Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 6, June -2017

Clustering analysis based learning of Web Mining

[1]Aparna Upadhyay, [2]Mr.Ravindra Gupta , [3]Dr. Varsha Namdev

[1]*Student of Department of Computer Science, S.R.K.U.,BHOPAL*

ABSTRACT:- *The World Wide Web has a giant amount of different forms of data and mining the data leads to knowledge discovery which is used in various fields. These discoveries need a proper way to be analysed for further use such as in machine learning, artificial intelligence etc. Clustering is a conventional method of analysing web data and giving best solutions by different evaluation methods. There are various clustering algorithms present but the accuracy and efficiency is what needed in analysis. In this paper, the comparisons of two of the major clustering algorithms i.e. k-means and Hierarchical algorithm is done and the best algorithm is shown through external evaluation method.*

Keywords – *Web Mining, K- means algorithm, Hierarchical algorithm, Euclidean distance function, Precision and Recall.*

1. Introduction

Web mining - is the application of data mining techniques to discover patterns from World Wide Web. Web mining can be divided into three different types – **Web usagemining**, **Web content mining** and **Web structure mining.** Web mining techniques could be used to solve the information overload problems above directly or indirectly. However, we do not claim that Web mining techniques are the only tools to solve those problems. Other techniques and works from different research areas, such as database (DB), information retrieval (IR), natural language processing (NLP), and the Web document community, could also be used. By the direct approach we mean that the application of the Web mining techniques directly addresses the above problems. For example, a Newsgroup agent that classifies whether the news is relevant to the user. By the indirect approach we mean that the Web mining techniques are used as a part of a bigger application that addresses the above problems. For example, Web mining techniques could be used to create index terms for the Web search services. The Web mining research is a converging research area from several research communities, such as database, IR, and AI research communities especially from machine learning and NLP.

The Web’s size and its unstructured and dynamic content, as well as its multilingual nature, make the extraction of useful knowledge a challenging research problem. Furthermore, the Web generates a large amount of data in other formats that contain valuable information. For example, Web server logs’ information about user access patterns can be used for information personalization or improving Web page design. Machine learning techniques represent one possible approach to addressing the problem.

Artificial intelligence and machine learning techniques have been applied in many important applications in both and data mining research has become a significant subfield in this area. Machine learning techniques also have been used in information retrieval (IR) and text mining applications. The various activities and efforts in this area are referred to as Web mining. The term Web mining was coined by Etzioni (1996) to denote the use of data mining techniques to automatically discover Web documents and services, extract information from Web resources, and uncover general patterns on the Web. Over the years, Web mining research has been extended to cover the use of data mining and similar techniques to discover resources, patterns, and knowledge from the Web and Web-related data (such as Web usage data or Web server logs). The classification is based on two aspects: the purpose and the data sources. Retrieval research focuses on retrieving relevant, existing data or documents from a large database or document repository, while mining research focuses on discovering new information or knowledge in the data. For example, data retrieval techniques are mainly concerned with improving the speed of retrieving data from a database, whereas data mining techniques analyse the data and try to identify interesting patterns. It should be noted, however, that the distinction between information retrieval and text mining is not clear. Many applications, such as text classification and text clustering, are often considered both information retrieval and text mining. In fact, almost all text mining techniques have been investigated by the information retrieval community, notably the Text Retrieval Conference (TREC). Because information retrieval research has the primary goals of indexing and searching, we consider areas such as document clustering to be an instance of text mining techniques that is also part of the retrieval process. Similarly, Web retrieval and Web mining share many similarities. Web document clustering has been studied both in the context of Web retrieval and of Web mining. On the other hand, however, Web mining is not simply the application of information Web Mining: Machine Learning for Web Applications Purpose Finding new patterns or knowledge previously Unknown. A classification of retrieval and mining techniques and applications Data Mining Text Mining Web Mining Data information sources I Any data 1 Textual data 1 Retrieving known data or documents efficiently and Data Retrieval effectively Web Retrieval Information Retrieval and text mining techniques to Web pages; it also involves nontextual data such as Web server logs and other transaction-based data. From this point of view, Web retrieval and Web mining are considered overlapping areas, in which the main criterion for classification is the specific purpose of the application.

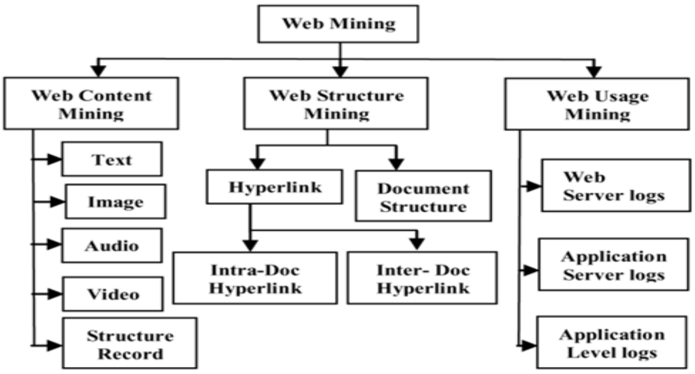


Fig.1.Architecture Diagram of Web mining.

2. Clustering Analysis

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics A. Types of Clustering Cluster: It is said to be "Collection of data objects". Where the two types of similarities of clustering's are: • Intraclass similarity - Objects are similar to objects in same cluster • Interclass dissimilarity - Objects are dissimilar to objects in other clusters B. Methods of Clustering • Partitioning methods • Hierarchical methods •Density-based methods • Grid-based methods • Model-based method

2.1 Hierarchical Methods Connectivity based clustering, also known as hierarchical clustering, is based on the core idea of objects being more related to nearby objects than to objects farther away. As such, these algorithms connect "objects" to form "clusters" based on their distance. A cluster can be described largely by the maximum distance needed to connect parts of the cluster. At different distances, different clusters will form, which can be represented using a dendogram, which explains where the common name "hierarchical clustering" comes from: these algorithms do not provide a single partitioning of the data set, but instead provide an extensive hierarchy of clusters that merge with each other at certain distances. • Hierarchical Agglomerative Methods The hierarchical agglomerative clustering methods are most commonly used. The construction of a hierarchical agglomerative classification can be achieved by the following general algorithm. Find the 2 closest objects and merge them in a cluster.

Find and merge the next two closest points, where a point is either an individual object or a cluster of objects. If more than one cluster remains, return to step 2.

• Agglomerative approach • Divisive approach Individual methods are characterized by the definition used for identification of the closest pair of points, and by the Clustering and its Applications means used to describe the new cluster when two clusters are merged.

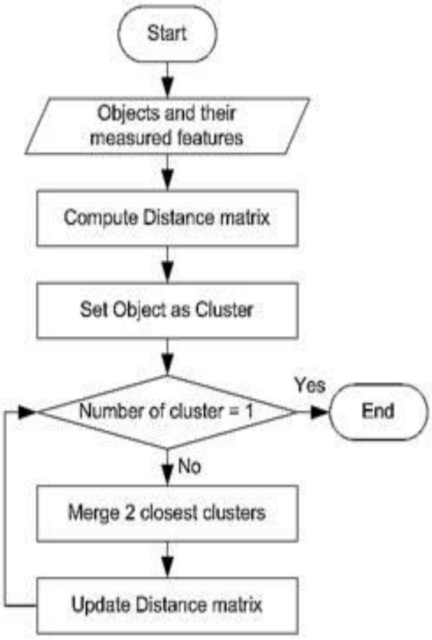


Fig.1.1 Flow diagram of h

2.2K-means Algorithm

K-Means algorithm is a type of partitioning method Group instances based on attributes into k groups High intra-cluster similarity; Low inter-cluster similarity Cluster similarity is measured in regards to the mean value of objects in the cluster. • First, select K random instances from the data – initial cluster centres • Second, each instance is assigned to its closest (most similar) cluster center • Third, each cluster center is updated to the mean of its constituent instances • Repeat steps two and three till there is no further change in assignment of instances to clusters

• K-Means Algorithm Properties

• There are always K clusters.

• There is always at least one item in each cluster. • The clusters are non-hierarchical and they do not overlap.

Every member of a cluster is closer to its cluster than any other cluster because closeness does not always involve the 'center' of clusters.

• The K-Means Algorithm Process

The dataset is partitioned into K clusters and the data points are randomly assigned to the clusters resulting in clusters that have roughly the same number of data points. For each data point:

Calculate the distance from the data point to each cluster.

If the data point is closest to its own cluster, leave it where it is. If the data point is not closest to its own cluster, move it into the closest cluster.

Repeat the above step until a complete pass through all the data points' results in no data point moving from one cluster to another. At this point the clusters are stable and the clustering process ends.

• The choice of initial partition can greatly affect the final clusters that result, in terms of inter-cluster and intracluster distances and cohesion.

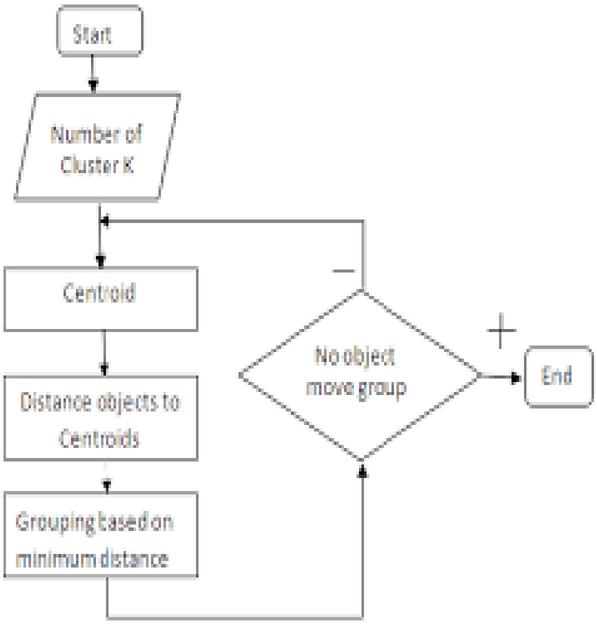

Start
Number of Cluster K
Centroid
Distance objects to Centroids
Grouping based on minimum distance
No object move group
–
+
End

Fig.1.2 Flow Chart of k-means algorithm

3. Proposed Methodology

In the proposed work the K-means clustering is performed on the dataset and the clusters are obtained. Next the hierarchical clustering is performed to obtain the desired number of clusters and in both the algorithms distance functions such as Euclidean and Manhattan are used to calculate the distance of the obtained observations.In cluster analysis, the common approach is to apply the F1-Measure to the precision and recall of *pairs*, often referred to as "pair counting f-measure".

Pair-counting has the nice property that it doesn't directly compare clusters, so the result is well defined when one result has m cluster, the other has n clusters. However, *pair counting needs strict partitions*. When elements are not clustered or assigned to more than one cluster, the pair-counting measures can easily go out of the range 0-1.

The experiment is carried by evaluation of the famous Fisher's Iris dataset by using two clustering algorithms- k-means and hierarchical clustering. In Data clustering (a sub field of Data Mining), k-means and hierarchical based clustering algorithms are popular due to its excellent performance in clustering of large data sets. This paper presents two different comparative studies which includes algorithms like k-means and hierarchical clustering and analysing the best one.

The foremost objective of this paper is to divide the data objects into k number of different homogeneity and each cluster should be heterogeneous to each other. However both the algorithms are not free with errors. The first half experiment is about the clustering analysis with both k-means and Hierarchical algorithm and the second half is about the comparisons between them through external evaluation methods.

3.1 Implementation of Algorithms

Kmeans algorithm

Kmeans is probably the most widely used clustering technique. It belongs to the class of iterative centroid based divisive clustering algorithm. It is different from hierarchical clustering in that it requires the number of clusters, k, to be determined in advance.

Algorithm Description

Kmeans is an algorithm for partition (or cluster) N data points into K disjoint subsets containing data points so as to minimize the sum-of-squares criterion:

Where J =

$$\sum_{j=1}^{n} \sum_{n \epsilon s_j} |x_n - \mu_j|^2$$

Where J is a vector representing the nth data point and is the geometric centroid of the data points in.

- Arbitrarily make any partition and clustering the data points into k clusters.
- Compute the centroid of each cluster based on all the data points within that cluster.
- If a data point is not in the cluster with the closest centroid, switch that data point to that cluster.
- Repeat steps 2 and 3 until convergence is achieved. By then each cluster is stable and no switch of data point arises.

Distance Functions in kmeans
Euclidean distance function
ALGORITHM
In mathematics the Euclidean distance or Euclidean metric is the "general" distance between two points that one would measure with a dimension, and is given by the Pythagoras formula. By using this formula as distance, Euclidean space becomes a metric space.
Euclidean distance is
D (i, j) =

$$\sqrt{(a_{i1} + a_{j1})^2 + (a_{i2} + a_{j2})^2 + \cdots \ldots + (a_{in} + a_{jn})^2}$$

Where i = $(a_{i1}, a_{i2}, \ldots \ldots \ldots a_{in})$ and j = $(a_{j1}, a_{j2}, \ldots \ldots . . a_{jn})$ two n dimension object.

Manhattan Distance Formula
The function of the Manhattan distance enumerate the distance that can be travelled to get from one data point to another if a network path as follows. The Manhattan distance between two elements is the sum of the differences of the corresponding components.
D(i,j) = $|a_{i1} + a_{j1}| + |a_{i2} + a_{j2}| \ldots . . |a_{in} + a_{jn}|.$
Where i = $(a_{i1}, a_{i2}, \ldots \ldots \ldots a_{in})$ and j = $(a_{j1}, a_{j2}, \ldots \ldots . . a_{jn})$ two n dimension object.
Hierarchical clustering
These methods build clusters by recursively portioning the case, either in a top-down or bottom-up. These methods can be divided as follows:

- Agglomerative hierarchical clustering – each object initially represents a cluster of its own. Then cluster are successively merged until the desired cluster structure is obtained.
- Divisive hierarchical clustering – All objects initially belong to one cluster then the group is divided into subgroups, divided successively into their own subgroups. This process continues until the desired cluster structure is obtained.

Manhattan distance function:
Compute the distance that probable travelled to get from one data point to the further if a grid-like path is followed. The Manhattan distance between two elements is the sum of the differences of the corresponding components.

This distance between a point X= $(X_1, X_2, \text{etc.})$ and a point Y= $(Y_1, Y_2\ \text{etc.})$ is: d = $\sum_{i=1}^{n} |X_i + y_i|$

Where n is the number of variables, and Xi and Yi are the values of the variable, a points X and Y respectively.

F-measure function
The F-measure can be used to balance the contribution of false negatives by weighting recall through a parameter {\displaystyle \beta \geq 0} $\beta \geq 0$. Let **precision** and **recall** (both external evaluation measures in themselves) be defined as follows:

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

$$R = \frac{TP}{TP + FN}$$

P is precision rate and {\displaystyle R} R is recall rate. We can calculate the F-measure by using the following formula:

$$F_\beta = \frac{(\beta^2 + 1) \cdot P \cdot R}{\beta^2 \cdot P + R}$$

The k-means and hierarchical algorithms are written in R on RStudio version 1.0.143. All the experiments were run on 2.19 GHz Intel Core™ i5 with 4 GB RAM and running Windows 8.1.

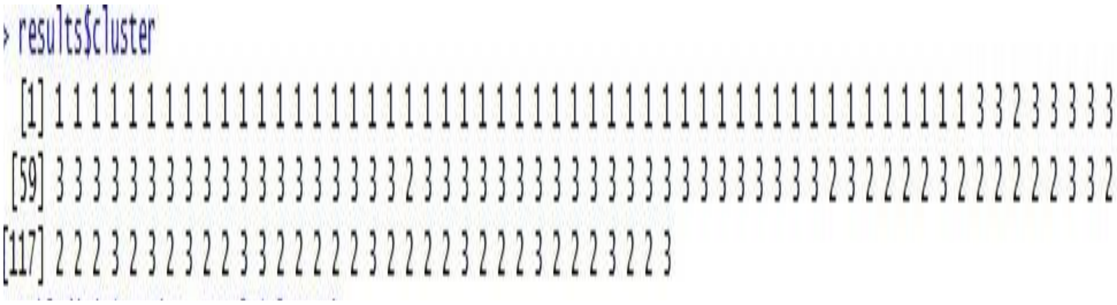
4. DataSet Used

The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper the use of multiple measurements in taxonomic problems as an example of linear discriminant analysis. It is sometimes called Anderson's Iris data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species. Two of the three species were collected in the Gaspe Peninsula "all from the same pasture, and picked on the same day and measured at the same time by the same person with the same apparatus".
The data set consists of 50 samples from each of three species of Iris (Iris setose, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimetres. Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other.

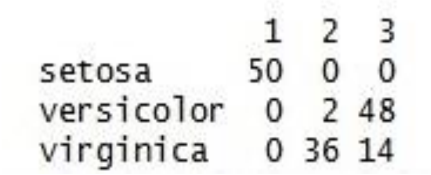
Based on Fisher's linear discriminant model, this data set became a typical test case for many statistical classification techniques in machine learning such as support vector machines.
The use of this data set in cluster analysis however is not common, since the data set only contains two clusters with rather obvious separation. One of the clusters contains *Iris setosa*, while the other cluster contains both *Iris virginica* and *Iris versicolor* and is not separable without the species information Fisher used. This makes the data set a good example to explain the difference between supervised and unsupervised techniques in data mining: Fisher's linear discriminant model can only be obtained when the object species are known: class labels and clusters are not necessarily the same.

4.1 Experimental Results

Kmeans clustering analysis results



```
> results$cluster
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 2 3 3 3 3 3
 [59] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 3 2 2 2 2 3 2 2 2 2 2 2 3 3 2
[117] 2 2 2 3 2 3 2 3 2 2 3 3 2 2 2 2 2 3 2 2 2 2 3 2 2 2 3 2 2 2 3 2 2 3
```



```
            1  2  3
setosa     50  0  0
versicolor  0  2 48
virginica   0 36 14
```

The above table shows that Iris setosa’s 50 species resides in cluster 1.

48 of versicolor species resides in cluster 3.

36 of the virginica species reside in cluster 2.

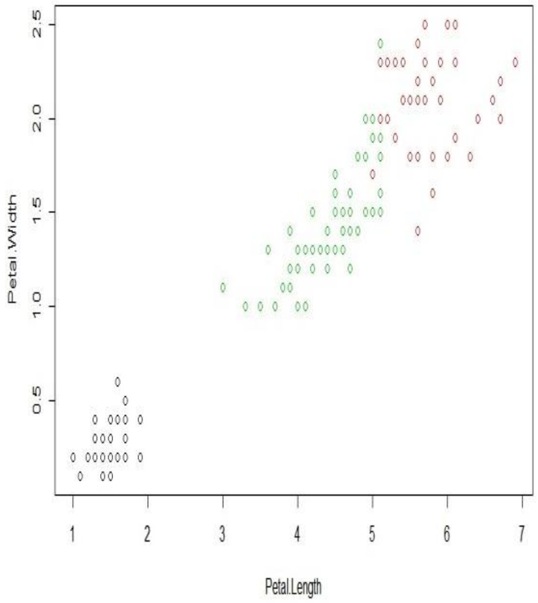


Fig.5 Scatter plot between Petal.Width and Petal.Length

The scatter plot suggests that cluster 1 intercluster distance from cluster 2 and 3 is greater and 2 and 3 have lower intercluster. This means that cluster 1 is well separated.

Hierarchical clustering results

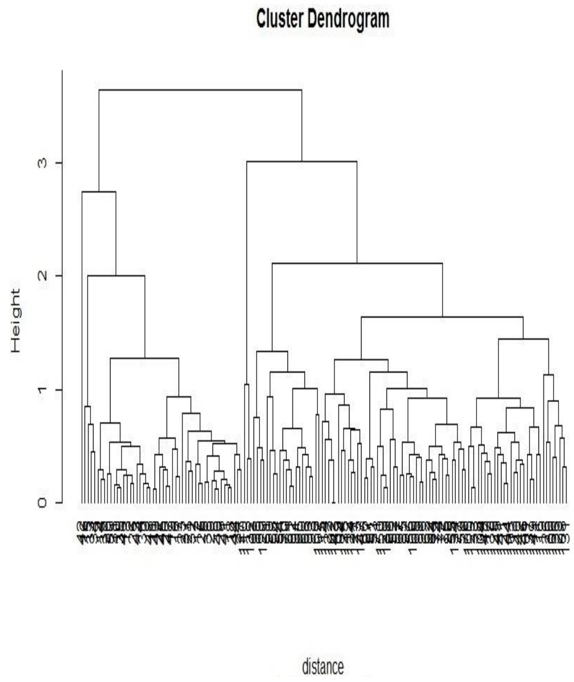
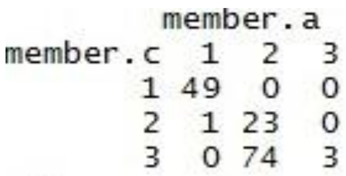


Fig.5.1 The cluster Dendogram



```
            member.a
member.c  1   2   3
       1 49   0   0
       2  1  23   0
       3  0  74   3
```

The above table shows the cluster membership of member.a and member.c

Cluster aggregate

Group	V1	V2	V3	V4
1	*- 0.9987207*	*0.9032290*	*- 1.29875725*	*- 1.25214931*
2	*- 0.3995253*	*- 1.3551557*	*0.06155712*	*- 0.03738991*
3	*0.7600769*	*- 0.1523959*	*0.80729525*	*0.80847629*

Table.1 Cluster aggregate

The above table shows the cluster aggregate results.

Silhouette Plot

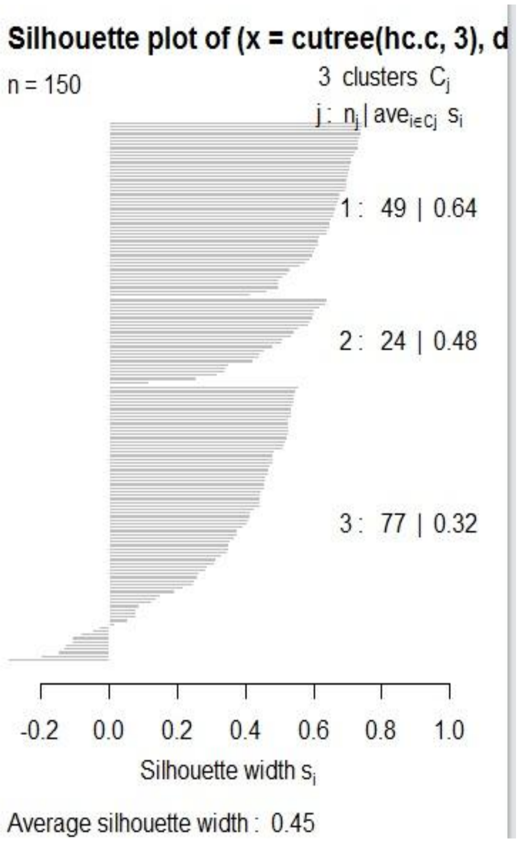


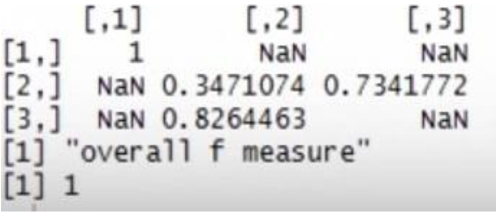
Fig.5.2 The silhouette plot

The Silhouette plot representations shows how much the clusters are closer to each other. If S_i value is high the clusters are closer to each other else not.

Fmeasure Function results

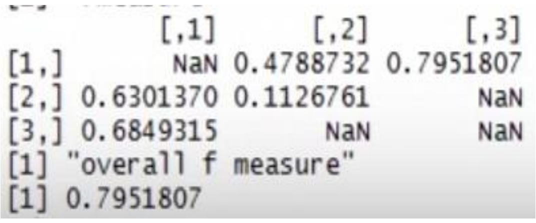
After performing both the clustering analysis, the fmeasure function is applied which calculates the precision, recall and overall f-measure of the particular algorithm.

The overall f measure given by Hierarchical algorithm is 1.



```
      [,1]      [,2]      [,3]
[1,]     1       NaN       NaN
[2,]   NaN 0.3471074 0.7341772
[3,]   NaN 0.8264463       NaN
[1] "overall f measure"
[1] 1
```

The overall f measure given by kmeans algorithm is 0.7951807



```
          [,1]      [,2]      [,3]
[1,]       NaN 0.4788732 0.7951807
[2,] 0.6301370 0.1126761       NaN
[3,] 0.6849315       NaN       NaN
[1] "overall f measure"
[1] 0.7951807
```

The overall fmeasure if 1 is considered to be the best. So, the comparative results show that Hierarchical clustering gives the best results.

5. Error Comparison Analysis

The error comparison analysis of both K-means and Hierarchical Clustering algorithm-

Hierarchical Clustering Algorithm	Clustering Error (%)
Euclidean Distance Function	34
Manhattan Distance Function	32

Table 1.2Result table of Hierarchical Clustering Algorithm.

K- means Clustering Algorithm	Clustering Error (%)
Euclidean Distance Function	11.3330
Manhattan Distance Function	10.6777

Table 1.3 Result table of Hierarchical Clustering Algorithm.

6. Conclusion

In this experiment, the problem was to predict best clustering algorithms by comparing various clustering techniques. The first half is based on clustering analysis and presenting the nature of clusters through various methods. The second half is based on comparisons of both algorithms.
Performances of these clustering methods are measured by the percentage of the incorrectly classified data instances and fmeasure function. If the overall fmeasure is 1, then the clustering algorithm is considered as the best algorithm.

Clustering has various other applications in different fields such as image processing etc. Data clustering is a data exploration method that allows objects with same characteristics to be grouped together in order to facilitate their further processing. Data clustering has various engineering application such as the recognition of part families for cellular procedure. The k-means clustering algorithm is one of the most accepted data clustering algorithms. It requires the number of cluster in the data to be pre-specified. Searching suitable number of clusters for a given data set is normally a trial-and-error process made more difficult by the subjective nature of deciding what constitutes correct clustering.

References

[1] J. Han and M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann Publisher, 2001

[2] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, M.-C. Hsu, "Free Span: Frequent Pattern-Projected Sequential Pattern Mining", Proc. 2000 Int. Conf. on Knowledge Discovery and Data Mining (KDD'00), Boston, MA, August 2000

[3] M. Spiliopoulou, L. C. Faulstich, K. Winkler, A Data Miner analysing the Navigational Behaviour of Web Users, Proceedings of workshop on Machine Learning in User Modelling of the ACAI'99, Creta, Greece, July, 1999.

[4] J. Srivastava, R. Cooley, M. Deshpande, P. Tan, and Web Usage Mining: Discovery and Applications of Usage Patterns form Web Data, SIGKDD Explorations, Vol.1, No.2, and Jan. 2000.

[5]Samah Fodeh · Bill Punch · Pang-Ning Tan (2011), ―On ontology-driven document clustering using core semantic features‖, Received: 10 December 2009 / Revised: 6 September 2010 / Accepted: 26 November 2010, Springer-Verlag London Limited 2011

[6] Third International Workshop on Advanced Issues of E-Commerce and Web-based Information Systems San Jose, CA, USA, June 21-22, 2001 http://www.chutneytech.com/wecwis2001.html

[6] Third WEBKDD workshop on data mining for web applications.

[7] Rekha Baghel and Dr. Renu Dhir (2010), ―A Frequent Concepts Based Document Clustering Algorithm‖, International journal of Computer Applications (0975-8887), Volume 4-No.5, July 2010

[8] Dhruv Gupta, Mark Digiovanni, Hiro Narita, and Ken Goldberg, "Jester 2.0: Evaluation of a New Linear Time Collaborative Filtering Algorithm", SIGIR „99 Berkley, CA, USA, ACM, 1999.

[9] Lin W., Alvarez S. A., and Ruiz C., "Efficient Adaptive Support Association Rule Mining for Recommender Systems", Data Mining and Knowledge Discovery, vol. 6, pp. 83–105, 2002. [24] Liu B., Hsu W. and Ma Y., "Mining association rules with

[10] Mobasher B., Dai H., Luo T. and Nakagawa M., "Effective personalization based on association rule discovery from web usage data", WIDM'01, USA, 2001.

[11] Mohammed J. Zaki Christopher ,D. Carothers and Boleslaw K. Szymanski, "VOGUE: A Variable Order Hidden Markov Model with Duration based on Frequent Sequence Mining ", ACM Transactions on Knowledge Discovery from Data, vol.4(1),article 5, January, 2010.

[12] Murthy, "Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey", Data Mining and Knowledge Discovery pp. 345–389, 1998.

[13] Olcay Taner Yıldız and Onur Dikmen, "Parallel univariate decision trees", Pattern Recognition Letters, Vol. 28, Issue. 7, pp. 825-832, May 2007.

[14] Pablo Loyola, Pablo E. Rom´an and Juan D. Vel´asquez, "Clustering-Based Learning Approach for Ant Colony Optimization Model to Simulate Web User Behaviour", IEEE, 20111.