International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 9 Issue 3 –SEPTEMBER 2014.

High Quality Assessment of Similarity by Using Multiple View Points

C.S Mahaboobbee[*1], and M.Venkatesh Naik[#2]

[*]*Student, Dept of CSE, CRIT, Affiliated to JNTUA University, ANANTAPURAMU, AP, India*

[#] *Asst Professor, Dept of CSE, CRIT, Affiliated to JNTUA University, ANANTAPURAMU, AP, India*

[1]sadiyaghouse786@gmail.com

[2]venkateshnaikm0@gmail.com

***Abstract*— Data grouping, data partitioning and hierarchical clustering are the three types of well known clustering methods. The data grouping approach is meant for making a set of overlapping clusters. The K-means algorithm, a kind of partitioned clustering needs dataset and number of clusters as two required arguments. Credit card fraud detection is one of the areas in which K-means is being used. Evnethough it is simple and effective algorithm, it suffers from the drawbacks of low performance, initialization and sensitive to cluster size. Hence always, The process of clustering with highest quality is an optimization process. In order to attain highest quality clusters, we go for Similarity measure approach.**

In this paper we propose a multi-view point based similarity measure which considers multiple viewpoints while comparing objects for clustering. This measure can have more informative assessment of similarity thus making clusters with highest quality. We also proposed two criterion approaches for achieving highest intra-cluster similarity and lowest inter-cluster similarity.

Keywords: Similarity measure, text mining, document clustering

I. INTRODUCTION

Clustering data described by categorical attributes [3] is a challenging task in data mining applications. Unlike numerical attributes, it is difficult to define a distance between pairs of values of the same categorical attribute, since they are not ordered. Clustering is a popular data mining technique that enables to partition data into groups (clusters) in such a way that objects inside a group are similar, and objects belonging to different groups are dissimilar. When objects are described by numerical (real, integer) features, there is a wide range of possible choices. Objects can be considered as vectors in a n-dimensional space, where n is the number of features. Then, many distance metrics can be used in n-dimensional spaces. Clearly, these distance metrics do not distinguish between the different values taken by the attribute, since they only measure the equality between pair of values. This is a strong limitation for a clustering algorithm, since it prevents to capture similarities that are clearly identified by human experts.

Clustering is "unsupervised classification" or "unsupervised segmentation". The aim is to assign instances to classes that are not defined a priori and that are supposed to somehow reflect the underlying structure" of the entities that the data represents. Most of the problems encountered with the clustering algorithms involve dealing with the large number of dimensions and a large number of objects becoming prohibhitive due to time complexity, the effectiveness of an algorithm depends upon the similarity measure.

II. EXISTING SYSTEM

Document clustering is a form of text mining meant for grouping documents into various clusters. A document is treated as an object a word in the document is referred as a term. A vector is built to represent each document. The existing document clustering algorithms include probabilistic based methods , nonnegative matrix factorization and information theoretic co-clustering.

The most widely used clustering algorithm [1] is ED-Euclidean distance which is measured as

$$\mathbf{Dist}\ (\mathbf{d_i}, \mathbf{d_j}) = \|\mathbf{d_i} - \mathbf{d_j}\|$$

Based on the ease of use and the simplicity, K-Means is most widely used clustering algorithm. ED is the measure used in K-Means algorithm to measure the distance between objects to make them into clusters. The cluster centroid is computed as

$$\mathbf{Min} \sum_{r=1}^{k} \sum_{d_i \in S_r} \|\mathbf{d_i} - \mathbf{C_r}\|^2$$

Cosine similarity measure is another algorithm used in hi-dimensional documents. This measure is also being used in Spherical K-Means which is a variant of K-Means. The

difference between the two flavors of K-Means that use cosine similarity measure and ED measure respectively is that the former focuses on vector directions while the latter focuses on vector magnitudes. Graph partitioning is yet another approach in which It considers the document corpus as graph and uses min-max cut algorithm which represents centriod as:

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

Another graph partitioning approach CLUTO documents are clustered based on the nearest neighbor graph

$$\text{Sim}_{gJacc}(u_i,u_j) = \frac{u_i^t u_j}{\|ui\|^2 + \|uj\|^2 - utiuj} \quad (4)$$

For document clustering other approaches can be used which are phrase based and concept based. The common algorithm used by both of them is "Hierarchical agglomerative Clustering". The drawback of these approaches is that their computational cost is very high. For clustering XML documents also there are measures. One such measure is named "Structural Similarity" which differs from text document clustering.

III. Proposed System

The implemented work in this paper is based on Multi-view point based similarity measure. It does mean that it uses more than one view point while finding similarity between objects and clustering them into various groups. The similarity between the two documents can be given as

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

where di and dj are the two points in cluster Sr, dh is considered the similarity between them which is equal to cosine angle of ED of those points.

The procedure for similarity matrix is as given below:

1. Procedure BUILDMVSMATRIX(A)
2. For r ← 1 : c do
3. $D_{s/sr} \leftarrow \Sigma_{di \epsilon Sr}$ di
4. Ns/sr ← |S\Sr|
5. End for
6. For r ← 1 : n do
7. R ← class of di
8. For j ← 1 : n do
9. If dj ϵ Sr then
10. aij ← $d_j^t dj - d_i^{t\,Ds/Sr}{}_{nS/Sr} - d_j^{t\,Ds/sr}{}_{nS/Sr} + 1$
11. else
12. aij ← $d_j^t dj - d_i^{t\,Ds/Sr}{}_{nS/Sr} - d_j^{t\,Ds/sr - Dj}{}_{nS/Sr\,-1} + 1$

end if
end for
end for
return A={a_{ij}} mxn
end procedure

Algorithm 1: Procedure for Similarity matrix

The validity is calculated as an average of all the rows. If validation score is higher, it reflects that the similarity is higher and thus eligible for clustering.

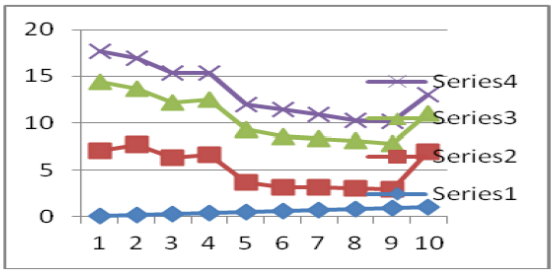

Fig 1 : Validity of Cosine similarity and Multi-view point Based Similarity

From the above figure it can be clearly observed that the performance of Multi-view point based similarity is higher when compared to Cosine Similarity.

The incremental clustering algorithm for clustering the documents has been implemented in two phases: Refinement and Initialization.

Initialization involves selecting k documents as seeds for making the initial positions. The refinement phase makes each iteration to form best clusters. Each iteration in refinement phase visits n number of documents in random fashion. Once the verification process is done for each document, it is moved to the cluster if the document is considered to be similar. When no documents are there the iterations come to an end.

The following Bench mark datasets have been used to test the efficiency of our approach:

Data	Source	c	n	m	Balance
fbis	TREC	17	2,463	2,000	0.075
hitech	TREC	6	2,301	13,170	0.192
k1a	WebACE	20	2,340	13,859	0.018
k1b	WebACE	6	2,340	13,859	0.043
la1	TREC	6	3,204	17,273	0.290
la2	TREC	6	3,075	15,211	0.274
re0	Reuters	13	1,504	2,886	0.018
re1	Reuters	25	1,657	3,758	0.027
tr31	TREC	7	927	10,127	0.006
reviews	TREC	5	4,069	23,220	0.099
wap	WebACE	20	1,560	8,440	0.015
classic	CACM/CISI/ CRAN/MED	4	7,089	12,009	0.323
la12	TREC	6	6,279	21,604	0.282
new3	TREC	44	9,558	36,306	0.149
sports	TREC	7	8,580	18,324	0.036
tr11	TREC	9	414	6,424	0.045
tr12	TREC	8	313	5,799	0.097
tr23	TREC	6	204	5,831	0.066
tr45	TREC	10	690	8,260	0.088
reuters7	Reuters	7	2,500	4,977	0.082

c: # of classes, n: # of documents, m: # of words
Balance= (smallest class size)/(largest class size)

Table 1: Benchmark documents datasets

The evaluation results are best compared to M-means, Min Max Cut Algorithm, graph EJ CLUTO's graph with extended Jacquard, graphCS which is nothing but CLUTO's graph with Cosine Similarity, SpkMeans which is nothing but Spherical K-Means with Cosine Similarity, MVSC Ir the proposed algorithm with Ir iteration.

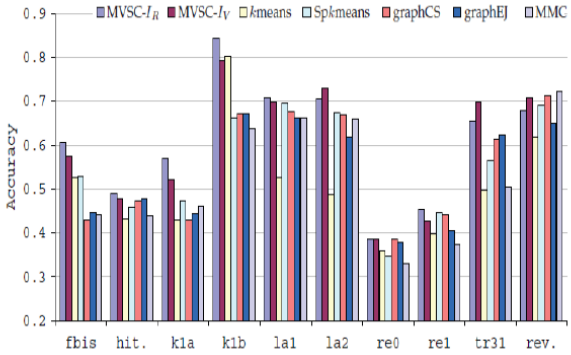

Fig 2 : Evaluation results for different clustering algorithms for first 10 data sets

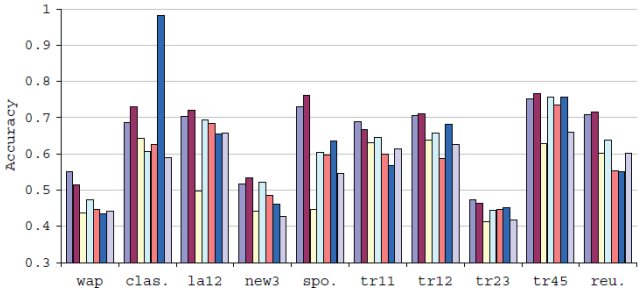

Fig 3: Evaluation results for different clustering algorithms for next 10 data sets

IV. CONCLUSION

The similarity measure is capable of providing informative assessment and bestows high quality clusters. The proposed approach achieves highest similarity between objects of same cluster and lowest similarity between the objects of different clusters. The data for experimental evaluation considers benchmark datasets.

REFERENCES

[1] K.Ramesh, C.Vasumurthy, Prof.D.Venkatesh, High Quality Assessment of Similarity by Using Multiple View Points, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering Vol. 2, Issue 7, July 2013

[2] I. Guyon, U. von Luxburg, and R. C. Williamson, "Clustering:Science or Art?" *NIPS'09 Workshop on Clustering Theory*, 2009.

[3] D. Ienco, R. G. Pensa, and R. Meo, "Context-based distance learning for categorical data clustering," in Proc. of the 8th Int. Symp. IDA, 2009, pp. 83– 94.

[4] Leo Wanner (2004). "Introduction to Clustering Techniques".Available online at: http://www.iula.upf.edu/materials/040701wanner.pdf [viewed: 16 August 2012].

[5] 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 Recognit. Lett., vol. 28, no. 1, pp. 110 – 118, 2007.