IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 1, Issue 3, June-July, 2013
ISSN: 2320 - 8791
www.ijreat.org

An Innovative Approach through New Algorithm by the Amalgamation of Clusters for High Dimensional Domain

K. Mohana Prasad[1], Dr. R. Sabitha[2]

[1]Research *Scholar, Department of CSE, Sathyabama University, Chennai 600119* ,

[2]*Head and professor, Dept of IT, Jeppiaar Engineering College, Chennai 600119*

Abstract: Clustering has become a vital area of research in this modern era. Algorithms have been applied to a greater extent to check the usability and sustainability of clusters. A cost effective, more reliable algorithm is the need of the hour and this paper aims in providing one such algorithm. The algorithm used by the researcher is unique in approach and methodology and it is proved as an effective tool to diagnose clusters. This paper aims in providing high dimensional domain through this unique algorithm. This paper also states vividly the existing algorithms their vice and virtues, usability reliability etc., this paper would undoubtedly provide a novel technique in deriving clusters, which would certainly pave way for expansion of research in the Data mining and clustering areas of research.

***Keywords*- Usability, unique, high dimensional domain, expansion of research.**

1. INTRODUCTION

Clustering is a prominent and widely used technique in Data Mining. The requirement of clustering is to find extensive structures in data and organize them into meaningful subgroups for analysis. Each clustering algorithms created is vital. It is developed using totally different techniques and research fields. Over the centuries no matter how many different algorithms are published, the most extensively used algorithm implementing clustering technique is k-means algorithm. The k-means algorithm has a few basic drawbacks such as sensitiveness, cluster size and state of the art algorithm performance in many domains can be worse but it is fast and easy to combine with other methods in larger systems.

The common problem in clustering is to treat an optimization process. Clustering algorithm has a unique approach about the assumption and the intrinsic structure of the data and the correct formula used. Hence, the exactness of this approach depends on the effectiveness and the appropriateness of the algorithm. Let us consider the example, the original k means algorithm has sum of the squared error function that uses the Euclidean distance. Another recent scientific study states that the k-means is the favorite algorithm used by the practitioners.

Although widely used k- means algorithm also has drawbacks, it is sensitive to initialization and it cannot be considered for a large number of clusters. The algorithm is used for better performance and is preferable, which has limited complexity and limited usage. But k- means is fast and easy to combine other steps also into the larger cluster of systems used.

EUCLIDEAN DISTANCE:

The Euclidean distance was one particular from a particular class called as the Bergman's divergence. It was also proposed that Bergman's hard- clustering algorithm is the kind of algorithm in which any kind of Bergman distance can be applied. Kullback-Leibler divergence is also a good example of non symmetric measure. It was also argued that the symmetric and the non negativity assumption of similarity measures was actually a state-of-the-art clustering approaches. Simultaneously, state-of-the art clustering approaches. Similarly clustering still requires more robust dissimilarity or similarity measures, recent works illustrate this.

TABLE 1
Notations

NOTATION	DESCRIPTION
N	**Number of Documents**
M	**Number of Terms**
C	**Number of Classes**
K	**Number of Clusters**
D	**Document Vector, ‖d‖**
S= {d1,d2,…..,dn}	**Set of all the documents**
Sr	**Set of Documents in Cluster r**

$D = \sum di \in S\ di$	**Composite Vector of all the Documents**
$D = \sum di \in Srdi$	**Composite Vector of cluster r**
C = D/n	**Centroid Vector of all the Documents**
$Cr = \frac{Dr}{nr}$	**Centroid Vector of Cluster r,** $nr = \|s\|$

The nature of similarity plays an important role in the success or failure of clustering method. Our first objective is to derive a novel method for measuring similarity between data objects in sparse and high-dimensional domain. The proposed systems formulate new clustering criterion functions and introduce new clustering algorithms, which are fast and scalable like k-means algorithm. Kull-back divergence was a special case of Bergman divergences used for delivering a good set of clusters that result on documented datasets. Kull- back divergence is a good example of non symmetric measure. Though the power of capturing dissimilarity in data can also be found in the discriminative power of some distance that is measured

2. RELATED WORK

The basic notations in Table 1 will be used to represent documents and related concepts .Document vectors are often subjected to some schemes, such as the standard Term Frequency- Inverse Document Frequency (TF-IDF), and have unit length. The principle of clustering is to arrange data objects into separate clusters, where intra cluster similarity and inter cluster dissimilarity is maximized. Even the state-of-the-art clustering does not have any specific form of measurement for any model. One of the popular measures of Euclidean distance used in traditional k-means algorithm to minimize the Euclidean distance between cluster's object and its centroid and is given by

$$Dist(di, dj) = \| di - dj \|$$

$$min \sum_{r=1}^{k} \sum_{di \in s} \| di - dj \|^2$$

It was also concluded that the non-Euclidean and non metric measures can be found in clustering, then by informative learning of statistical learning of data. Pelilo also argued that the symmetry and the non- negative assumption of similarity measures was actually a limitation of the state- of- the art for the clustering approaches.

In this paper we try to eliminate the limitation of these approaches. Our first objective is for obtaining new method for measuring similarity between data objects and in high- dimensional domain. In this paper we formulate new clustering functions and introduce their new clustering algorithm.

However, for data in rare and in the high-dimensional space, such as document clustering, cosine is more widely used. Particularly similarity of two document vectors di and dj, Sim (di, dj), is defined as the angle between them. For unit vectors, this equals to their inner product

$$sim\ (di, dj) = \cos(di, dj) = di^t dj$$

The Cosine measure is used in variant of k-means called spherical k- means. The Euclidean distance is minimized in the k- means algorithm. Spherical k- means tends to maximize the Euclidean distance.

$$min \sum_{r=1}^{k} \sum_{di \in Sr} \frac{d_i^t Cr}{\| Cr \|}$$

The major difference is that k- means focuses on the vector magnitudes while the spherical k- means focuses on the vector directions. Besides that the direct application on the in spherical k- means cosine has a wide area of document. Min- max cut algorithm tries to minimize the criterion function

$$min \sum_{r=1}^{k} \frac{sim\ (S_r, S\ \backslash S_r)}{Sim\ (S_r, S_r)}$$

$$where\ Sim(S_q, S_r) = \sum_{d_i \in S_q,\ d_j \in S_r} Sim(d_i, d_j)$$

Where the cosine used is minimizing the criterion and is equal to the following equation.

$$min = \sum_{r=1}^{k} \frac{D_r^t D}{\| Dr \|^2}$$

There are various graph partitioning methods with different cutting strategies and criterion functions such as average weight and normalization cut all of which can be applied to document clustering. The most prominent and popularly used clustering technique which can also be implemented in the software package is called as the CLOUTO. This method first stores the nearest neighbour graph and then splits the graph into the mini- cut algorithm. Apart from the cosine variant the Jaccard similarity can also be implemented, that can be represented in the similarity between the nearest documents.

$$u_i, u_j(d_i = {}^{u_i}/_{\| u_i \|}, d_j = {}^{u_j}/_{\| u_j \|})$$

The extended Jaccard coefficient is

$$Sim\ e_{jacc}\ (u_i, u_j) = \frac{u_i^t u_j}{\| u_i \|^2 + \| u_j \|^2 - u_i^t u_j}$$

Comparing the Euclidean distance and the cosine similarity, the extended Jaccard coefficient takes into account both the magnitude and direction of the document vectors. If the documents are instead represented by the corresponding unit vectors that also have the same value and same effect as the cosine similarity.

Strehl compared four measures: Euclidean, cosine, Pearson correlation, the extended Jaccard and concluded that extended Jaccard and the cosine are the best graphs used on web documents.

In CLUTO's graph method, the concept of similarity is different from previous discussion. The value of cosine similarity in two documents may be same but it should not connect between the neighbourhood values. Ahmad and Day composed a technique for computing the distance between two values of an attribute based on relationship with other attributes. Also LENCOET similarly described a context based distance learning method for group of data. But from the whole attribute set, they have selected only relevant subset of attributes to use as a context for calculating distance between its two values

There is a similarity in phrase- based and concept-based documents when related to text data. To identify similar documents Lakkaraju introduced a conceptual tree method. Using the above method documents are representing as concept tree with the help of classifiers.

By combining suffix tree model and vector space model, Chim and Deng have introduced a phrase- based document similarity for clustering. Later they used hierarchical agglomerative clustering algorithm to perform a clustering task. There is a drawback of computational complexity in this approach due to the needs of building suffix tree and calculating pair wise similarities explicitly before clustering. A special technique is introduced for capturing structural similarity among xml documents.

Due to simple interpretation and easy computation, through its effectiveness, the cosine similarity remains most popular. Hence a novel measure is proposed to evaluate similarity between documents and consequently formulate a new criterion for document clustering. We can also use more than one reference point to construct a new concept of dissimilarity.

3. MULTIVIEW POINT-BASED SIMILARITY

3.1 OUR NOVEL SIMILARITY MEASURE

Without changing the meaning the expression of cosine following similarity can be expressed in the form as below.

$$Sim\ (d_i, d_j) = \cos(d_i - 0, d_j - 0) = (d_i - 0)^t\ (d_j - 0)$$

Here 0 indicates vector 0 representing the origin point. The one and only measure from the above formula indicates 0. From the point of origin we can find the similarity between two documents d_i *and* d_j with respect to the angle between two points. We can also use more than one point of reference to construct a new concept of similarity. If we look the points from different points of view, we can make conclusion of how close or distant a pair of points are located.

The direction and distance to d_i *and* d_j are represented with the third point d_h as difference vector$(d_i - d_h)$ *and* $(d_j - d_h)$. We can also define the similarity between two points as below.

$$Sim(d_i, d_j)_{d_i, d_j \in S_r} = \frac{1}{n - n_r} \sum_{d_h \in S / S_r} Sim\ (d_i - d_h, d_j - d_h)$$

The similarity above is defined as a closed relation to the clustering problem. From the same cluster we have to measure two objects, but the points from where to establish this measure must be outside of cluster. This method is called Multi View Point based similarity or MVS. Now, the similarity measure of document vector d_i *and* d_j between two points can also be determined as $MVS(d_i, d_j \mid d_i, d_j \in S_r)$ *or* $MVS(d_i, d_j)$

In vector, the final form of MVS is determined by dot product of different vectors as given below:

$$MVS(d_i, d_j \mid d_i, d_j \in S_r)$$

$$= \frac{1}{n-n_r}\sum_{d_h \in S/S_r}(d_i - d_h)^t\ (d_j - d_h)$$

$$= \frac{1}{n-n_r}\sum_{d_h} Cos\,(d_i - d_h, d_j - d_h) \parallel d_i - d_h \parallel \parallel d_j - d_h \parallel$$

The product of cosine of angle between d_i *and* d_j *from* d_h and Euclidean distance from d_h to two points is equal to the similarity between two points d_i *and* d_j inside clusterS_r, viewed from a point outside the cluster. The above point is assumption that d_h is not in same cluster withd_i *and* d_j. But if the distance between $\parallel d_i - d_h \parallel$ and $\parallel d_j - d_h \parallel$ is small, there is a chance that d_h also belongs to the same cluster S_r which contains d_i and d_j will provide dissimilarity of inter cluster by taking the average over all the view points of S_r, we can be able to find the dissimilarities between d_i and d_j which does not belongs to S_r. But sometimes it will give misleading information if it starts from origin point. To reduce the effect of misleading, view points are constrained by taking the averaging steps. Therefore this type leads to an assessment of similarity that single origin point is based on similarity measure.

3.2 ANALYSIS AND PRACTICAL EXAMPLES OF MVS

For Clustering of data, MVS could be very effective. To illustrate that MVS is compared with cosine similarity of group of structure in document collection.

$MVS\ (d_i, d_j \mid d_i, d_j \in S_r)$

$$= \frac{1}{n-n_r}\sum_{d_h \in S/S_r}\{d_i^t d_j - d_i^t d_h - d_j^t d_h + d_h^t d_h)$$

$$= d_i^t d_j - \frac{1}{n-n_r} d_i^t\, D_S/S_r - \frac{1}{n-n_r} d_i^t\, D_S/S_R + 1$$

$$= d_i^t d_j - d_i^t C_{S/S_r} - d_j^t C_{S/S_r} + 1$$

The outer composite vector with cluster r, is the composite vector of all documents outside cluster r, and is defined as$D_{S/S_r} = \sum_{d_h \in S/S_r} d_h$. The outer centroid vector of cluster r is defined as $C_{s/S_r} = D_{S/S_r}/(n - n_r)$.

$MVS(d_i, d_j)$ *and* $MVS(d_i, d_l)$ can be declared as

$$d_i^t d_j - d_j^t C_{S/S_r} > d_i^t d_l - d_i^t C_{S/S_r}$$

$$\Leftrightarrow COS(d_i, d_j) - COS(d_j, C_{S/S_r})\left\|C_{S/S_r}\right\| > COS(d_i, d_l) - COS(d_l, C_{S/S_r})|C_{S/S_r}|$$

The above condition is applicable when $d_l$is closer to d_i where$COS(d_i, d_j) < COS(d_i, d_l)$. The outer centroid value of C_{S/S_r} is closer to d_j based on MVS. If d_l is closer toC_{S/S_r}, then there is a chance that it belongs to another cluster and it is closer tod_i.

The validity test is proposed for MVS and Cs to prove it further. A Similarity matrix$A = \{a_{ij}\}_{n\times n}$ is created for each type.

The procedure to construct MVS value is as follows:

STEP 1: Construct a MVSMATRX (A).
STEP 2: Assign Centroid vector for set of all documents in Cluster r.
STEP 3: Find the sum of d_i values; where i$\in S_r$.
STEP 4: Find the Number of documents for set of all documents in cluster r.
STEP 5: If the number of Documents=1 in the first set i, then initialize the document vector i into cluster vector.
STEP 6: If the number of Documents=1 in the second set j, and second set belongs to set of documents in cluster r, then find $Sim(d_i, d_j)$ and initialize to a_{ij} matrix.
STEP 7: Otherwise find $COS(d_i, d_j)$value and initialize to a_{ij} matrix.
STEP 8: Return matrix$A = \{a_{ij}\}_{n\times n}$ to A
STEP 9: Stop the procedure of MVSMATRIX.

After constructing the MVSMATRIX, select the q_r value for each document d_i of a row a_i in A. Then get the validity score ofd_i. The final value of validity can be calculated by average value of rows A. For the Clustering task, the Similarity measure has the higher value. To prove this Task by validity test, we are taking the example of two real world document Data Sets. The famous collection of *reuters7* subset is chosen as first example. The Distribution 1.0 of Reuters-21578 is news article. For Text Categorization Reuters-21578 is widely used. To form reuters7 2500 documents are selected from largest 7 categories.

Another data set i.e. second set is K1b. It has the collection of 2340 WEBPAGES from yahoo and is now available in CLUTO toolkit. The final session of the document were weighted by TF-IDF and is converted into Unit Vector. The set of documents were pre-processed by stop-word removal and stemming. The diagrammatic representation of characteristics of reuters7 and K1b are shown below.

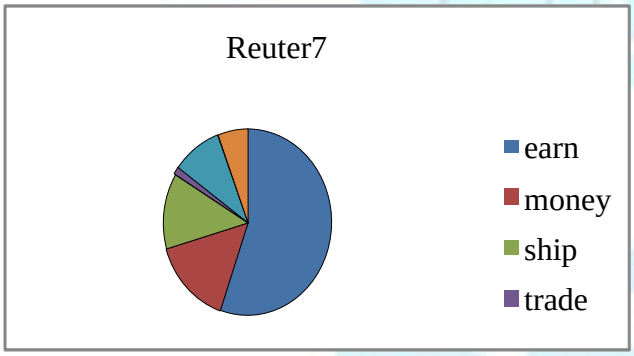

Fig 1. Characteristics of reuters7

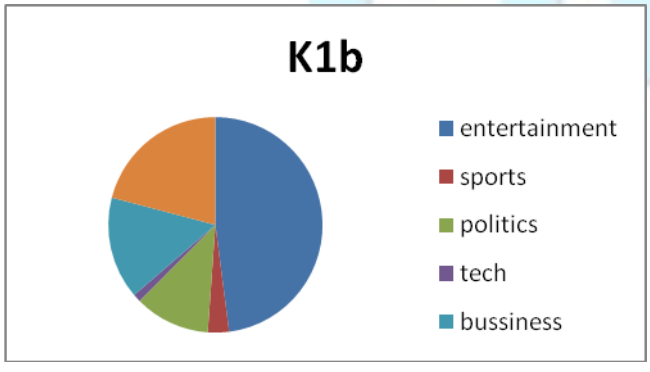

Fig 2. Characteristics of K1b data sets

On computing the validity score of Cs and MVS on two data sets it has a percentage parameter. Finally the validity test proves that MVS is clearer than Cs for both data sets. We can also make conclusion that any two sets of data of equal size has 67% document neighbours based on Cs and 80% based on MVS. Finally the new Multi view-point based similarity has more advantage than cosine measure. The validity test of MVS and CS is as shown below.

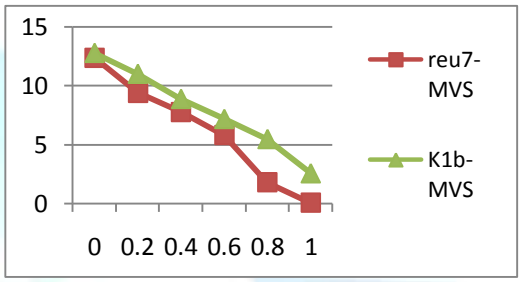

Fig.3. MVS Validity Test

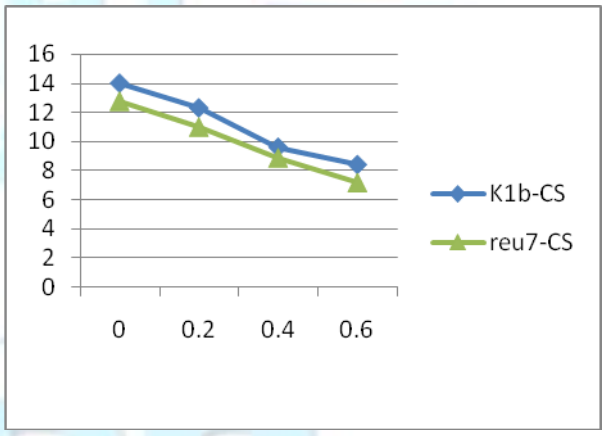

Fig.4. CS Validity Test

4. MULTIVIEW POINT-BASED CLUSTERING

4.1 Two Clustering Criterion Functions I_r and I_v

Based on similarity measure, the clustering criterion function is formed. The weighted sum of average pair wise similarities of documents in the same cluster is defined asI_r. In general form Function F is expressed as follows.

$$F = \sum_{r=1}^{k} n_r \left[\frac{1}{n_r^2} \sum_{d_i, d_j \in S_r} Sim(d_i, d_j)\right]$$

To perform in simple, fast and efficient way the above function can be declared using optimization procedure as

$$\sum_{d_i d_j \in S_r} Sim(d_i, d_j)$$

$$= \sum_{d_i d_j \in S_r} d_i^t d_j - \frac{2n_r}{n - n_r} \sum_{d_i \in S_r} d_i^t \sum_{d_h \in S / S_r} d_h + n_r^2$$

$$= \frac{n + n_r}{n - n_r} \| D_r \|^2 - \frac{2n_r}{n - n_r} D_r^t D + n_r^2$$

Therefore

$$F = \sum_{r=1}^{k} \frac{1}{n_r} \left[\frac{n + n_r}{n - n_r} \|D_r\|^2 - \left(\frac{n + n_r}{n - n_r} - 1\right) D_r^t D\right] + n$$

Since n is constant maximizing $\bar{F}$ can be declared as below

$$\bar{F} = \sum_{r=1}^{k} \frac{1}{n_r} \left[\frac{n + n_r}{n - n_r} \|D_r\|^2\right] - \left(\frac{n + n_r}{n - n_r} - 1\right) D_r^t D$$

While $\bar{F}$ is compared with min-max cut, it contains an intra cluster similarity measure $\|D_r\|^2$ and inter cluster similarity measure $D_r^t D$. The aim of this procedure is to maximize the weighted difference. The value of $\bar{F}$ is weighted by taking the inverse of cluster size. This procedure is quite sensitive to cluster size. We know that from clustering algorithm the set of weight factor $\lambda = \{\lambda_r\}_1^k$ is used to regulate the size of cluster in clustering solution.

$$\overline{F_\lambda} = \sum_{r=1}^{k} \frac{\lambda_r}{n_r} \left[\frac{n + n_r}{n - n_r} \|D_r\|^2\right] - \left(\frac{n + n_r}{n - n_r} - 1\right) D_r^t D$$

Let us consider $\lambda_r = n_r^{\alpha}$, then the criterion function I_R is defined as

$$I_R = \sum_{r=1}^{k} \frac{\lambda_r}{n_r^{1-\alpha}} \left[\frac{n + n_r}{n - n_r} \|D_r\|^2\right] - \left(\frac{n + n_r}{n - n_r} - 1\right) D_r^t D$$

The above results will give a good result for clustering when $\alpha \in (0,1)$. The cluster quality I_R leads sensitiveness to size and tightness of the cluster. Instead of cluster's centroid an alternative approach is used to prevent the pair wise similarity between MVS based on cluster and CS. So the Objective Function G is defined as below.

$$G = \sum_{r=1}^{k} \left[\frac{n + \|D_r\|}{n - n_r} \|D_r\| - \left(\frac{n + \|D_r\|}{n - n_r} - 1\right) \frac{D_{rD}^t}{\|D_r\|}\right] + n$$

Since n is constant, we can eliminate that value. The maximizing G is equivalent to maximizing I_V as described below.

$$I_V = \sum_{r=1}^{k} \left[\frac{n + \|D_r\|}{n - n_r} \|D_r\| - \left(\frac{n + \|D_r\|}{n - n_r} - 1\right) \frac{D_{rD}^t}{\|D_r\|}\right]$$

The weighted difference between $\|D_r\|$ and $D^{D_r^t}/_{\|D_r\|}$ are calculated using I_V which represent intra and inter cluster similarity. To optimize the performance of cluster GREEDY algorithm has been introduced.

4.2 OPTIMIZATION ALGORITHM AND COMPLEXITY

Since clustering framework is defined as MVSC, we have to define its criterion function as MVSC-I_R and MVSC-I_V. Its main aim is to optimize document clustering using I_R and I_V. In general I_V can be follow as below.

$I_V = \sum_{r=1}^{k} I_r(n_r, D_r)$; Where $\sum_{r=1}^{k} I_r(n_r, D_r)$ is the objective value of cluster r. Similarly I_R can also be described as I_V.

The above general form has two major steps as Initialization and Refinement. The aim of

initialization is to select the initial parameters from K – arbitrary documents. The number of iterations is defined in refinement. The n documents were visited one by one in random order for iteration. While checking, if the document has an improvement, it moves to another cluster which has highest improvement. If it does not show better result than current cluster, then it won't move to the other cluster. The iteration process can also terminate without moving the document to the new cluster.

In this process k-means will update after all n documents have reassigned, but the incremental cluster algorithm immediately updates.

The cost optimization procedure is described below.

1. We have to search for an optimum cluster to move individual document O :(nz.k).
2. The next process is to update composite vector as O: (m.k).
   Where nz is the total no of non- zero entities in vector document.

The value of nz is 10 times larger than m for document domain. The computational complexity required for clustering with I_R and I_V is O (nz.k.τ); where τ is the number of iterations.

5. PERFORMANCE EVALUVATION OF MVSC

The advantage of our proposed system can be evaluated based on their performance in experiment on data. the main aim of this part is to compare MVSC-I_R and MVSC-I_V with the existing algorithm. The similarity measure of existing algorithm has Euclidean distance, cosine similarity and extended Jaccard co-efficient.

5.1 DOCUMENT COLLECTION

The data sets of reuter7 and k1b as described above also includes 18 documents of text collection in clustering and they combine with CLUTO by the toolkit's author. They were already defined by standard procedures including stop – word removal, stemming, and removal of too rare and too frequent words, TF-IDF weighting and normalization.

5.2 EXPERIMENTAL SETUP AND EVALUVATION

The performance of MVSCS can be illustrated by comparing them with five other clustering with some data sets. The seven clustering algorithms used for comparing clustering methods are given below.

1. MVSC using criterion function I_R : MVSC-I_R.
2. MVSC using criterion function I_V : MVSC-I_V.
3. Standard K-mean with Euclidean distance: k-means.
4. Spherical K-mean with Euclidean distance: Spk-means.
5. CLUTO's graph method with CS: graph-CS.
6. CLUTO's graph method with Extended Jaccard: graph-EJ.
7. Spectral Min-Max cut Algorithm: MMC.

The programs of MVSCS-I_R and MVSCS-I_V can be implemented in Java. While performing the regulating factor α is set to 0.3 as appropriate value inI_R. The CLUTO toolkit contains other algorithm in c library interface at free of cost. The cluster number is predefined for each data set and is equal to number of true class; k=c.

Since the above declared algorithms are initialization dependent, we can't able to find the global optimum value. The trial and error method is applied on seven clustering algorithm randomly to choose the best of value of objective function. Each test runs consist of 10 trails on clustering method and average value is calculated for test – runs. The clustering solution is evaluated by comparing the document after a test run. To illustrate the performance of clustering three types of external evaluation method are used. They are

1. FScore
2. Normalized Mutual Information (NMI) and
3. Accuracy

The weighted combination of Precision 'p' and Recall 'R' values used in information retrievals gives the FScore and is determined as

$$FScore = \sum_{i=1}^{k} \frac{n_i}{n} max_j(F_{ij})$$

$$F_{ij} = \frac{2 \times P_{ij} \times R_{i,j}}{p_{i,j} + R_{i,j}}$$

$$P_{i,j} = \frac{n_{i,j}}{n_j}$$

$$R_{i,j} = \frac{n_{i,j}}{n_i}$$

n_i = no of documents in class i.

n_j = no of document assigned to cluster j.

$n_{i,j}$ = no of documents shared in class i and cluster j.

The partition of true class and cluster assignment shares the information of NMI measure.

$$NMI = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} n_{i,j} \log\left(\frac{n.n_{i,j}}{n_i n_j}\right)}{\sqrt{\left(\sum_{i=1}^{k} n_i \log\frac{n_i}{n}\right)\left(\sum_{j=1}^{k} n_j \log\frac{n_j}{n}\right)}}$$

Then the fraction of documents labels the accuracy measure and it has one – to – one relation between true classes and clusters.

$$\text{Accuracy} = \frac{1}{n} max_q \sum_{i=1}^{k} n_{i,q(i)}$$

Using Hungarian algorithm the accuracy value of q has found. The above 3 metrics has the range from 0 to 1 and the better solution of clustering will yield a greater value.

5.3 RESULTS

The result of clustering is based on FScore and NMI value. The accurate value of 7 clustering algorithm for some data sets are shown in below diagram.

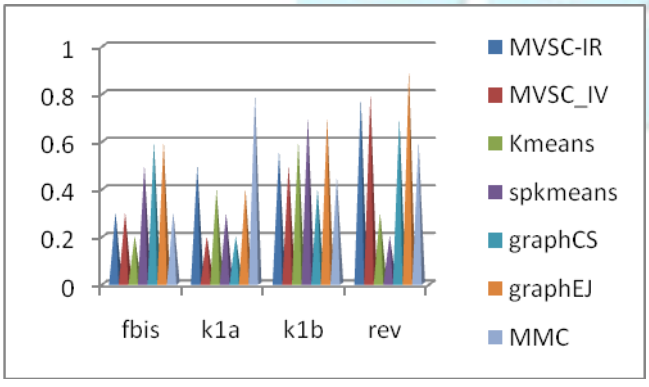

Fig.5. Clustering Results in Accuracy

In the above diagram graphEJ will produce outstanding result on classic and on reviews MMC and graphics, yields a good result. The other data does not yield a good results based on spk mean. To yield a good result both of the MVSC approaches in top two algorithms. To perform a paired t-test, MVSC-Ir and MVSC-Iv was paired with remaining algorithm to carry out a statistical justification on clustering. The P-value of the paired t-test is calculated and the value is less than 0.05, it yields a significant value else the comparison is significant. The paired t-test yields the advantage of MVSC-Ir and MVSC-Iv is statically significant over other methods. So it overcomes the result of graphEJ.

5.6 EFFECT OF α ON MVSC-IR'S PERFORMANCE

The partitional clustering method is sensitive to cluster size and balance based on criterion function. The parameter **α** which exists in I_R is called a regulating factor, **α**$\in (0,1)$. Since the metrics evaluation of data sets is not meaningful by taking average value of datasets. So we have to transform the metrics into relative metrics before taking the average value of datasets.

The relative FScore value of MVSC-I_R can be found by

$$relative_{-FScore} = \frac{max_{\alpha_i}\{FScore(I_R; S, \alpha_i)\}}{FScore(I_R; S, \alpha_i)}$$

S= document Collection

$\alpha = \alpha_i$

We can also apply the same procedure to NMI and Accuracy to get the relative- NMI and relative-Accuracy. If the value of α_i is 1, MVSC-I_R yields good result. If it is greater than 1, MVSC-I_R yields worse result.

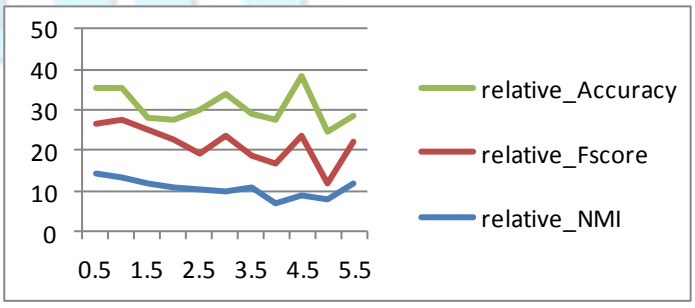

Fig.6. MVSC-IR'S PERFORMANCE WITH RESPECT TO α

If the range of **α is** from 0.3 to 0.7 regarding the **t**ype of evaluation metrics, MVSC-I_R yields 5%best case result.

6 CONCLUSION AND FUTURE WORK

The MVS method has been proposed in this paper and is named as Multi-view point based on similarity. We have proved that MVS is suitable for test documents when compared to cosine similarity based on theoretical analysis and empirical equations. The two criterion functions and similar clustering algorithms are based on MVS were found in this paper. To improve the performance of clustering we have proposed a different clustering algorithm based on similarity measure.

The main aim of this paper is to find the fundamental concept of similarity measure from Multi-view points. The partition clustering of documents are focused mainly in this paper. For hierarchal clustering algorithms, it is possible to apply the criterion function in future. The applications of MVS and clustering algorithm were also shown finally. In this paper we also explore that how they work on high dimensional domains and we have explored only some sets of data for application in future.

REFERENCES

1. D. Lee and J. Lee, "Dynamic Dissimilarity Measure for Support Based Clustering," IEEE Trans. Knowledge and Data Eng., vol. 22, no. 6, pp. 900-905, June 2010.
2. M. Pelio, "What is a cluster? Perspectives from Game Theory, proc. Clustering Theory, 2009.
3. D. Ienco, R.G. Pensa and R. Meo, "Context-Based Distance Learning for Categorical Data Clustering," Proc. Eighth Int'l Symp. Intelligent Data Analysis (IDA), pp. 83-94, 2009.
4. P. Lakkaraju, S. Gauch, and M. Speretta, "Document Similarity Based on Concept Tree Distance," Proc. 19th ACM Conf. Hypertext and Hypermedia, pp. 127-132, 2008.
5. H. Chim and X. Deng, "Efficient Phrase-Based Document Similarity for Clustering," IEEE Trans. Knowledge and Data Eng., vol. 20, no. 9, pp. 1217-1229, Sept. 2008.
6. A. Ahmad and L. Dey, "A Method to Compute Distance Between Two Categorical Values of Same Attribute in Unsupervised Learning for Categorical Data Set," Pattern Recognition Letters, vol. 28, no. 1, pp. 110-118, 2007.
7. Y. Gong and W. Xu, Machine Learning for Multimedia Content Analysis. Springer-Verlag, 2007.
8. S. Zhong, "Efficient Online Spherical K-means clustering," Proc.IEEE Int'l Joint Conf. Neural Networks (IJCNN), pp. 3180-3185, 2005.
9. J. Friedman and J. Meulman, "Clustering Objects on Subsets of Attributes," J. Royal Statistical Soc. Series B Statistical Methodology, vol. 66, no. 4, pp. 815-839, 2004
10. S. Zhong and J. Ghosh, "A Comparative Study of Generative Models for Document Clustering," Proc. SIAM Int'l Conf. Data Mining Workshop Clustering High Dimensional Data and Its Applica- tions, 2003.
11. C. Ding, X. He, H. Zha, M. Gu, and H. Simon, "A Min-Max Cut Algorithm for Graph Partitioning and Data Clustering," Proc. IEEE Int'l Conf. Data Mining (ICDM), pp. 107-114, 2001.
12. I. Dhillon and D. Modha, "Concept Decompositions for Large Sparse Text Data Using Clustering," Machine Learning, vol. 42, nos. 1/2, pp. 143-175, Jan. 2001.