Efficiency and Effectiveness of Clustering Algorithms for High Dimensional Data

Smita Chormunge
Research Scholar
GITAM University,
Hyderabad, INDIA

Sudarson Jena
GITAM University
Hyderabad, INDIA

ABSTRACT
Clustering high dimensional data is challenging due to its dimensionality problem and it affects time complexity and accuracy of clustering methods. This paper presents the F-measure and Euclidean distance based performance efficiency and effectiveness of K-means and Agglomerative hierarchical clustering methods on Text and Microarray datasets by varying cluster values. Efficiency concerns about computational time required to build up dataset and effectiveness concerns about accuracy to cluster the data. Experimental results on different datasets demonstrate that K-means clustering algorithm is favourable in terms of effectiveness where as Agglomerative hierarchical clustering is efficient in time for text datasets used for empirical study.

Keywords
Clustering, K-means, Agglomerative hierarchical, F-measure, Precision, Recall.

1. INTRODUCTION
Data mining has two fundamental task classification and clustering. Classification is supervised learning, where as Clustering is unsupervised learning method and used in statistical data analysis, pattern recognition, DNA Analysis, image analysis, information retrieval and bioinformatics. Clustering is also widely used in image segmentation, for mining text data, in spatial database, analysis of heterogeneous , Web mining, Clustering high-dimensional data for genes data [2]. Cluster analysis group's elements in such way that elements in a group should be similar to one another and unrelated to the elements in other groups. One can consider that better the clustering when there is a greater the homogeneity within a group and greater the difference between groups [3].

To cluster large dataset is computationally expensive to traditional clustering algorithms. Datasets can be large in different ways; large number of elements in the dataset; each element can have many attributes, and there can be many clusters to discover. Many of clustering algorithms suffer with dimensionality problem. Lot of research has been done in Clustering low dimensional data. When we consider high dimensional data like microarray data these clustering methods fails to handle such kind of data. There are numerous of clustering algorithms introduced for clustering data. It is broadly classified into partitioning and Hierarchical. Partitioning subdivided into K-means and K-medoids. CLARA and CLARANS are popular to deal with large datasets. Hierarchical further classified into Agglomerative and Divisive. BIRCH, Chameleon, ROCK and CURE are examples of hierarchical method which deal with large amount of data. Hierarchical clustering group's data objects in hierarchical manner, it partition cluster from singleton clusters to individual as per similarity criteria or individual to singleton cluster [4]. Other categories of clustering methods are Model based clustering, Density based clustering, Grid–based clustering, and Constrained based clustering. Some issues are address while clustering the data such as scalability to large datasets, handling high dimensional data, to work with outliers, computational time complexity to find clusters of irregular shape and data order dependency.

Recently hierarchical clustering has been adopted in word selection in the context of text classification [13]. Hierarchical clustering is better quality clustering approach for document clustering, but it has limitation because of its high time complexity. Whereas K-means have a linear time complexity which is linear in the number of documents, but are thought to produce inferior clusters. Sometimes K-means and agglomerative hierarchical approaches are combined to get the best results [1]. A gene expression data set usually contains thousands of genes. this data are often highly correlated and clusters may be exceedingly intersected with each other. In text dataset for example tr12.wc have 5803 attributes such high dimensional data can be serve obstacles for classification algorithms. To handle this data it affects the quality and efficiency of clustering algorithms.

In this paper we evaluate the performance efficiency and effectiveness of K-means and Agglomerative hierarchical clustering methods on text and microarray datasets. Analyzed datasets for different number of cluster values 10, 20 and 30 for each dataset to both clustering methods. Here 10 folds cross validation strategy used to get the precise results. For calculating computational time and Accuracy of clustering methods Euclidean distance function, F-measure, Precision and recall metrics are used. Empirical study performed on Microarray and Text datasets which have more features that ranges from 243 to 7129. Extensive experiments carried out to evaluate two clustering methods on Microarray and text datasets.

The rest of the paper is organized as follows. Section 2 describes the review of clustering methods and Evaluation metrics. In Section 3 performance evaluation of clustering methods on Microarray and text datasets are described. Experimental results discuss in Section 4. Section 5 concludes the paper.

2. CLUSTERING METHODS
In this section a review of K-means and Agglomerative hierarchical clustering methods are described.

2.1 K-means Clustering Method
The k-means algorithm is one of the popular and simple clustering method for implementation. It is partition based clustering method and used in different applications. K-means clustering method form groups without any prior knowledge objects and their relationships.

The k-means Algorithm [5]

1) Arbitrarily selects k as initial centers $C = c_1, \ldots, c_k$,
2) Set the cluster $\mathbf{c}_i$ for each $i \in \{1, \ldots, k\}$, to be the set of points in χ which is closer to $\mathbf{c}_i$ than the cluster $\mathbf{c}_j$ for all $j \neq i$.
3) For each i ∈ {1, . . . , k}, set center point in a set C_i of cluster i to $\mathbf{c}_i$. cluster center ci is recomputed as

$$\mathbf{c}i = \frac{1}{|Ci|} \sum_{x \in Ci} x$$

4) A step 2) and 3) repeats until clustering criteria meets.

K-means algorithm is popular because of two reason one is its linear complexity. The complexity of k-means is *O* (*T* * *s* * *m* * *N*) where *T* iterations performed on a sample size of *s* instances, for N attributes. Its adaptability to sparse data and best speed of convergence is also one reason of popularity of K-means [11]. But this method has some disadvantages, user have to give cluster values because the algorithm is unable to determine the appropriate number of clusters. In advanced, user has to specify cluster values which is input values to the algorithm. To get better results, user has to experiments with different values of clusters and finds the best value which is suits to their data. K-means efficiently handles nominal data and numerical data but inefficient to handle categorical data.

2.2 Hierarchical Clustering

Hierarchical clustering constructs a hierarchy. Hierarchy of different level looks like a tree which can be represented by graphical way, called dendrogram. The branches of tree monitor the groups and their similarity between the clusters. Different level of dendrogram, we get specified number of clusters, it arrange the similar objects together. Hierarchical clustering mainly classified into two types [2]:

Agglomerative: It is also called as bottom up approach: It initiates with each object forming its individual group. Similarity distance is calculated for each pair of clusters and based on this criterion, clusters are merged until termination condition reached. Clusters merge based on distance function between any two objects from different clusters.

Divisive: This is also called as a top down approach: It initiates forming one cluster to all objects and then this cluster is splits into smaller clusters by calculating distance function between objects until termination condition reached.

Simple Agglomerative Clustering Algorithm

1. Initialize the each data object is an individual cluster.
2. Compute the similarity between all pairs of clusters, i.e. calculate the similarity distance between two clusters a and b clusters.
3. Most similar clusters are merging.
4. Revise the similarity matrix to reproduce the pair-wise similarity between the original clusters and new cluster.
5. Repeats Steps 3 and 4 until cluster criterion meets.

Based on the similarity measure the hierarchical clustering methods could be further classified in to[6] *Single-link clustering* also called the nearest neighbor method, the link between two clusters whose two elements are closest to each other is made by a single element pair. *Complete-link clustering* also known as diameter, it considers the distance between two clusters whose elements are similar in same cluster but different from other cluster elements. *Average-link clustering* also known as minimum variance method, it considers the mean distance between elements of each cluster. Hierarchical clustering have some drawbacks one is its high-computational Another is lack of robustness where small change in data changes a structure of the hierarchical dendrogram. The greedy nature of this method not allows the modification for previous clustering in both approaches agglomerative and divisive [2]. Initial step of merging and splitting the cluster is important, once it cross step then it can never be corrected.

2.3 Evaluation Metrics

This section describes the metrics for evaluating performance of clustering methods. Here F-measure, Precision and Recall as a quality measure and Euclidean distance function to measure computation time is used. Depends on the measures used, the performance of different clustering method varies. Based on these measures we can consider the best clustering algorithm for the dataset which is being evaluated. Two main measures are used for measuring the quality of clustering method. One is internal quality measure which are not referring external knowledge and another is external quality measure clustering, for known classes it calculate the effectiveness of clustering methods, by comparing the groups created by the clustering algorithms. F-measure is one of the external metric which measures the effectiveness of clustering algorithms.

There are three categories for comparing clustering [7] the first is based on amount of information shared by two objects (entropy measure) or information based measures, it measures the information shared by two clustering. The second is computing recall, precision or other measures for the clusters which are most similar clusters. Third category is based on pair counting.

2.3.1 F measure

F-measure is an external measure for measuring goodness or accuracy of clustering methods. For computing F-score it depend on two factors Precision and recall. F-score is calculated by weighted average of recall and precision.

$$\text{Recall}(i, j) = N_{ij} / N_i$$

$$\text{Precision}(i, j) = N_{ij} / N_j$$

Where N_j is elements of cluster j and N_i is the number of elements of class i for class i and cluster j, N_{ij} is the numbers of elements of class i in cluster j [12].

Following equation calculates F-measure for class i and cluster j as follows

$$F(i,j) = \frac{(2 * \text{Recall}(i,j) * \text{Precision}(i,j))}{((\text{Precision}(i,j) + \text{Recall}(i,j))}$$

Calculated F measure is a result of weighted average of Precision and recall for each class i, as shown in equation 1.

$$Fc = \frac{\sum_i (|i| * F(i))}{\sum_i |i|} \quad (1)$$

Where |i| is the size of class i.

2.3.2 Euclidean Distance Function

It is a distance between two points in Euclidean space. It is computed by squared length of a vector x = [x_1 x_2] shown in

equation 2 and 3, which is a square of summation of co-ordinates. The two coordinates squared distance for example x = [x_1 x_2] and y = [y_1 y_2] is the sum of squared differences in their coordinates given in equation 4. Notation x, y is the vectors x and y, where as *d* refer the distance between two vectors x and y, it can be represented as [8]:

$$d_{x,y}^2 = (x_1 - y_1)^2 + (x_2 - y_2)^2 \quad (2)$$

The distance between two vectors is the square root

$$d_{x,y} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \quad (3)$$

Zero vector 0 = [0 0] when all coordinates of the vector are zero. In such case the distance between the vector x = [x_1 x_2] and zero vector is given by

$$d_{x,0} = \sqrt{x_1^2 + y_1^2} \quad (4)$$

The zero vector is also called as o*rigin* of the space and finally we can write $d_{x,0}$ as d_x.

3. PERFORMANCE EVALUATION OF CLUSTERING METHODS

Datasets used for empirical study discussed in this section. The summary of dataset used for empirical study shown in Table 1. We collected publically available well-known Microarray and text datasets. This study is on five microarray datasets such as Colon cancer, SRBCT, Lymphoma, CNS and Leukemia and text datasets are tr11, tr12, tr23, and DBWorld emails datasets for evaluation of clustering methods.

Gene selection problem is one of the typical application domains, which have thousands of features and it also correlated to each other. Clustering such high dimensional data is challenging task. We calculated average time required to build up datasets and quality of two clustering methods; K-means and Agglomerative clustering method. These clustering methods evaluated on collected datasets which have different features, instances and classes shown in Table 1.

Table 1. Summary of Datasets used for empirical study

Datasets	Features	Instances	Domain
DB World _subjects	243	64	Text
Colon Cancer	2000	62	Microarray
SRBCT	2308	83	Microarray
DB World _b_stem	3722	64	Text
Lymphoma	4026	62	Microarray
DB World_ bodies	4703	64	Text
tr12	5803	313	Text
tr23	5833	204	Text
tr11	6430	414	Text
Leukemia	7129	72	Microarray
CNS	29	60	Microarray

Colon Cancer [10] contains 62 samples collected from colon-cancer patients. Among them, 40 tumor biopsies are from tumors (labeled as "negative") and 22 normal (labeled as "positive") biopsies are from healthy parts of the colons of the same patients, the total number of genes to be tested is 2000.

Table 2. Evaluation of Clustering methods for clusters K=10

Datasets	K-means				Agglomerative			
	Time	Precision	Recall	F-measure	Time	Precision	Recall	F-measure
DB World _subjects	0.17	0.649	0.644	0.646	0.13	0.562	0.571	0.531
Colon Cancer	1.16	0.677	0.68	0.659	1.21	0.529	0.564	0.528
SRBCT	1.78	0.609	0.614	0.607	2.24	0.189	0.329	0.237
DB World _stemmed	0.84	0.677	0.678	0.677	1.67	0.319	0.565	0.407
Lymphoma	1.74	0.929	0.923	0.925	2.54	0.486	0.697	0.573
DB World_ bodies	1.2	0.647	0.644	0.644	2.13	0.319	0.565	0.407
tr12	20.52	0.222	0.232	0.165	10.9	0.17	0.29	0.136
tr23	28.31	0.344	0.432	0.38	5.49	0.184	0.429	0.258
tr11	49.56	0.323	0.315	0.213	19.76	0.182	0.316	0.156
Leukemia	2.61	0.794	0.795	0.794	5.6	0.438	0.662	0.527
CNS	3.55	0.5	0.5	0.5	3.57	0.51	0.593	0.523

Table 3. Evaluation of Clustering methods for Clusters K=20

Datasets	**K-means**				**Agglomerative**			
	Time	Precision	Recall	F-measure	Time	Precision	Recall	F-measure
DB World _subjects	0.13	0.792	0.732	0.741	0.14	0.509	0.534	0.484
Colon Cancer	1.27	0.785	0.733	0.747	1.09	0.487	0.574	0.467
SRBCT	2	0.776	0.742	0.738	2.24	0.304	0.373	0.313
DB World _stemmed	1.42	0.677	0.673	0.673	1.65	0.329	0.574	0.418
Lymphoma	2.77	0.965	0.962	0.961	2.52	0.486	0.697	0.573
DB World_ bodies	1.77	0.624	0.625	0.619	2.06	0.43	0.583	0.43
tr12	44.91	0.24	0.242	0.188	9.91	0.085	0.289	0.131
tr23	25.54	0.317	0.346	0.33	5.57	0.179	0.423	0.251
tr11	84.06	0.255	0.312	0.27	19.03	0.249	0.319	0.162
Leukemia	4.12	0.839	0.828	0.824	5.5	0.444	0.667	0.533
CNS	4.76	0.535	0.464	0.469	3.78	0.49	0.571	0.5

Table 4. Evaluation of Clustering methods for Clusters K=30

Datasets	**K-means**				**Agglomerative**			
	Time	Precision	Recall	F-measure	Time	Precision	Recall	F-measure
DB World _subjects	0.25	0.73	0.69	0.701	0.11	0.492	0.547	0.454
Colon Cancer	1.78	1	1	1	1.08	0.425	0.629	0.507
SRBCT	2.05	0.833	0.8	0.798	2.19	0.776	0.727	0.716
DB World _stemmed	3.17	0.882	0.872	0.87	1.65	0.332	0.576	0.421
Lymphoma	3.8	0.951	0.941	0.942	2.5	0.486	0.697	0.573
DB World_ bodies	8.16	0.679	0.679	0.645	3.98	0.332	0.576	0.421
tr12	49.79	0.217	0.237	0.202	10.77	0.084	0.29	0.131
tr23	36.56	0.327	0.296	0.305	5.52	0.178	0.422	0.25
tr11	90.84	0.245	0.3	0.265	19.03	0.247	0.316	0.159
Leukemia	5.9	0.868	0.808	0.806	5.32	0.849	0.808	0.776
CNS	6.27	0.58	0.517	0.54	3.69	0.558	0.581	0.536

SRBCT is a Gene's data which contains 2308 features and 83 samples. It is taken from the microarray experiments of Small

Round Blue Cell Tumors (SRBCT) [10]. Out of 83 samples 63 is training samples and 25 test samples. Lymphoma is a broad term encompassing a variety of cancers of the lymphatic system [10]. It contains total 4026 genes and the samples are 62. There are all together three types of lymphomas. The first category, Chronic Lymphocytic Lymphoma, the second type

Follicular Lymphoma and the third type Diffuse Large B-cell Lymphoma. CNS [10] represents a heterogeneous group of tumors about which little is known biologically. It contains 7129 genes and 42 numbers of samples. The Leukemia data set [10] contains 7129 genes on 72 samples. Two sample variants of leukemia is (AML, 25 samples, or ALL, 47 samples).

Text datasets tr12, tr11, tr23 are multi-class (1-of-n) attributes and 313 records, tr23 have 5833 features and 204 samples and tr11 contains 6430 features and 414 samples. DBWorld emails datasets [15], it is collection of 64 e-mails from DBWorld newsletter. DBWorld_bodies contains 4703 features,

DBWorld_subjects contains 243 and DBWorld_bodies_stemmed contains 3722 features.

For evaluation above datasets Weka software used it is a data mining tool [9]. The datasets are uploaded in software and calculated the computational time of each dataset, varying number of clusters as 10, 20 and 30 based on Euclidean distance function for both clustering method k-means and Hierarchical method. Further it calculates the Precision, Recall and F-measure for collected microarray datasets to observe quality of these clustering methods.

4. RESULTS AND ANALYSIS

This section present the experimental results obtained for evaluation of clustering methods varying cluster values 10, 20 and 30 on Microarray and text data based on quality metrics and distance function. Table 2 shows the tabular representation of results obtain for K-means and Agglomerative hierarchical clustering method to evaluate 10 clusters based on quality measures Precision, Recall, F-measure and Euclidean distance function to calculate time . Here 'K' denoted as number of clusters. Table 3 and Table 4 present the results of evaluation of clustering method for 20 and 30 cluster values respectively.

From Table 2 results it is observed that K-means is good for quality and takes less time for evaluation of microarray datasets. Whereas Agglomerative clustering method behaves poor in accuracy and takes more time to compute microarray datasets but takes less computational time to evaluate text datasets tr12, tr11, tr23 and DBWorld_subjects. Table 3 shows the results where k-means takes more time to evaluate text datasets than Agglomerative; in terms of accuracy K-means is better even though increase in clusters. Table 4 represent the results of clusters 30, here time and accuracy varies based on datasets. Both clustering method shows variation in performance when we increase the cluster values.

A figure 1, 3 and 5 is a comparison graph of evaluation of clustering methods based on quality measure; F-measure for Microarray and text datasets for cluster values 10, 20 and 30 respectively. K-means method works effectively in accuracy for cluster value 10. Even after increasing cluster values 20 and 30 K-means shows better performance than Agglomerative method.

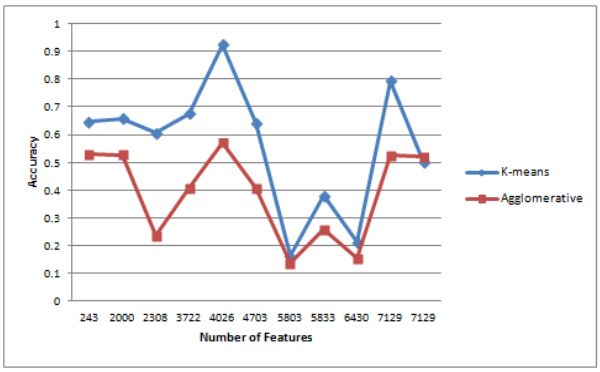

Fig. 1. Comparison graph of clustering methods based on F-measure for K=10

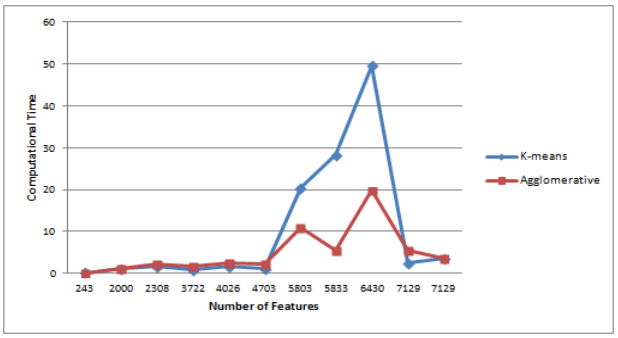

Fig. 2. Comparison graph of clustering methods based on Time for K=10

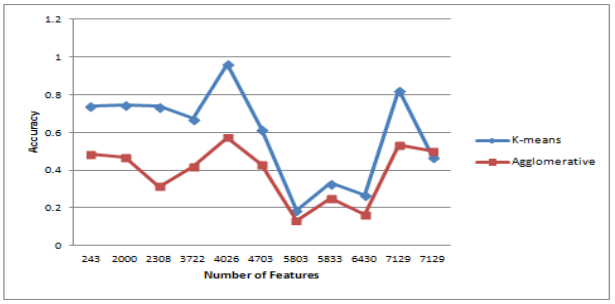

Fig. 3. Comparison graph of clustering methods based on F-measure for K=20

Figures 2, 4 and 6 is a comparison graph of evaluation of clustering methods based on Computational time for Microarray and text datasets for clusters 10, 20 and 30 respectively. Some variations are observed in results after increasing the cluster values. For cluster values 20 and 30 there is variation in time for both clustering methods. From these observations we found that the K-means is better in accuracy point of view than Agglomerative clustering method. Efficiency point of view there is variation in results for both clustering methods.

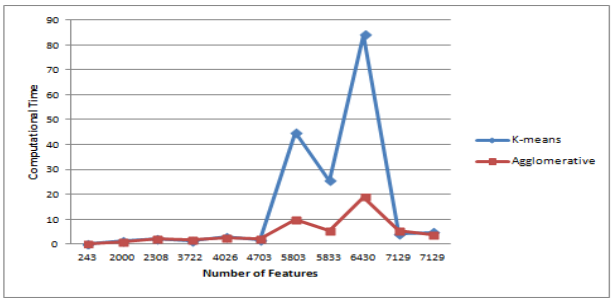

Fig. 4. Comparison graph of clustering methods based on Time for K=20

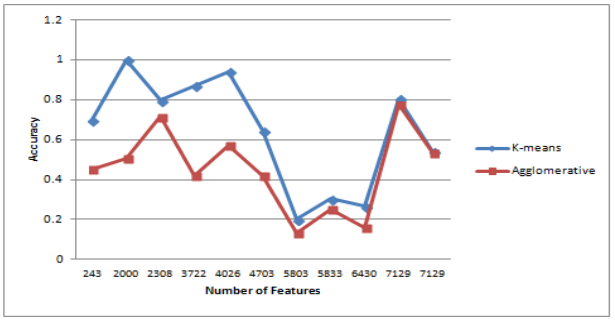

Fig. 5. Comparison graph of clustering methods based on F-measure for K=30

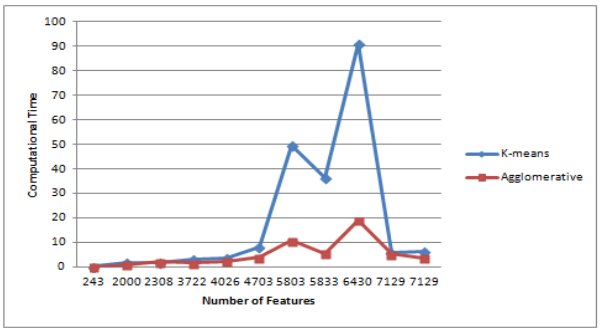

Fig. 6. Comparison graph of clustering methods based on Time for K=30

5. CONCLUSION

In this paper the performance efficiency and effectiveness of K-means and Agglomerative hierarchical clustering methods for high dimensional data based on Euclidean distance function and quality measures Precision, Recall and F-measure for different cluster values 10, 20 and 30 are evaluated. Analyzing all results it found that K-means method is effective in accuracy point of view for Microarray and Text datasets used for empirical study than Agglomerative hierarchical clustering method where as efficiency of clustering algorithms varies based on dataset used for empirical study. Further it plans to evaluate clustering algorithms for image and web data by using different quality metrics.

6. REFERENCES

[1] Michael Steinbach, George Karypis and Vipin Kumar,*A Comparison of Document Clustering Techniques.* KDD Workshop on Text Mining, 2000.

[2] Daxin Jiang,Chun Tang,Aidong Zhang,Cluster Analysis for Gene Expression Data: A survey,IEEE Transactions on Knowledge and Data Engineering,vol.16,no.11,pp-1370- 1386,Nov 2004, doi.ieeecomputersociety.org /10.1109/TKDE.

[3] Michael Steinbach, Levent Ertöz, and Vipin Kumar The Challenges of Clustering High Dimensional Data.in New Vistas in Statistical Physics – Applications in Econophysics, Bioinformatics, and Pattern Recognition, Springer-Verlag, 2004.

[4] Rui Xu and Donald Wunsch, Survey of Clustering Algorithms,IEEE Transactions On Neural Networks, pp 645-678, Vol. 16, No. 3, May 2005.

[5] Takashi Onoda, Miho Sakai, Independent Component Analysis based Seeding method for k-means Clustering, IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, 2011, DOI 10.1109/WI-IAT.2011.29.

[6] Lior Rokach, Oded Maimon,Clustering Methods Data Mining and Knowledge Discovery Handbook,Springer,2005.

[7] Elke Achtert, Sascha Goldhofer, Hans-Peter Kriegel, Erich Schubert, Arthur Zimek,Evaluation of Clusterings -- Metrics and Visual Support ,Proceedings of the 28th International Conference on Data Engineering (ICDE), Washington, DC, 2012.

[8] Michael Greenacre,Raul Primicerio Measures of Distance between Samples: Euclidean.. Fundacion BBVA publication ,ISBN: 978-84-92937-50-9 pp-47-59, December 2013.

[9] Remco R. Bouckaert,Eibe Frank,Mark Hall,Richard Kirkby,Peter Reutemann,Alex Seewald,David Scuse, WEKA Manual for Version 3-7-10,July 31, 2013.

[10] http://csse.szu.edu.cn/staff/zhuzx/Datasets.html.

[11] Dhillon I. and Modha D., Concept Decompositi-on for Large Sparse Text Data Using Clustering. Machine Learning. 42, pp.143-175. 2001.

[12] Bourennani F,Ken Q. Pu,Ying Zhu,Visualization and Integration of Databases Using Self-Organizing Map, IEEE International Conference on Advances in Databases, Knowledge, and Data Applications, pp-155-160,2009, DOI 10.1109/DBKDA.2009.30.

[13] Song Q, Jingjie Ni and Wang G, A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data, IEEE Transactions On Knowledge And Data Engineering Vol 25 No:1,2013.

[14] http://tunedit.org/repo/Data/Text-wc available at: Machine Learning & Data Mining Algorithms.

[15] https://archive.ics.uci.edu/ml/datasets/DBWorld+e-mails available at: DBWorld e-mails Data Set.