,y) to identify the distance. It has the time complexity O(nk), where n is number of objects in the dataset and k is number of desired clusters

II. EXPERIMENT AND RESULT



AGE	GENDER	TOT.BILIRUBIN	DIRECT BILIRUBIN	AAP	ALT/SGPT	AST/SGOT	TP	ALB
65	Female	0.7	0.1	187	16	18	6.8	3.3
62	Male	10.9	5.5	699	64	100	7.5	3.2
62	Male	7.3	4.1	490	60	68	7	3.3
58	Male	1	0.4	182	14	20	6.8	3.4
72	Male	3.9	2	195	27	59	7.3	2.4
46	Male	1.8	0.7	208	19	14	7.6	4.4
26	Female	0.9	0.2	154	16	12	7	3.5
29	Female	0.9	0.3	202	14	11	6.7	3.6
17	Male	0.9	0.3	202	22	19	7.4	4.1
55	Male	0.7	0.2	290	53	58	6.8	3.4
57	Male	0.6	0.1	210	51	59	5.9	2.7
72	Male	2.7	1.3	260	31	56	7.4	3
64	Male	0.9	0.3	310	61	58	7	3.4
74	Female	1.1	0.4	214	22	30	8.1	4.1
61	Male	0.7	0.2	145	53	41	5.8	2.7
25	Male	0.6	0.1	183	91	53	5.5	2.3
38	Male	1.8	0.8	342	168	441	7.6	4.4
33	Male	1.6	0.5	165	15	23	7.3	3.5
40	Female	0.9	0.3	293	232	245	6.8	3.1
40	Female	0.9	0.3	293	232	245	6.8	3.1
51	Male	2.2	1	610	17	28	7.3	2.6
51	Male	2.9	1.3	482	22	34	7	2.4
62	Male	6.8	3	542	116	66	6.4	3.1
40	Male	1.9	1	231	16	55	4.3	1.6
63	Male	0.9	0.2	194	52	45	6	3.9
34	Male	4.1	2	289	875	731	5	2.7
34	Male	4.1	2	289	875	731	5	2.7
34	Male	6.2	3	240	1680	850	7.2	4
20	Male	1.1	0.5	128	20	30	3.9	1.9
84	Female	0.7	0.2	188	13	21	6	3.2
57	Male	4	1.9	190	45	111	5.2	1.5
52	Male	0.9	0.2	156	35	44	4.9	2.9
57	Male	1	0.3	187	19	23	5.2	2.9
38	Female	2.6	1.2	410	59	57	5.6	3

Fig 1: Input ILPD

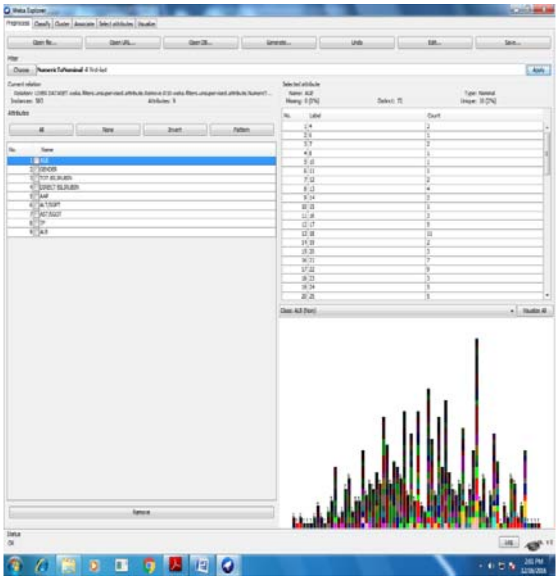


Fig 2: Preprocessed ILPD

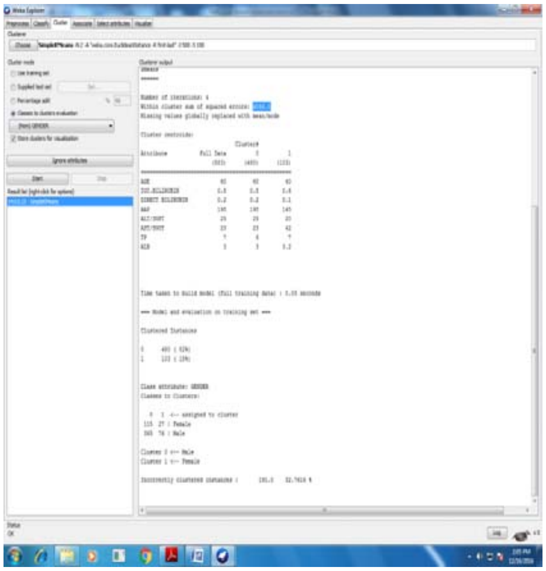


Figure 3: Simple K-Means clustered result

The ILPD contributed by Professor.Bendi Venkata Ramana and weka 3-6-9 was used for the purpose of evaluation. From the simulation of the experiment results, we can conclude that clustering accuracy of Hierarchical clustering is good compared to other clusters.

Table -1 Experiment Result

Clustering Algorithm	Correctly classified instances	Incorrectly classified instances	Mean absolute error	Accuracy	Time taken in secs
Simple	392	191	0.3276	67.23	0.02

K-Means					
Make Density	410	173	0.2967	70.32	0.03
Hierarchical	440	143	0.2452	75.47	1.56
Filtered	412	171	0.2933	70.66	0.02
Farthest first	420	163	0.2797	72.04	0.0

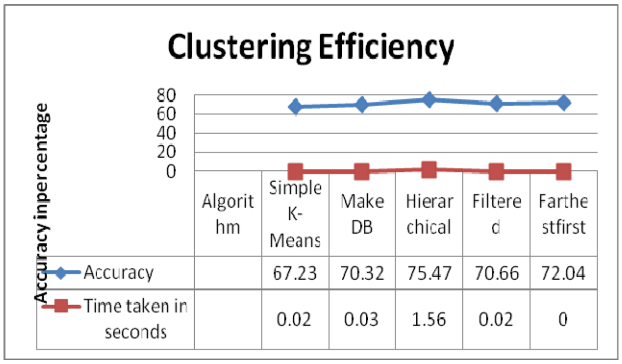


Figure 4: Clustering efficiency of chosen clusterers

IV.CONCLUSION

The Mean absolute error for hierarchical clustering is less compared to other considered algorithms. The percentage of clustering accuracy is highest among the other algorithms. Considering time factor the time taken by hierarchical clustering is longest, this indicates that clustering efficiency is approximately equal to time taken.

V. ACKNOWLEDGMENT

Authors would like to thanks to UCI Machine Learning Repository for providing the ILPD (Indian Liver Patient Dataset) datset.

REFERENCES

[1] Bendi Venkata Ramana, Prof. M. S. Prasad Babu and Prof. N. B. Venkateswarlu, 'A Critical Comparative Study of Liver Patients from USA and INDIA: An Exploratory Analysis' , International Journal of Computer Science Issues, ISSN :1694-0784,May2012.

[2] Anju Gulia , Dr. Rajan Vohra , Praveen Rani , 'Liver Patient Classification Using Intelligent Techniques' International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5110-5115 , ISSN:0975-9646

[3] Harsha Pakhale1,Deepak Kumar Xaxa, 'A Survey on Diagnosis of Liver Disease Classification' International Journal of Engineering and Techniques - Volume 2 Issue 3, May – June 2016, ISSN: 2395-1303

[4] P.Rajeswari, G.Sophia Reena,' Analysis of Liver Disorder Using Data mining Algorithm',Technology Vol. 10 Issue 14 (Ver. 1.0) November 2010

[5] Priyanka Sharmal, Deepika Comparative Analysis of Various Clustering Algorithms Used in WEKA', International Journal of Advance Research in Computer Science and Management Studies ISSN: 2321-7782

[6] A. Dharmarajan, T. Velmurugan,' Lung Cancer Data Analysis by k-means and Farthest First Clustering Algorithms' Indian Journal of Science and Technology, Vol 8(15), DOI: 10.17485/ijst/2015/v8i15/73329, July 2015, ISSN (Print) : 0974-6846 ISSN (Online) : 0974-5645

[7] A.S.Aneeshkumar, C.Jothi Venkateswaran, ‘ Estimating the Surveillance of Liver Disorder using Classification Algorithms’ International Journal of Computer Applications (0975 – 8887) olume 57– No.6, November 2012

[8] D.Sindhuja, R. Jemina Priyadarsini,’A Survey onClassification Techniques in Data Mining for Analyzing Liver Disease Disorder’ International Journal of Computer Science and Mobile Computing, Vol.5 Issue.5, May- 2016, pg. 483-488, International Journal of Computer Science and Mobile Computing, Vol.5 Issue.5, May- 2016, pg. 483-488

[9] S. Karthik, A. Priyadarishini, J. Anuradha and B. K. Tripathy, ‘Classification and Rule Extraction using Rough Set for Diagnosis of Liver Disease and its Types’, Advances in Applied Science Research, 2011, 2 (3): 334-345, ISSN: 0976-8610