COMPARATIVE ANALYSIS OF HEPATITIS DISEASE USING VARIOUS CLUSTERING ALGORITHM

R.UMA.,MCA.,M.Phil.,M.Ed.,
Assistant professor,
Department of computer Application
M.PAVITHRA,Research scholar
Department Of (CS&IT)
Nadar Saraswathi College Of Arts and Science Theni.

ABSTRACT

Data mining is an information discovery database that is the process of discovers an interest and useful pattern and connection in large volumes of records. It is very helpful in pattern alike through which a huge volume of information can be analysed and offer the knowledgeable result for the benefit of the researchers. Health care is one among the data ware house benefit to care provider, patients, health organizations, researchers and insurers in which data analysis is performed to bring out effective treatments and best practices.In earlier years lacks of analysing the hepatitis diseases in many countries, where the millions of people die for hepatitis disease. Hepatitis means harm to the liver with irritation of the liver cells. Many patients died due to lacking quantity of information that may help in effective and efficient decision-making. Largelyof the people are affected by hepatitis to do that dataset for hepatitis disease with key factors is obtained from various medical repositories. In this paper, the hepatitis disease datasets are analysed by using data mining clustering algorithms are Simple K-Means, Hierarchical and Expectation - maximization algorithm are comparing to find the time , accuracy, precision of each.

Keywords: Data mining, Clustering, Randomizing

I.INTRODUCTION:

Data mining is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. It is an interdisciplinary subfield of computer science.[1] The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves data base and data management aspects, data pre-processing, model and inference considerations, post-processing of discovered structures, visualization, and online updating. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD.The notion of a "cluster" cannot be precisely defined, which is one of the reasons why there are so many clustering algorithms. There is a common denominator: a group of data objects. However, different researchers employ different cluster models, and for each of these cluster models again different algorithms can be given. The notion of a cluster, as found by different algorithms, varies significantly in its properties. Understanding these "cluster models" is key to understanding the differences between the various algorithms. [2]

Hepatitis is an inflammation of the liver. The condition can be self-limiting or can progress to fibrosis scarring, cirrhosis or liver cancer. Hepatitis viruses are the most common cause of hepatitis in the world but other infections, toxic substances e.g. alcohol, certain drugs, and autoimmune diseases can also cause hepatitis.There are 5 main hepatitis viruses, referred to as types A, B, C, D and E. These 5 types are of greatest concern because of the burden of illness and death they cause and the potential for outbreaks and epidemic spread. In particular, types B and C lead to chronic disease in hundreds of millions of people and, together, are the most common cause of liver cirrhosis and cancer.[10]

Risk factors

- Travel or work in regions with high rates of hepatitis
- Attend child care or work in a child care center
- Have a clotting-factor disorder, such as hemophilia
- Use injected or non-injected illicit drugs
- Live with another person who has hepatitis
- Have oral-anal contact with someone who has hepatitis
- Hepatitis spreads through contact with blood, semen or other body fluids from an infected person.
- Share needles during intravenous (IV) drug use
- Have a job that exposes you to human blood

II. METHODOLOGY

DATA COLLECTION

Clustering is a process of partitioning a set of data or objects into a set of meaningful sub-classes, called clusters. It helps users to understand the natural grouping or structure in a data set. Clustering is an unsupervised classification and has no predefined classes. In this research the data which I used are being taken from UCI repositories. Mostly standard data set is used to carry out the medical diagnosis, but we can also have better alternative to use analytical dataset instead of standard dataset. The resultant data set thus obtained may be used for the clustering model to calculate the efficiency of the modified system. All these clustering algorithms are implemented with the help of WEKA tool for the diagnosis of heart diseases. Clustering algorithms have been used for analysing the hepatitis disease dataset.[8] The clustering Accuracy should be compared for three algorithms and randomize methods. These repositories and tools are too useful for the researchers for doing the research. They can directly use this information in clustering algorithms of data mining and conclude the result.

Working with Filters

The Filtered clustered meta-clustered offers the user the possibility to apply filters directly before the clustered is learned. This approach eliminates the manual application of a filter in the Pre-process panel, since the data gets processed on the fly. Useful if one needs to try out different filter setups.[5]

weka.filters.unsupervised.instance:

Resample creates a non-stratified subsample of the given dataset, i.e. random sampling without regard to the class information. Otherwise it is equivalent to its supervised variant.

Randomizing data

Since learning algorithms can be prone to the order the data arrives in, randomizing also called “shuffling” the data is a common approach to alleviate thisproblem.Especially repeated randomizations e.g., as during cross-validation, help to generate more realistic statistics.[9] WEKA offers two possibilities for randomizing a dataset: The randomize method of the weka.core.Instances object containing the data itself. This method requires an instance of thejava.util.Random class. How to correctly instantiate such an object isexplained below. Randomize filter (package weka.filters.unsupervised.instance).A very important aspect of Machine Learning experiments is that experimentshave to be repeatable. Subsequent runs of the same experiment setup haveto yield the exact same results. It may seem weird, but randomization is stillpossible in this scenario.

Random number generates never return a completelyrandom sequence of numbers anyway, only a pseudo-random one. In order toachieve repeatable pseudo-random sequences, seeded generators are used. Using the same seed value will always result in the same sequence then. The default constructor of the java.util.Random random number generator class should never be used; as such created objects will generate most likely different sequences. The constructor Random long, using a specified seed value, is the recommended one to use.[6]

In order to get a more dataset-dependent randomization of the data, the Get Random Number Generator (int) method of the weka.core.Instances classcan be used. This method returns a java.util.Random object that was seededwith the sum of the supplied seed and the hash code of the string representation of a randomly chosen weka.core.Instance of the Instances object (using a random number generator seeded with the seed supplied to this method.

Attribute list level in hepatitis disease

1. Class: Die, Live
2. Age: 10, 20, 30, 40, 50, 60, 70, 80
3. Sex: Male, Female
4. Steroid: No, Yes
5. Anti Viral: No, Yes
6. Fatigue: No, Yes
7. Malaise: No, Yes
8. Anorexia: No, Yes
9. Liver Big: No, Yes
10. Liver Firm: No, Yes
11. Spleen Palpable: No, Yes
12. Spiders: No, Yes
13. Ascites: No, Yes
14. Varices: No, Yes
15. Bilirubin: 0.39, 0.80, 1.20, 2.00, 3.00, 4.00
16. Alk Phosphate: 33, 80, 120, 160, 200, 250
17. Sgot: 13, 100, 200, 300, 400, 500,
18. Albumin: 2.1, 3.0, 3.8, 4.5, 5.0, 6.0
19. Protime: 10, 20, 30, 40, 50, 60, 70, 80, 90
20. Histology: No, Yes

Weka

Waikato Environment for Knowledge Analysis. Weka is a collection of machine learning algorithms for data mining tasks. These algorithms can either be applied directly to a data set or can be called from your own Java code. The Weka (pronounced Weh-Kuh) workbench contains a collection of several tools for visualization and algorithms for analytics of data and predictive modeling, together with graphical user interfaces for easy access to this functionality.[7]

ADVANTAGES

• It is also suitable for developing new machine learning schemes.

• Weka loads data file in formats of ARFF, CSV, C4.5, binary. Though it is open source, Free, Extensible, Can be integrated into other java packages.

K-MEANS CLUSTERING:

K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. K-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

Given a set of observations ($\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$), where each observation is a *d*-dimensional real vector, *k*-means clustering aims to partition the *n* observations into k ($\leq n$) sets $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares WCSS sum of distance functions of each point in the cluster to the K center. In other words, its objective is to find:

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \left\| \mathbf{x} - \boldsymbol{\mu}_i \right\|^2$$

K-Means is relatively an efficient method. However, we need to specify the number of clusters, in advance and the final results are sensitive to initialization and often terminates at a local optimum. Unfortunately there is no global theoretical method to find the optimal number of clusters. A practical approach is to compare the outcomes of multiple runs with different *k* and choose the best one based on a predefined criterion. In general, a large *k* probably decreases the error but increases the risk of overfitting.

EM ALGORITHM:

The EM algorithm is used to find (locally) maximum likelihood parameters of a statistical model in cases where the equations cannot be solved directly. Typically these models involve latent variables in addition to unknown parameters and known data observations. That is, either missing values exist among the data, or the model can be formulated more simply by assuming the existence of further unobserved data points. For example, a mixture model can be described more simply by assuming that each observed data point has a corresponding unobserved data point, or latent variable, specifying the mixture component to which each data point belongs.n statistics, an expectation–maximization (EM) algorithm is an iterative method to find maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the *E* step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.

HIERARCHICAL CLUSTERING:

Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom. For example, all files and folders on the hard disk are organized in a hierarchy.

In data mining and statistics, 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.

Parameters:

Time: This is referred to as the time required to complete training or modeling of a dataset. It is represented in seconds.

Precision: Precision and recall is that one is more important than the other in many circumstances. Typical web surfers would like every result on the first page to be relevant high precision but have not the slightest interest in knowing let alone looking at every document that is relevant.

Accuracy: It determines the proportion of the total number of instances clustered to the instances which are correctly clustered. An obvious alternative that may occur to the reader is to judge an information retrieval system by its accuracy, that is, the fraction of its classifications that are correct.

PRECISION:

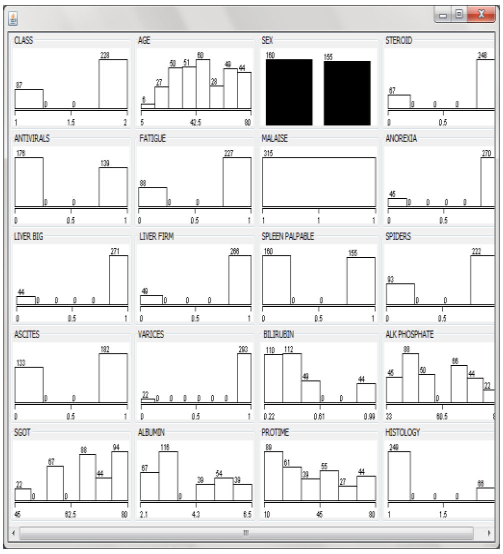
$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

ACCURACY:

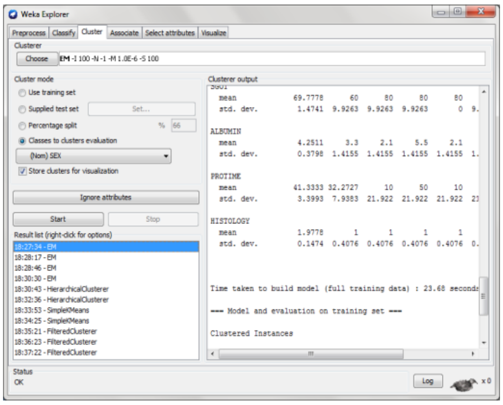
$$\text{accuracy} = (tp + tn)/(tp + fp + fn + tn)$$

III. EXPREMENTAL RESULT

In this paper to comparing the three algorithms and randomizing methods to take the data sets to repository and the applying the tool for weka.[4]The three algorithm to calculated precision and accuracy values for normal clustering outputs and randomizing output then finally to compared to best for ordinary clustering or randomizing method to it.[5]



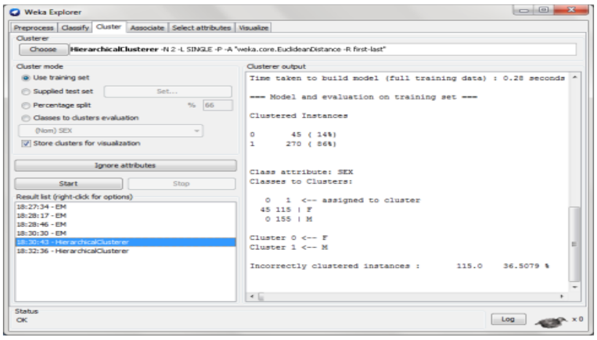
Visualize the data sets



```
Clusterer: EM -I 100 -N -1 -M 1.0E-6 -S 100
Cluster mode: ● Classes to clusters evaluation   (Nom) SEX   ☑ Store clusters for visualization
Result list (right-click for options)
18:27:34 - EM
18:28:17 - EM
18:28:46 - EM
18:30:30 - EM
18:30:43 - HierarchicalClusterer
18:32:36 - HierarchicalClusterer
18:33:53 - SimpleKMeans
18:34:25 - SimpleKMeans
18:35:21 - FilteredClusterer
18:36:23 - FilteredClusterer
18:37:22 - FilteredClusterer
Status: OK
```

```
SGOT
  mean              69.7778       60       80       80       80
  std. dev.          1.4741   9.9263   9.9263   9.9263        0  9.

ALBUMIN
  mean               4.2511      3.3      2.1      5.5      2.1
  std. dev.          0.3798   1.4155   1.4155   1.4155   1.4155  1.

PROTIME
  mean              41.3333  32.2727       10       50       10
  std. dev.          3.3993   7.9383   21.922   21.922   21.922  21

HISTOLOGY
  mean               1.9778        1        1        1        1
  std. dev.          0.1474   0.4076   0.4076   0.4076   0.4076  0.



Time taken to build model (full training data) : 23.68 seconds

=== Model and evaluation on training set ===

Clustered Instances
```

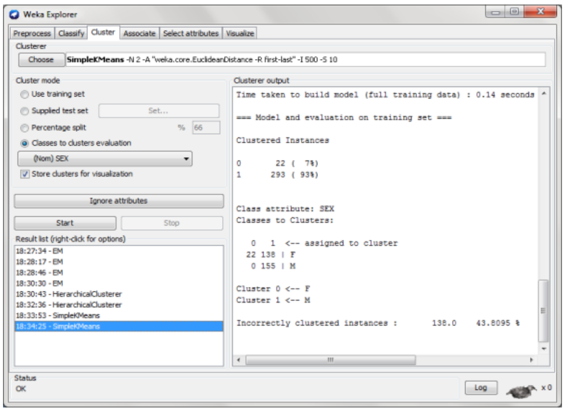
Fig1.1 shows the ordinary EM clustering outputs.



```
Clusterer: HierarchicalClusterer -N 2 -L SINGLE -P -A "weka.core.EuclideanDistance -R first-last"
Cluster mode: ● Use training set   (Nom) SEX   ☑ Store clusters for visualization
Result list (right-click for options)
18:27:34 - EM
18:28:17 - EM
18:28:46 - EM
18:30:30 - EM
18:30:43 - HierarchicalClusterer
18:32:36 - HierarchicalClusterer
Status: OK
```

```
Time taken to build model (full training data) : 0.28 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       45 ( 14%)
1      270 ( 86%)


Class attribute: SEX
Classes to Clusters:

  0   1  <-- assigned to cluster
 45 115 | F
  0 155 | M

Cluster 0 <-- F
Cluster 1 <-- M

Incorrectly clustered instances :       115.0    36.5079 %
```

Fig 1.2 shows the ordinary Hierarchical clustering outputs.

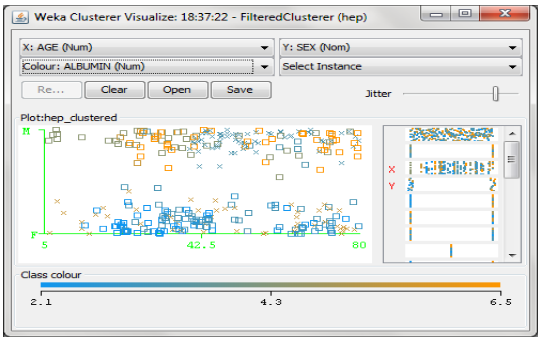


```
Clusterer: SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Cluster mode: ● Classes to clusters evaluation   (Nom) SEX   ☑ Store clusters for visualization
Result list (right-click for options)
18:27:34 - EM
18:28:17 - EM
18:28:46 - EM
18:30:30 - EM
18:30:43 - HierarchicalClusterer
18:32:36 - HierarchicalClusterer
18:33:53 - SimpleKMeans
18:34:25 - SimpleKMeans
Status: OK
```

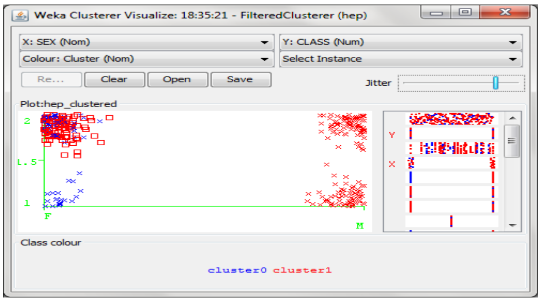
```
Time taken to build model (full training data) : 0.14 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       22 (  7%)
1      293 ( 93%)


Class attribute: SEX
Classes to Clusters:

  0   1  <-- assigned to cluster
 22 138 | F
  0 155 | M

Cluster 0 <-- F
Cluster 1 <-- M

Incorrectly clustered instances :       138.0    43.8095 %
```

Fig 1.3 shows the ordinary k-means clustering output.

Visualize the cluster assignment results 1:

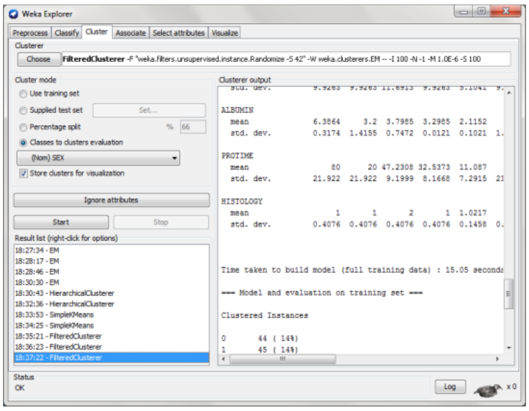


Visualize the cluster assignment results 2:



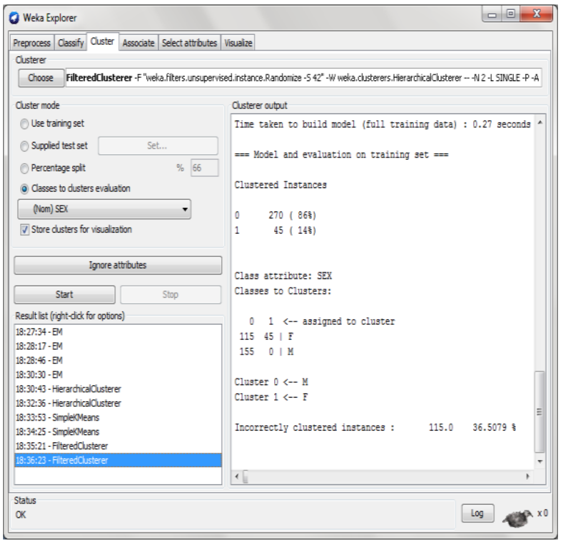
S. No	Clustering Algorithm	No of clustering	Incorrectly cluster instance	Incorrectly cluster instance (%)	Precision	Time	Accuracy
1	EM	2	200	60.49%	0.587	23.60	0.7904
2	Hierarchical	2	115	35.50%	0.2812	0.28	0.6349
3	Simple K-means	2	138	43.80%	0.1375	0.14	0.5619

Randomly shuffles the order of instances passed through it. Comparing the ordinary clustering method to give the clear accuracy level and reduce the timing for randomizing method.



```
Clusterer
FilteredClusterer -F "weka.filters.unsupervised.instance.Randomize -S 42" -W weka.clusterers.EM -- -I 100 -N -1 -M 1.0E-6 -S 100

Cluster mode
○ Use training set
○ Supplied test set
○ Percentage split      % 66
● Classes to clusters evaluation
(Nom) SEX
☑ Store clusters for visualization

Result list (right-click for options)
18:27:34 - EM
18:28:17 - EM
18:28:46 - EM
18:30:30 - EM
18:30:43 - HierarchicalClusterer
18:32:36 - HierarchicalClusterer
18:33:53 - SimpleKMeans
18:34:25 - SimpleKMeans
18:35:21 - FilteredClusterer
18:36:23 - FilteredClusterer
18:37:22 - FilteredClusterer

Clusterer output
ALBUMIN
  mean              6.3864     3.2  3.7985  3.2985  2.1152
  std. dev.         0.3174  1.4155  0.7472  0.0121  0.1021  1.

PROTIME
  mean                  80      20 47.2308 32.5373  11.087
  std. dev.         21.922  21.922  9.1999  8.1668  7.2915  21

HISTOLOGY
  mean                   1       1       2       1  1.0217
  std. dev.         0.4076  0.4076  0.4076  0.4076  0.1458  0.



Time taken to build model (full training data) : 15.05 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       44 ( 14%)
1       45 ( 14%)

Status
OK
```

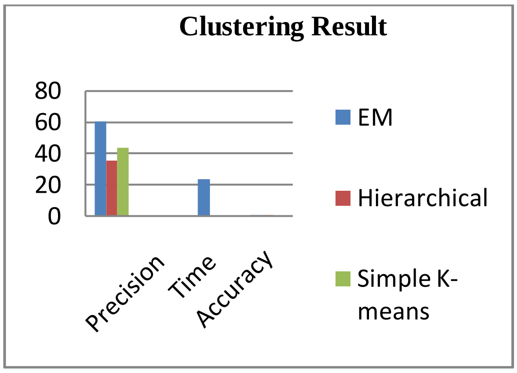
Fig 1.4 shows the randomizing EM clustering output.

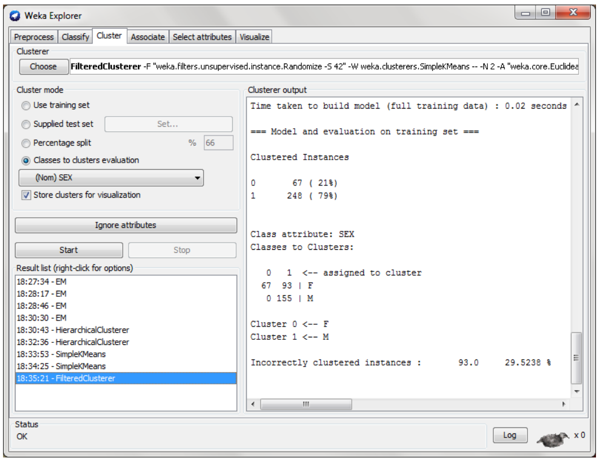


```
Clusterer
FilteredClusterer -F "weka.filters.unsupervised.instance.Randomize -S 42" -W weka.clusterers.HierarchicalClusterer -- -N 2 -L SINGLE -P -A

Cluster mode
○ Use training set
○ Supplied test set
○ Percentage split      % 66
● Classes to clusters evaluation
(Nom) SEX
☑ Store clusters for visualization

Result list (right-click for options)
18:27:34 - EM
18:28:17 - EM
18:28:46 - EM
18:30:30 - EM
18:30:43 - HierarchicalClusterer
18:32:36 - HierarchicalClusterer
18:33:53 - SimpleKMeans
18:34:25 - SimpleKMeans
18:35:21 - FilteredClusterer
18:36:23 - FilteredClusterer

Clusterer output
Time taken to build model (full training data) : 0.27 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      270 ( 86%)
1       45 ( 14%)


Class attribute: SEX
Classes to Clusters:

   0   1  <-- assigned to cluster
 115  45 | F
 155   0 | M

Cluster 0 <-- M
Cluster 1 <-- F

Incorrectly clustered instances :      115.0    36.5079 %

Status
OK
```

Fig 1.5 shows the randomizing Hierarchical output.

Table 1.1 shows the ordinary clustering algorithm values



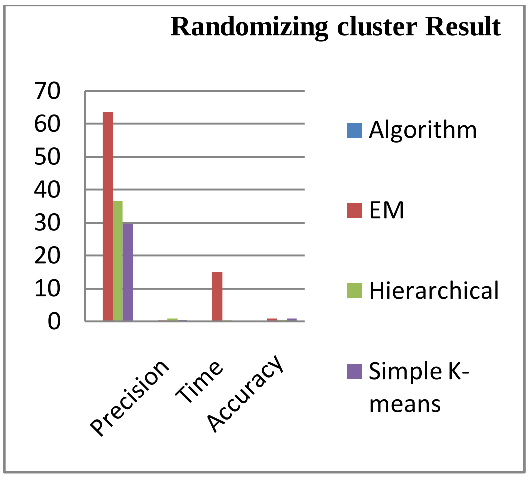


```
Clusterer
FilteredClusterer -F "weka.filters.unsupervised.instance.Randomize -S 42" -W weka.clusterers.SimpleKMeans -- -N 2 -A "weka.core.Euclidea

Cluster mode
○ Use training set
○ Supplied test set
○ Percentage split      % 66
● Classes to clusters evaluation
(Nom) SEX
☑ Store clusters for visualization

Result list (right-click for options)
18:27:34 - EM
18:28:17 - EM
18:28:46 - EM
18:30:30 - EM
18:30:43 - HierarchicalClusterer
18:32:36 - HierarchicalClusterer
18:33:53 - SimpleKMeans
18:34:25 - SimpleKMeans
18:35:21 - FilteredClusterer

Clusterer output
Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       67 ( 21%)
1      248 ( 79%)


Class attribute: SEX
Classes to Clusters:

   0   1  <-- assigned to cluster
  67  93 | F
   0 155 | M

Cluster 0 <-- F
Cluster 1 <-- M

Incorrectly clustered instances :       93.0     29.5238 %

Status
OK
```

Fig 1.6 shows the randomizing Simple k-means output.

S. No	Clustering Algorithm	No of clustering	Incorrectly cluster instance	Incorrectly cluster instance (%)	Precision	Time	Accuracy
1	EM	2	200	63.49%	0.275	15.05	0.8017
2	Hierarchical	2	115	36.50%	0.7187	0.27	0.4181
3	Simple K-means	2	93	29.52%	0.4187	0.02	0.704

Table 1.2 shows the randomizing clustering output values



IV. CONCLUSION

In this research, hepatitis disease prediction data set was developed using three type of clustering algorithms to predict the in correctly cluster instances, time and accuracy of the normalize and randomized methods. Cluster is applying for the three types of clustering algorithms are EM, Hierarchical and simple k-means to normally analysing results are time and accuracy levels are very high values to be performed it. Filtered cluster are used in randomized methods. The randomized methods gives accurate results in EM, Density based Hierarchical and simple k-means algorithm, when compare to normal clustering output to produce the less computation time.

V. FUTURE WORK

The proposed work can be further enhanced and expanded for the automation of Hepatitis disease prediction. Real data from Health care organizations and agencies needs to be collected and all the available techniques will be compared for the accuracy. In this paper the problem of constraining and summarizing different algorithms of data mining used in the field of medical prediction are discussed. The focus is on using different algorithms and combinations of several target attributes for a comparative analysis of variousclustering algorithmsusinghepatitis diseaseusing data mining.In future work, we have planned to propose an effective disease prediction system to predict the hepatitis disease with better accuracy using different data mining techniques and compare the performance of algorithm with other related data mining algorithms.

VI. REFERENCES:

1. https://en.wikipedia.org/wiki/Cluster_analysis
2. www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/.../datamining.html
3. https://en.wikipedia.org/wiki/Data_mining
4. “Knowledge Discovery in the Data Sets of Hepatitis Disease for Diagnosis and Prediction to Support and Serve Community”, Mohammed Abdullah Al-Hagery, Adeebah Saleh Alfaiz, Fatimah Suliman Alorini, Muzun Saleh Althunayan, Mohammed Abdullah Al-Hagery, et al International Journal of Computer and Electronics Research [Volume 4, Issue 6 2015]
5. “Disease Prediction in Data Mining Technique”, S.Vijiyarani, S.Sudha, International Journal of Computer Applications & Information Technology Vol. II, Issue I, January 2013 (ISSN: 2278-7720)
6. “Disease Predicting System Using Data Mining Techniques”, M.A.Nishara Banu , B Gomathy, International Journal of Technical Research and Applications e-ISSN: 2320-8163, www.ijtra.com Volume 1, Issue 5 (Nov-Dec 2013), PP. 41-45.
7. “Data Mining on DNA Sequences of Hepatitis B Virus”, Kwong-Sak Leung, Kin Hong Lee, Jin-Feng Wang, Eddie Y.T. Ng, Henry L.Y. Chan, Stephen K.W. Tsui, Tony S.K. Mok, Pete Chi-Hang Tse, and Joseph Jao-Yiu Sung,Ieee/Acm Transactions On Computational Biology And Bioinformatics, Vol. 8, No. 2, March/April 2011.
8. Comparative Study for Analysis the Prognostic in Hepatitis Data: Data Mining Approach Fadl Mutaher Ba-Alwi, Houzifa M. Hintaya”, Volume 4, Issue 8, August-2013 680 ISSN 2229-5518
9. An Effective Evolutionary Clustering Algorithm: Hepatitis C case study M. H. Marghny ,Rasha M. Abd El-Aziz, Ahmed I. Taloba, International Journal of Computer Applications (0975 – 8887) Volume 34– No.6, November 2011
10. https://en.wikipedia.org/wiki/Hepatitis