Volume 5, Issue 8, August 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

# Multi-View Point based Similarity Measure Using Fuzzy-C Means Clustering

[1]Brahmani Parvataneni, [2]K V Sambasiva Rao
[1]M.Tech Student, NRI Institute of Technology, Pothavarappadu, Andhra Pradesh, India
[2]Dean, Department of CSE, NRI Institute of Technology, Pothavarappadu, Andhra Pradesh, India

***Abstract:*** ***Theoretical analysis and scientific illustrations display that Multi-Viewpoint based Similarity, or MVS is possibly more suitable for written text records than the well-known cosine similarity. Depending on MVS, two requirements features, $I_R$ and $I_V$, and their specific clustering techniques, MVSC-$I_R$ and MVSC-$I_V$, have been presented. In contrast to other state-of-the-art clustering techniques that use different types of likeness evaluate, on a huge variety of papers, data sets and under different assessment analytics, the proposed algorithms display that they could offer significantly improved clustering efficiency. In this paper we propose to develop efficient and effective clustering algorithm for processing similar type of data items in application framework based on related topics present in processed data sets. Our experimental results show efficient and communication related to application development in real time processes.***

***Index Terms: Fuzzy C-means Clustering, Datasets, and Multi View Point Clustering.***

## I. INTRODUCTION

Clustering is one of the most exciting and important topics in information exploration. The aim of clustering is to find intrinsic components in information, and arrange them into meaningful subgroups for further study and analysis. There have been many clustering methods released every year. They can be suggested for very unique analysis areas, and developed using absolutely different approaches and methods. Nevertheless, according to majority of folks [1], more than 50 years after it was presented, the simple algorithm k-means still continues to be as one of the top 10 data mining methods these days [2].

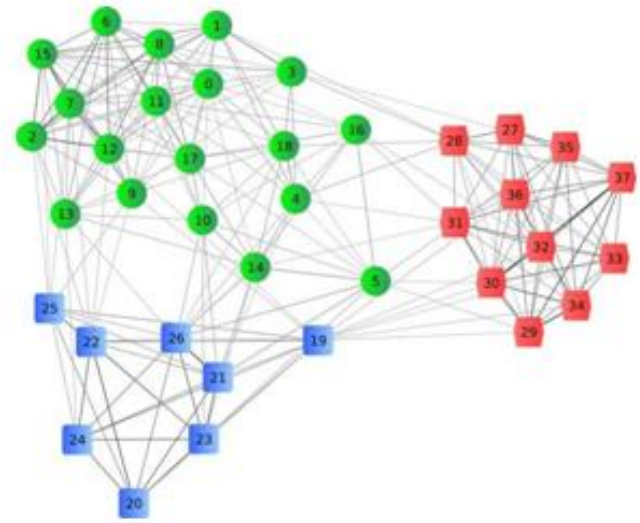

Fig. 1. Data clustering in data analysis.

Fig. 1 show sample cluster formation based on nodes presented in overall structure. Green color items may achieved same group, red and blue color may formed as separate bundle. Needless to bring up, k-means has more than a few basic drawbacks, such as sensitiveness to initialization and to cluster dimension, and its efficiency can be more intense than other state-of-the-art methods in many websites [3]. Regardless of that, its convenience, understandability, and scalability are the reasons for its remarkable reputation. A criteria with adequate efficiency and functionality in most of application scenarios could be much better one with better performance in some situations but restricted utilization due to great complexness. While providing affordable outcomes, k-means is quick and simple to combine with other methods in bigger techniques.

A typical strategy to the clustering problem is to treat it as a marketing procedure. A maximum partition is found by improving a particular operate of likeness (or distance) among information. Generally, there is an implied

 

supposition that the real implied framework of information could be properly described by the likeness system described and embedded in the clustering requirements operate. Hence, effectiveness of clustering methods under this approach depends on the suitability of the likeness evaluate to the information at hand. For example, the unique k-means has sum-of-squared-error purpose operate that uses Euclidean distance [4] [5]. In a very rare and high-dimensional domain like written text records, rounded k-means, which uses cosine similarity (CS) instead of Euclidean range as the evaluate, is considered to be more appropriate.

As outlined in the sharp case may not be quickly generalized for unclear clustering. This is due to the fact that in fuzzy methods an example does not are part of a group completely but has restricted account principles in most clusters [6]. More about clustering methods can be discovered in. Clustering considerable quantities of information requires a long time. Further, new unlabeled information places, which will not fit in storage, are becoming available. To group them, either sub testing is required to fit the information in storage or time will be significantly affected by hard drive accesses making clustering an unpleasant choice for information analysis. Another resource of huge information places is streaming information where you do not shop all the information, but process it and remove it. There are some very huge information places for which a little branded information is available and relax of the data is unlabeled i.e. for example, computer attack detection. Our purpose is to obtain a novel technique for measuring likeness between information things in rare and high-dimensional sector, particularly written text records.

## II. BACKGROUND APPROACH

The likeness between two records di and dj is identified w.r.t. the position between the two factors when looking from the source. To create a new idea of likeness, it is possible to use more than just one referrals factor. We may have a more precise evaluation of how near or remote a couple of factors are, if we look at them from many different opinions. A presumption of group subscriptions has been created before to evaluate [7]. The two things to be calculated must be in the same group, while the factors from where to set up this measurement must be outside of the group.

$$\begin{aligned}&\mathrm{MVS}(d_i,d_j|d_i,d_j\in S_r)\\&=\frac{1}{n-n_r}\sum_{d_h\in S\setminus S_r}(d_i-d_h)^t(d_j-d_h)\\&=\frac{1}{n-n_r}\sum_{d_h}\cos(d_i-d_h,d_j-d_h)\|d_i-d_h\|\|d_j-d_h\|.\end{aligned}$$

Fig. 2. MVS Final Form.

Fig.2 is the final form of MVS which depends on formulation of individual similarities with in the sum. As shown in the Fig. 2, the likeness between two factors $d_i$ and $d_j$ within group $S_r$, considered from a factor $d_h$ outside this group, is similar to the item of the cosine of the position between $d_i$ and $d_j$ looking from $d_h$ and the Euclidean ranges from $d_h$ to these two factors. This meaning is in accordance with the supposition that $d_h$ is not in the same group with $d_i$ and $d_j$. Little sized the ranges $\|d_i - d_h\|$ and $\|d_j - d_h\|$ are, the greater the opportunity that $d_h$ is actually in the same group with $d_i$ and $d_j$, and the likeness depending on $d_h$ should also be minute indicate this prospective.

**procedure** BUILDMVSMATRIX($A$)
  **for** $r \leftarrow 1 : c$ **do**
    $D_{S\setminus S_r} \leftarrow \sum_{d_i \notin S_r} d_i$
    $n_{S\setminus S_r} \leftarrow |S \setminus S_r|$
  **end for**
  **for** $i \leftarrow 1 : n$ **do**
    $r \leftarrow$ class of $d_i$
    **for** $j \leftarrow 1 : n$ **do**
      **if** $d_j \in S_r$ **then**
        $a_{ij} \leftarrow d_i^t d_j - d_i^t \frac{D_{S\setminus S_r}}{n_{S\setminus S_r}} - d_j^t \frac{D_{S\setminus S_r}}{n_{S\setminus S_r}} + 1$
      **else**
        $a_{ij} \leftarrow d_i^t d_j - d_i^t \frac{D_{S\setminus S_r} - d_j}{n_{S\setminus S_r} - 1} - d_j^t \frac{D_{S\setminus S_r} - d_j}{n_{S\setminus S_r} - 1} + 1$
      **end if**
    **end for**
  **end for**
  **return** $A = \{a_{ij}\}_{n\times n}$
**end procedure**

Fig. 3. Process of multi view clustering in real time data sets.

The overall likeness between $d_i$ and $d_j$ is identified by getting regular over all the opinions not that belong to group $S_r$. The procedure for building the MVS matrix is given in Fig. 3. It is possible to claim that while most of these opinions are useful, there may be some of them providing deceiving details just like it may occur with the source factor. However, given a huge enough variety of opinions and their wide range, it is affordable to believe that most of them will be useful.

 

## III. PROPOSED APPROACH

This constraint makes Fuzzy-C Means, or FCM to be exceptionally touchy to clamor. The general rule of the method introduced in this paper is to consolidate the area data into the FCM calculation amid arrangement. Keeping in mind the end goal to join the spatial connection into FCM calculation, the destination capacity of afore mentioned comparison may seems equal and other critical issues will be planned and the proposed calculation focused around regularization term in predefined and different incidents is punished by a regularization term, which is propelled by the above NEM (Neighborhood Expectation Maximization) calculation and adjusted focused around the paradigm of FCM calculation. The new target capacity of the Possibilistic Fuzzy C- Means, or PFCM is characterized as takes after:

$$J_{PFCM} = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik|})^{q} d^{2}(x_k, v_i) + \gamma \sum_{k=1}^{n}\sum_{j=1}^{n}\sum_{i=1}^{c}(u_{ik})^{q}(1-u_{ij})^{q} w_{kj},$$

Fig. 4. Objective Function

The Conventional FCM partitions a set of object data into a number of c clusters based on the minimization of a quadratic objective function, formulated in Fig.4 [8]. The parameter q controls the impact of the punishment term. The relative criticalness of the regularizing term is conversely corresponding to the sign to-clamor (SNR) of the picture. Lower SNR would oblige a higher estimation of the parameter q, and the other way around. At thepoint when q = 0, JPFCM measures up to JFCM. The significant contrast between NEM calculation and PFCM calculation is that the punishment term in the NEM is boosted to get the arrangements while in the PFCM it ought to be minimized so as to fulfill the standard of FCM calculation. Moreover, the punishment term in the PFCM calculation has the weighting example to control the level of fluffiness in the ensuing participation capacity in spite of the punishment term in the NEM calculation that is fresh [9]. This new punishment term is minimized when the enrollment esteem for a specific class is vast and the participation values for the same class at neighboring pixels is additionally huge, and the other way around. As it were, it obliges the pixel's participation estimation of a class to be associated with those of the neighboring pixel.

## IV. EXPERIMENTAL EVALUATION

To confirm the key benefits of our suggested techniques, we evaluate their efficiency in tests on document data. The purpose of this area is to evaluate MVSC-IR and MVSC-IV with the current techniques that also use specific likeness actions and requirements features for document clustering. The likeness actions to be compared include Euclidean range, cosine likeness, and extended Jacquard coefficient.

The information corpora that we used for tests include of 20 standard papers information places. Besides reuters7 and k1b [10] [11], which have been described in information previously, we included another 18 written text selections so that the evaluation of the clustering techniques is more thorough and comprehensive.

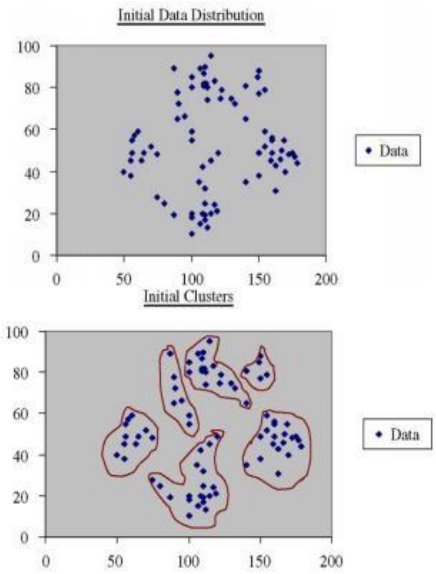


Fig. 5. Cluster results with processing application development.

It has been known that requirements function- based partitional clustering techniques can be delicate to group dimension and balance. In the ingredients of IR, parameter α which is known as the controlling aspect, α ϵ [0,1]. To analyze how the dedication of α could impact MVSC-IR's performance, we analyzed MVSC-IR with different values of α from 0 to 1, with 0.1 step-by-step period [12]. The assessment was done in accordance with the clustering

 

outcomes in NMI (Normalized Mutual Information), FScore, and Precision, each averaged over all the 20 given information places. Since the evaluation analytics for different data places could be very different from each other, simply taking the common over all the information places would not be very meaningful. Fig. 5. Shows the cluster results with processing application development. Hence, we applied the technique used to convert the analytics into comparative analytics before calculating.

1. Initialize $U=[u_{ij}]$ matrix, $U^{(0)}$
2. At k-step: calculate the centers vectors $C^{(k)}=[c_j]$ with $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

3. Update $U^{(k)}$ , $U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left( \frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

4. If $\| U^{(k+1)} - U^{(k)} \| <$ then STOP; otherwise return to step 2.
5. Initialize $U=[u_{ij}]$ matrix, $U^{(0)}$

Fig. 6. Algorithm for fuzzy c means process.

Powerful Indicates clustering strategy is the new methodology to group the information things into variety of groups, which is unidentified originally. Variety of categories (clusters) is some beneficial integer. The collection is done by measuring the range between item and centroid. The procedure for the Fuzzy C-Means algorithm is given in the Fig. 6 [13]. Objects are iteratively arranged into the current categories or a new cluster development is done with those things centered up on the threshold restrict. Thus the objective of dynamic clustering is to classify the information. It could enhance the possibilities of discovering the global optima with cautious choice of preliminary group. In this algorithm information things are saved in additional storage and transferred to primary storage individually. Only the group representatives are saved completely in primary storage to alleviate area restrictions. Therefore, an area need of these criteria is very small, necessary only for the centroids of the categories. This algorithm is non-iterative and therefore it is time need is also little.

## V. CONCLUSION

Clustering decides the connections between information objects in the data source. The things are arranged or arranged based on the key of "maximizing the infraclass similarity and reducing the interclass similarity". It discovers out something useful from data source. Clustering has its roots in many areas, such as information exploration, research, biology, and device learning etc. Clustering methods can be divided into various types: Dividing methods, Hierarchical methods, Solidity centered methods, Grid-based methods; Design centered methods, Probabilistic methods, and Chart theoretic and Unclear methods. The Powerful mean algorithm are the significant concentrate of this dissertation work. Dynamic mean criteria generate good groups automatically because there is no need to described the number of groups before head but in Powerful mean criteria each data factor can be a participant of one and only one group at a time. In other terms we can say that the sum of account grades of each information point in all groups is similar to one and in all the staying groups its account quality is zero. In our thesis dynamic criteria is customized using fuzzy criteria. By implementing fuzzy criteria over Powerful criteria we can show the account of each information factor in all groups. By applying Unclear criteria over Powerful criteria clustering can be at an extremely quicker rate. It is appropriate to a large amount of information saved in databases. The overall results are significant in displaying that Powerful criteria display membership of each information factor in every groups.

## REFERENCES

[1] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg,"Top 10 algorithms in data mining," Knowl. Inf. Syst., vol. 14, no. 1, pp. 1–37, 2007.

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

 

Clustering Theory, 2009.
[3] M. Pelillo, "What Is a Cluster? Perspectives from Game Theory," Proc. NIPS Workshop Clustering Theory, 2009. [4] I. Dhillon and D. Modha, "Concept decompositions for large sparse text data using clustering," Mach. Learn., vol. 42, no. 1-2, pp. 143–175, Jan 2001.
[5] S. Zhong, "Efficient online spherical K-means clustering," in IEEE IJCNN, 2005, pp. 3180–3185. S. Zhong, "Efficient Online Spherical K-means Clustering," Proc. IEEE Int'l Joint Conf. Neural Networks (IJCNN), pp. 3180-3185, 2005.
[6] 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.
[7] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2nd ed. New York: John Wiley & Sons, 2001.
[8] Modeling Decision for Artificial Intelligence: 8th International Conference, pg 153, edited by Vicenc Torra, Yasuo Narukawa, Jianping Yin, Jun Long.
[9] J. C. Bezdek (1981): "Pattern Recognition with Fuzzy Objective Function Algorithms", Plenum Press, New York Tariq Rashid: "Clustering"
[10] E.-H. Han, D. Boley, M. Gini, R. Gross, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, and J. Moore, "Webace: a web agent for document categorization and exploration," in AGENTS '98: Proc. of the 2nd ICAA, 1998, pp. 408–415.
[11] G. Karypis, "CLUTO a clustering toolkit," Dept. of Computer Science, Uni. of Minnesota, Tech. Rep., 2003, http://glaros.dtc. umn.edu/gkhome/views/cluto.
[12] Y. Zhao and G. Karypis, "Empirical and theoretical comparisons of selected criterion functions for document clustering," Mach. Learn., vol. 55, no. 3, pp. 311–331, Jun 2004.
[13] Bezdek, James C. (1981). Pattern Recognition with Fuzzy Objective Function Algorithms. ISBN 0-306-40671-3.
[14] Ahmed, Mohamed N.; Yamany, Sameh M.; Mohamed, Nevin; Farag, Aly A.; Moriarty, Thomas (2002). "A Modified Fuzzy C-Means Algorithm for Bias Field Estimation and Segmentation of MRI Data".

## ABOUT AUTHOR

**Brahmani Parvataneni** received B.Tech degree in Computer Science Engineering at DVR & Dr HS MIC College of Technology Affiliated to Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh, India, Where she is also pursuing the M.Tech Degree in the division of Computer Science Engineering at NRI Institute of Technology, Agiripalli Affiliated to Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh, India. Her Research Interests include information retrieval, data mining.

**Dr. K.V.Sambasiva Rao,** is working as a Dean, Department of CSE, NRI Institute of Technology, Pothavarappadu, Agiripalli, Andhra Pradesh. He has a total of 25 years of teaching experience and 5 years of Research Experience. He has produced 4 PhD's and guided 22 M.Tech students. He has published a total of 53 papers in National as well as International Journals. His Research Interests include Artificial Intelligence, Data Mining, Cloud Computing, and Big data.

 