ISSN No. 0976-5697

Volume 1, No. 4, Nov-Dec 2010

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Fuzzy Clustering in Web Content Mining

Mr.Ranjit R.Keole*
M E (CSE), Department of Computer Science & Engg.
Prof. Ram Meghe Institute of Tech and Research,Badnera,
Amaravati,India
ranjitkeole@gmail.com

Dr. G.R.Bamnote
Prof. & Head of Department of Computer Science & Engg.
Prof. Ram Meghe Institute of Tech. and Research,Badnera,
Amaravati,India
grbamnote@rediffmail.com

Abstract: Clustering is useful technique in the field of textual data mining. Cluster analysis divides objects into meaningful groups based on similarity between objects. Copious material is available from the World Wide Web (WWW) in response to any user-provided query. It becomes tedious for the user to manually extract real required information from this material. Large document collections, such as those delivered by Internet search engines, are difficult and time-consuming for users to read and analyze. The detection of common and distinctive topics within a document set, together with the generation of multi-document summaries, can greatly ease the burden of information management. This paper focus on this problem of mining the useful information from the collected web documents using fuzzy clustering of the text collected from the downloaded web documents.

Keyword: hard c –means, fuzzy c –means, possibilistic c-means, hyper-spherical fuzzy c-means, text mining

I. INTRODUCTION

Clustering is an unsupervised classification of objects (data instances) into different groups. In particular we are talking about the partitioning of a dataset into subsets (clusters), so that the data in each subset (ideally) share some common property. This property is usually defined as proximity according to some predefined distance measure. The goal is to divide the dataset in such a way that objects belonging to the same cluster are as similar as possible, whereas objects belonging to different clusters are as dissimilar as possible. The computational task of classifying the data set into k clusters is often referred to as k-clustering. Although estimating the actual number of clusters (k) is an important issue we leave it untouched in this work. Fuzzy clustering [2, 3] in contrast to the usual (crisp) methods does not provide hard clusters, but returns a degree of membership of each object to all the clusters. The interpretation of these degrees is then left to the user that can apply some kind of a thresholding to generate hard clusters or use these soft degrees directly. With more than two billion pages created by millions of Web page authors and organizations, the World Wide Web is a tremendously rich knowledge base. The knowledge comes not only from the content of the pages themselves, but also from the unique characteristics of the Web, such as its hyperlink structure and its diversity of content and languages. A considerably large portion of information present on the World Wide Web (WWW) today is in the form of unstructured or semi-structured text data bases. The WWW instantaneously delivers huge number of these documents in response to a user query. However, due to lack of structure, the users are at a loss to manage the information contained in these documents efficiently. The WWW continues to grow at an amazing rate as an information gateway and as a medium for conducting business. Web mining is the extraction of interesting and useful knowledge and implicit information from artifacts or activity related to the WWW.

In this context, the importance of data/text mining and knowledge discovery is increasing in different areas like: telecommunication, credit card services, sales and marketing etc [1]. Text mining is used to gather meaningful information from text and includes tasks like Text Categorization, Text Clustering, Text Analysis and Document Summarization. Text Mining examines unstructured textual information in an attempt to discover structure and implicit meanings within the text.

One main problem in this area of research is regarding organization of document data. This can be achieved by developing nomenclature or topics to identify different documents. However, assigning topics to documents in a large collection manually can prove to be an arduous task. Documents into the related topics. Clustering is the proven technique for document grouping and categorization based on the similarity between these documents [4]. Documents within one cluster have high similarity with each another, but low similarity with documents in other clusters.

II. FUZZY CLUSTERING ALGORITHMS

In this section, we present some of the fuzzy clustering Algorithms mainly based on the descriptions in [5]. We devote the majority of space to the hard c-means, fuzzy c-means, H-FCM and possibilistic c-means.

All algorithms described here are based on objective functions, which are mathematical criteria that quantify the quality of cluster models. The goal of each clustering algorithm is the minimization of its objective function. The following syntax will be used in the equations, algorithms and their explanations: \

J … .Objective function

$X=\{X_{1\ldots}\ Xn_{\}}$ … …dataset of all objects (data instances)

$C = \{C_1 \ldots C_C\}$……set of cluster prototypes (centroid vectors)

$d_{ij} = \|X_1\text{-}C_1\|$…..distance between object X_1 and centre C_1

μ_{ij} ……….weight of assignment of object xj $_{\text{to}}$ cluster i.

$\mu_{j=(}\mu_{ij}, \ldots \mu_{cj})^{T\ldots\ldots\ldots}$membership vector of object x_j

$U=(\mu_{ij)=}(\mu_1 \ldots \mu_n)$…partition matrix of size c x n

A. *Hard C –Means (HCM)*

Hard c-means is better known as k-means and in general this is not a fuzzy algorithm. However, its overall structure is the basis for all the others methods. Therefore we call it hard c-means in order to emphasize that it serves as a starting point for the fuzzy extensions.

The objective function of HCM can be written as follows:

$$J_h = \sum_{i=1}^{c}\sum_{j=1}^{n} \mu_{ij}\, d_{ij}^2 \tag{2.1}$$

As mentioned HCM is a crisp algorithm, therefore:

$\mu_{ij} \in \{0, 1\}$ also required that each object belongs to exactly one cluster $\sum_{i=1}^{c} \mu_{ij} = 1,\ \forall j \in \{1, \dots, n\}$.

Before outlining the algorithm, we must know how to Calculate new membership weights

$$\mu_{ij} = \begin{cases} 1, \text{if } \operatorname{argmin}_{l=1}^{c} \text{dij} \\ 0, \text{otherwise} \end{cases} \tag{2.2}$$

And based on the weights, how to derive new cluster centres.

$$C_i = \frac{\sum_{j=1}^{n} \mu_{ij} x_j}{\sum_{j=1}^{n} \mu_{ij}} \tag{2.3}$$

The algorithm can now be stated very simply as shown below.

INPUT:

A set of learning objects to be clustered and the number of desired clusters *c*

OUTPUT:

Partition of learning examples into *c* clusters and membership values μ_{ij} for each example X_i and cluster i.

ALGORITHM *(2.1) The hard c-means algorithm:*

```
(randomly) generate clusters centres
repeat
    for each object recalculate membership weights
    using equation (2.2)
    recomputed the new centres using equation (2.3)
until
    no change in C can be observed.
```

The HCM algorithm has a tendency to get stuck in a local minimum, which makes it necessary to conduct several runs of the algorithm with different initializations. Then the best result out of many clusterings can be chosen based on the objective function value.

B. *Fuzzy C –Means (FCM)*

In hard clustering, data is divided into distinct clusters, where each data element belongs to exactly one cluster. In fuzzy clustering, data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the strength of the association between that data element and a particular cluster. Topics that characterize a given knowledge domain are somehow associated with each other. Those topics may also be related to topics of other domains. Hence, documents may contain information that is relevant to different domains to some degree. With fuzzy clustering methods documents are attributed to several clusters simultaneously and thus, useful relationships between domains may be uncovered, which would otherwise be neglected by hard clustering methods.

Probabilistic fuzzy cluster analysis [2,3] relaxes the requirement $\mu_{ij} \in \{0,1\}$, which now becomes $\mu_{ij} \in [0,1]$.

However $\sum_{i=1}^{c} \mu_{ij} = 1,\ \forall j \in \{1, \dots, n\}$ still holds. FCM optimizes the following objective fuction:

$$J_f = \sum_{i=1}^{c}\sum_{j=1}^{n} {\mu_{ij}}^m\, d_{ij}^2 \tag{2.4}$$

Parameter *m*, *m*>1, is called the fuzzyfier or the weighting exponent. The actual value of *m* determines the 'fuzziness' of the classification. It has been shown [4] that for the case *m*=1, J_f becomes identical to J_h and thus FCM becomes identical to hard c-means.

The transformation from the hard c-means to the FCM is very straightforward; we must just change the equation for calculating memberships (2.2) with:

$$\mu_{ij} = \frac{1}{\sum_{l=1}^{c} \left(\frac{d_{ij}^2}{d_{ij}^2}\right)^{\frac{1}{m-1}}} = \frac{d_{ij}^{\frac{-2}{m-1}}}{\sum_{l=1}^{c} d_{ij}^{\frac{-2}{m-1}}} \tag{2.5}$$

And function for recomputing clusters (2.3) with

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}^m X_j}{\sum_{j=1}^{n} u_{ij}^m} \tag{2.6}$$

Equation (2.5) clearly shows the relative character of the probabilistic membership degree. It depends not only on the distance of the object Xj to the cluster C_i, but also on the distances between this object and other clusters. Although the algorithm stays the same as in HCM, we get probabilistic outputs if we apply above changes. The (probabilistic) fuzzy c-means algorithm is known as a stable and robust classification method. Compared with the hard c-means it is quite insensitive to its initialization and it is not likely to get stuck in an undesired local minimum of its objective function in practice. Due to its simplicity and low computational demands, the probabilistic FCM is a widely used initializer for other more sophisticated clustering methods.

C. *Possibilistic C-Means (PCM)*

Although often desirable, the relative property of the probabilistic membership degrees can be misleading. High values for the membership of object in more than one cluster can lead to the impression that the object is typical for the clusters, but this is not always the case. Consider, for example, the simple case of two clusters shown in figure 2.1. Object X_1 has the same distance to both clusters and thus it is assigned a membership degree of about 0.5. This is plausible. However, the same degrees of membership are assigned to object X2 even though this object is further away from both clusters and should be considered less typical. Because of the normalization the sum of the memberships has to be 1. Consequently X_2 receives fairly high membership degrees to both clusters. For a correct interpretation of these memberships one has to keep in mind that they are rather degrees of sharing than of typicality, since the constant weight of 1, given to an object, must be distributed over the clusters.

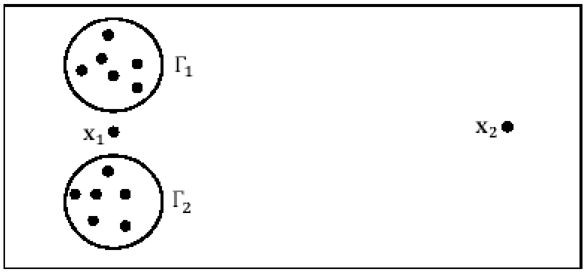

Figure 2.1: Example of misleading interpretation of the FCM membership degree.

Therefore PCM, besides relaxing the condition for μ_{ij} to $\mu_{ij} \in \{0, 1\}$ as in case of FCM, also drops the normalization requirement: $\sum_{i=1}^{c} \mu_{ij} = 1, \forall j \in \{1, \ldots, n\}$. The probabilistic objective function J_f that just minimizes squared distances would be inappropriate because with dropping of the normalization constraint a trivial solution exists for $\mu_{ij} = 0$, for all i∈ { 1,…c}, and j ∈ { 1,…n} , , i.e., all clusters are empty. In order to avoid this solution, penalty a term is introduced that forces the memberships away from zero.

Objective function J_f is modified to:

$$J_p = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m} d_{ij}^{2} + \sum_{j=1}^{c} \eta_i \sum_{j=1}^{n} \left(1 - u_{ij}\right)^{m} \tag{2.7}$$

Where $\eta_i > 0$ for all i ∈ {1,…., c}.

In the PCM algorithm, the equation for calculating cluster centres stays the same as in FCM (2.6). But the equation for Recalculating membership degree changes from (2.5) to:

$$u_{ij} = \frac{1}{\left(\frac{d_{ij}^2}{\eta_i}\right)^{\frac{1}{m-1}}} \tag{2.8}$$

This also slightly changes the original procedure since we must recompute η_I using the equation (2.9) before calculating the weight μ_{ij}.

$$\eta_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} d_{ij}^{2}}{\sum_{j=1}^{n} u_{ij}^{m}} \tag{2.9}$$

Properties of PCM [5] are the following:

Cluster Coincidence: since PCM is not forced to partition data exhaustively it can lead to solutions where two or more clusters occupy the same space (same objects with the same membership weighting).

Cluster Repulsion: objective function J_p is, in general, fully optimized only if all clustered centres are identical. Because of that, other, not optimal solutions are found just as a side effect of J_p getting stuck in a local optimum.

D. Other Reviewed Algorithm

During the review of fuzzy clustering algorithms we considered also the following algorithms. We will not precisely describe them in this paper, since we decided that they are not the best choice .An interesting reader can find their descriptions in [6] or [5].

(a) Gustafson-Kessel Algorithm: while FCM and PCM can only detect spherical clusters GKA can identify also clusters of different forms and sizes. It is more sensitive to initialization and has higher computational costs.

(b) Fuzzy Shell Clustering: can, in contrast to all the algorithms above, identify also non-convex shaped clusters. They are especially useful in the area of image recognition. We think that this property in not needed in text clustering.

(c) Kernel-based Fuzzy Clustering: are variants of fuzzy clustering algorithms that modify the distance function to handle non-vectorial data, such as sequences, trees or graphs, without the requirement to completely modify the algorithm itself. In text clustering we are dealing with vectors so there is no need for such an advanced method.

E. Hyper-Spherical Fuzzy C-Means (H-FCM)

Recently the Fuzzy c-Means (FCM) algorithm is modified for clustering text documents based on the cosine similarity coefficient rather than on the Euclidean distance. The modified algorithm works with normalized *k*-dimensional data vectors that lie in hyper-sphere of unit radius and hence has been named Hyper-spherical Fuzzy c-Means (H-FCM). The H-FCM algorithm for document clustering has shown that it outperforms the original FCM algorithm as well as the hard k-Means algorithm.

The objective function the H-FCM minimizes is similar to the FCM one, the difference being the replacement of the squared norm by a dissimilarity function $D_{i\alpha}$:

$$J_m\ (U,V) = \sum_{i=1}^{N} \sum_{\alpha=1}^{c} {u_{\alpha i}}^{m} D_{i\alpha} = \sum_{i=1}^{N} \sum_{\alpha=1}^{c} {u_{\alpha i}}^{m} \ (1 - \sum_{j=1}^{k} x_{ij}.v_{\alpha j}\ ) \tag{2.10}$$

The cosine coefficient ranges in the unit interval and when data vectors are normalized to unit length it is equivalent to the inner product. The dissimilarity function $D_{i\alpha}$ in equation (1) consists of a simple transformation of the cosine similarity coefficient, *i.e.* $D_{i\alpha} = 1 - S_{i\alpha}$.

$$u_{\alpha j} = \left[\sum_{\beta=1}^{c} \left(\frac{D_{i\alpha}}{D_{i\beta}}\right)^{\frac{1}{(m-1)}}\right]^{-1} = \left[\sum_{\beta=1}^{c} \left(\frac{1 - \sum_{j=1}^{k} x_{ij}.v_{\propto j}}{1 - \sum_{j=1}^{k} x_{ij}.v_{\beta j}}\right)^{\frac{1}{(m-1)}}\right]^{-1}$$

$$v_\alpha = \sum_{i=1}^{N} u_{\propto i}{}^{m} X_i \,.\left[\sum_{j=1}^{K}\left(\sum_{i=1}^{N} u_{\propto i}{}^{m} X_{ij}\right)^2\right]^{-\frac{1}{2}} \tag{2.11}$$

The update expression for the membership of data element x_i in cluster α, denoted as $u_{\alpha i}$ and shown in equation (2.11), is also similar to the original FCM expression since the calculation of $D_{i\alpha}$ does not depend explicitly on $u_{\alpha i}$. However, a new update expression for the cluster centroid v_α, shown in equation (2.12), had to be developed. Like the original algorithm, H-FCM runs iteratively until a local minimum of the objective function is found or the maximum number of iterations is reached**.**

F. Finding the Optimum Number of Clusters

The H-FCM algorithm requires the selection of the number of clusters *c*. However, in most clustering applications the optimum *c* is not known *a priori*. A typical approach to find the best *c* is to run the clustering algorithm for a range of *c* values and then apply validity measures to determine which *c* leads to the best partition of the data set. The validity of individual clusters is usually evaluated based on their compactness and density. In low-dimensional spaces it is acceptable to assume that valid clusters are compact, dense and well separated from each other. However, text documents are typically represented as high-dimensional sparse vectors. In such problem space, the similarity between documents and cluster centroids is generally low and hence, compact clusters are not expected. Therefore, the approach mentioned above for finding the optimum *c* is inappropriate. A question that arises is how the H-FCM algorithm is able to discover meaningful document clusters considering such low similarity patterns. As observed for the hard k-Means algorithm, the good performance of the H-FCM is justified by the fact that documents within a given cluster are always more similar to the corresponding centroid than documents outside that cluster, regardless of the number of clusters that has been selected. It is believe that in the high-dimensional document space the issue of finding the optimum number of clusters is not so relevant. The choice of *c* should rather address the desired granularity level, since the higher the number of clusters the more specific will be the topics covered by the documents in those clusters.

III. CONCLUSION AND FUTURE WORK

This paper presents an overview of fuzzy clustering algorithms that could be potentially suitable for document Clustering, we have surveyed HARD C –MEANS (HCM), Fuzzy C –MEANS (FCM), Possibilistic c-means (PCM), and HYPER-SPHERICAL FUZZY C-MEANS (H-FCM) cluster ing algorithms. The HCM algorithm has a tendency to get stuck in a local minimum, which makes it necessary to conduct several runs of the algorithm with different initializations. In hard clustering, data is divided into distinct clusters, where each data element belongs to exactly one cluster. In fuzzy clustering, data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the strength of the association between that data element and a particular cluster. PCM is not forced to partition data exhaustively it can lead to solutions where two or more clusters occupy the same space ie.same objects with the same membership weighting. The H-FCM generates clusters with a higher level of granularity and that the resulting clusters hierarchy successfully links clusters of the same topic.

There are many areas in text mining; where one may carry the work to enhance various areas. Out of these, the labeling of the clusters is a very daunting challenge of this time. No remarkable effort has been made in this regard to get good result. That is why automatic labeling of the clusters is not so much accurate. A keen and concerted work has been done to remove this hurdle. It will certainly serve as a lime length for future researchers.

IV. REFERENCES

[1] Hsinchun Chen and Michael Chau, "Web Mining: Machine learning for Web Applications", Annual Review of Information Science and Technology 2003.

[2] Dunn, J., C., A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact Well-Separated Clusters, Journal of Cybernetics 3, pp. 32-57,1973.

[3] Bezdek, J., C., Pattern Recognition with Fuzzy Objective Function Algoritms, Plenum Press, New York, 1988.

[4] Mr. Rizwan Ahmad and Dr. Aasia Khanum, "Document Topic Generation in Text Mining by Using Cluster Analysis with EROCK", International Journal of Computer Science & Security (IJCSS), Volume (4) : Issue (2) Aug 2008.

[5] Valente de Oliveira, J., Pedrycz, W., Advances in Fuzzy Clustering and its Applications, John Wiley & Sons, pp 3-30, 2007.

[6] Höppner, F., Klawonn, F., Krise, R., Runkler, T., Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition, John Wiley & Sons, pp 5-114, 2000.

[7] M.E.S. Mendes Rodrigues and L. Sacks, "A Scalable Hierarchical Fuzzy Clustering Algorithm for Text Mining", Department of Electronic and Electrical Engineering University College London Torrington Place, London, WC1E 7JE, United Kingdom, 2004.

[8] D.R. Cutting, D.R. Karger, J.O. Pederson and J.W. Tukey (1992). Scatter/gather: a cluster-based approach to browsing large document collections. In: Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'92, pp. 318-329, Copenhagen, Denmark, June 1992.

[9] A. Schenker, M. Last and A. Kandel (2001). A term-based algorithm for hierarchical clustering of Web documents. In: Proceedings of the Joint 9th IFSA World Congress and 20th NAFIPS International Conference, vol.5, pp. 3076-3081, Vancouver, Canada, July 2001.

[10] M.E.S. Mendes, W. Jarrett, O. Prnjat and L. Sacks (2003). Flexible searching and browsing for telecoms learning material. In: Proceedings of the 2003 International Symposium on Telecommunications, IST'2003, Isfahan, Iran, August 2003.

[11] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, Angela Y. Wu, "An Efficient k-Means Clustering Algorithm: Analysis and Implementation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 7, July 2002.

[12] Linas Baltruns, Juozas Gordevicius, "Implementation of CURE Clustering Algorithm", SIGMOD Seattle, WA, USA ACM February 1, 2005.

[13] Tian Zhang, Raghu Ramakrishan, Miron Livny, "BIRCH: An Efficent Data Clustering Method for Very

Large Databases” SIGMOD ’96 6/96 Montreal, Canada IQ 1996 ACM.

[14] Shaoxu Song and Chunping Li, “Improved ROCK for Text Clustering Using Asymmetric Proximity”, SOFSEM 2006, LNCS 3831, pp. 501–510, 2006.

[15] Linas Baltruns, Juozas Gordevicius, “Implementation of CURE Clustering Algorithm”, February 1, 2005.

[16] Raymond Y.K. Lau, Senior Member, IEEE, Dawei Song, Yuefeng Li, Member, IEEE, Terence C.H. Cheung, Member, IEEE, and Jin-Xing Hao, “Toward a Fuzzy Domain Ontology Extraction Method for Adaptive e-Learning,” IEEE Trans. On Knowledge and Data Engineering, Vol. 21, No. 6, Jun 2009.

[17] Shady Shehata, Member, IEEE, Fakhri Karray, Senior Member, IEEE, and Mohamed S. Kamel, Fellow, IEEE, “An Efficient Concept-Based Mining Model for Enhancing Text Clustering”, IEEE Trans. On Knowledge and Data Engineering, Vol. 22, No. 10, Oct 2010.

[18] Jingwen Tian, Meijuan Gao, and Yang Sun, “Study on Web Classification Mining Method Based on Fuzzy Neural Network”, Proceedings of the IEEE International Conference on Automation and Logistics Shenyang, China August 2009.