Unsupervised Approach for Document Clustering using Modified Fuzzy C mean Algorithm

K.Sathiyakumari[#1], V.Preamsudha [*2], G.Manimekalai[#3]
[#1&2]*Assistant Professors, PSGR Krishnammal College for Women, Coimbatore, India.*
[3]*M.Phil Scholar, PSGR Krishnammal College for Women, Coimbatore,India.*

***Abstract*—** *Clustering is one the main area in data mining literature. There are various algorithms for clustering. There are several clustering approaches available in the literature to cluster the document. But most of the existing clustering techniques suffer from a wide range of limitations. The existing clustering approaches face the issues like practical applicability, very less accuracy, more classification time etc. In recent times, inclusion of fuzzy logic in clustering results in better clustering results. One of the widely used fuzzy logic based clustering is Fuzzy C-Means (FCM) Clustering. In order to further improve the performance of clustering, this thesis uses Modified Fuzzy C-Means (MFCM) Clustering. Before clustering, the documents are ranked using Term Frequency–Inverse Document Frequency (TF–IDF) technique. From the experimental results, it can be observed that the proposed technique results in better clustering results when compared to the existing technique*

***Keywords*—** *Data mining, MFCM algorithm, Purity, Entropy, TF-IDF.*

I. INTRODUCTION

J. Han, M. Kamber [1] clustering deals with finding a structure in a collection of unlabeled data. A cluster is therefore a collection of objects, which are ‘‘similar’’ between them and are ‘‘dissimilar’’ to the objects belonging to other clusters. At present, clustering algorithms can be categorized into several types, such as partitional method, hierarchical method, density-based method, grid-based method and model-based method. Clustering is the organizing data into sensible grouping is an essential factor for understanding and learning. For instance, a common scheme of scientific classification puts organisms into a system of ranked taxa: domain, kingdom, Phylum, class, etc. Cluster analysis is defined as the study of methods for grouping, or clustering, objects depending on the measured or perceived intrinsic characteristics or similarity. A category label is not used in cluster analysis.

Thus clustering of document is an automatic grouping of text documents into clusters such that documents within a cluster have high resemblance in comparison to one another, but are different to documents in other clusters. Hierarchical document clustering categorizes clusters into a tree or a hierarchy that benefits browsing. The parent-child relationship among the nodes in the tree can be considered as a topic-subtopic relationship in a subject hierarchy such as the Yahoo! directory.

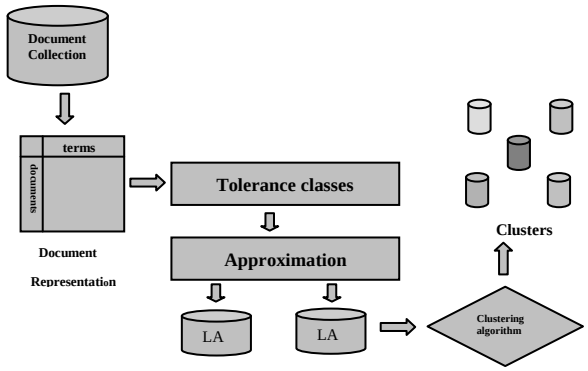

Fig 1: Document Clustering Process

In this paper, we mainly focus on the confusion matrix using this method. Presently, most clustering algorithms treat all data samples equally in the clustering process, such as hard C-Means (HCM) and its fuzzy extension, i.e. modified fuzzy C-Means (MFCM) [2]. However, different samples may play different roles in the clustering process, because the samples distribute no uniformly and asymmetrically. Moreover, a sample may contribute to the clustering results differently in different processes. Hence, it is very useful to give an appropriate sample weight in cluster analysis. For that purpose, sample weighting clustering algorithm have been proposed in literature [3–4].

The modified fuzzy c means clustering is one of the most popular clustering algorithms. It uses to constraints the membership function. 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.

II. FUZZY C MEAN ALGORITHM

Conventional clustering techniques create partitions in which each pattern belongs to one and only one cluster. Therefore, the clusters in a hard clustering are

disjoint. Fuzzy clustering approach associates each pattern with every cluster using a membership function. The output of such techniques is a clustering, but not a partition.

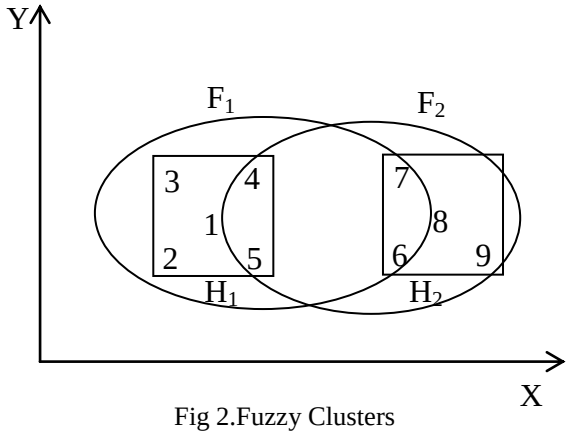

Fig 2.Fuzzy Clusters

In each cluster the ordered pairs (i, μ_i) denotes the i^{th} pattern and its membership value to the cluster μ_i. In the assignment of the pattern to the cluster the confidence will be higher as the membership values are larger. Solid clustering can be attained from a fuzzy partition through the threshold of the membership value.

In the year 1973 Dunn developed the Fuzzy C Means algorithm and later in 1981 it was enhanced by Bezdek.Fuzzy C Means algorithm is extensively used in pattern recognition. Fuzzy C Means algorithm uses the iteratively process, which rejuvenates cluster centers for individual data point. Fuzzy C Means algorithm repetitively iterates the cluster enters to the exact location with in data set elements. The performance of Fuzzy C Mean algorithm is based on the initial centroids selected. The mean of all data points in the Fuzzy C Means algorithm is calculated as the centroid of a cluster and is weighted by their degree corresponding to the cluster.

Fuzzy *c*-means (FCM) approach is one of the most widely used fuzzy clustering algorithms. In the case of eliminating the local minima, this algorithm is more significant than the hard *k*-means algorithm but still FCM can converge to local minima of the squared error criterion. The notable difficulty in the fuzzy clustering technique is the design of membership function. It has many alternate approaches which is comprised of those depends upon similarity decomposition and the centroids of clusters.

The C-Means algorithm [5] is a fuzzy clustering technique that works something like the above method but provides additional flexibility regarding membership. An individual will belong to *one or more* classes or clusters with different membership degrees. This idea arises from the fact that it is ambiguous to tell whether a point must go into a certain cluster and not into another (consider points with equal membership for two clusters, for instance). To deal with this ambiguity, it is necessary to introduce some fuzziness into the formulation of the problem. Instead of having precise, crisp boundaries for a cluster representing a binary threshold which indicates whether a point definitely belongs to a cluster or not, fuzzy membership functions compute a membership degree of each point for every cluster. C-Means will define clusters from a set of input points using this loose membership strategy, which constitutes the most famous algorithm that has been developed for this purpose.

III MODIFIED FUZZY C MEAN CLUSTERING (MFCM)

One of the important characteristics of an image is that neighbouring data have similar feature values, and the probability that they belong to the same cluster is great. The spatial information is important in clustering, but the standard FCM algorithm does not fully utilized it, to exploit the spatial information, a modified membership function is defined as follow:

$$u_{ij} = \frac{u_{ij}^m S_{ij}^n}{\sum_{k=1}^{c} u_{kj}^m S_{kj}^n}$$

Where $S_{ij} = \sum_{k \in N(x_j)} u_{ik}$ is called spatial function, and $N(x_j)$ represents a square window centered on particulate document in the spatial domain. The spatial function S_{ij} represents the probability that the document x_j belongs to i^{th} cluster. The spatial function of a document for a cluster is large if the majority of its neighbourhood belongs to the same cluster. In a homogenous region, the spatial functions enhance the original membership, and the clustering result remains unchanged. However, for misclassified documents, it will reduce the weighting of a noisy cluster by the labels of its neighbors. As a result, misclassified documents can be easily corrected.

There are two steps at each clustering iteration. The first step is to calculate the membership function in the spectral domain and the second step is to map the membership information of each pixel to the spatial domain and then compute the spatial function from that.

The iteration proceeds with the new membership that is incorporated with the spatial function. The iteration is stopped when the maximum difference between two cluster centroids at two successive

iterations is less than a threshold. After the convergence, defuzzification is applied to assign each document to a specific cluster for which the membership is maximal. The Modified FCM algorithm (MFCM) can be described as follows:

Step 1: Set the number of clusters c and the parameter m. Initialize the fuzzy cluster centroid vector $V = [v_1, v_2, ..., v_c]$ randomly and set ε = 0.01.

Step 2: Compute u_{ij} by

$$u_{ij} = \left(\sum_{k=1}^{c} \left(\frac{d(x_j, v_i)}{d(x_j, v_k)} \right)^{2/(m-1)} \right)^{-1}$$

Step 3: Compute v_i by

$$v_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m}$$

Step 4: Update u_{ij} by

$$u_{ij} = \frac{u_{ij}^m S_{ij}^n}{\sum_{k=1}^{c} u_{kj}^m S_{kj}^n}$$

Step 5: Update v_i by

$$v_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m}$$

Repeat Steps 4 and 5 until the following termination criterion is satisfied:

$$|v_{new} - v_{old}| < \varepsilon$$

Finally, the documents are clustered using Modified Fuzzy C-Means (MFCM) clustering algorithm and the ranking is performed using Term Frequency–Inverse Document Frequency (TF–IDF).

IV. EXPERIMENTAL RESULT

A. *Document Pre-Processing*

Not all the words in the documents are important, so they may degrade the classifier's performance. In addition, representing small set of documents that may have hundreds of different words using *bag-of words* approach will generate a huge feature space and thus will increase the processing time. To solve these problems, approaches to reduce the feature space dimension are needed. We used three approaches bellow as the same sequence:

1) As a result of consulting an expert in the domain field, we removed unhelpful sentences from the documents such as "Informed consent was obtained with the benefits, risks and alternatives for the procedure explained", which is found in all reports;

2) We have removed stop words from all data sets using stop-lists containing common words such as "the","a", "an"; the stop words used are corpus-based.

3) We stemmed the words using Porter's suffix stripping algorithm [6]. Words are considered the same if they share the same stem.

V. DOCUMENT REPRESENTATION

Each text document was automatically indexed for term frequency extraction. Stop words (*i.e.* insignificant words like 'a', 'and', 'where', 'or') were eliminated and stemming (*i.e.* removing word affixes such as 'ing', 'ion','s') was performed using Porter's stemming algorithm. Documents were represented as TF (Term Frequency) vectors according to the Vector Space model of IR and a pre-processing filter was applied to discard terms that appeared in a small percentage of documents, leading to significant dimensionality reduction without loss of clustering performance.

VI. PERFORMANCE MEASURES

There are many ways to measure how clustering algorithms perform. One is the confusion matrix. Entry (o, i) of the confusion matrix is the number of data points assigned to output class o and generated from input class i. The input map I is the map of the data points to the input classes. So, the information of the input map can be measured by the entropy H (I). The goal of clustering is to find an output map O that recovers the information. Thus, the conditional entropy H (I|O) is interpreted as the information of the input map given the output map O, i.e., the proportion of information not recovered by the clustering algorithm.

VII. EVALUATION MEASURES USING ENTROPY

Entropy is the degree to which each cluster consists of objects of a single class. For each cluster, the class distribution of the data is calculated initially, i.e., for cluster j we compute p_{ij}, the probability that a member of cluster i belongs to class j as $p_{ij} = m_{ij} / m_i$ where m_i is the number of objects in cluster i and m_{ij} is the number of objects of class j in cluster i. Using this class distribution, the entropy of each cluster i is calculated using the standard formula, $e_i = -\sum_{i=1}^{L} p_{ij} \log_2 p_{ij,}$ where L is the number of

classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each clusters weighted by the size of each cluster, i.e., $e = \sum_{i=1}^{k} \frac{m_i}{m} e_i$, where K is the number of clusters and m is the total number of data points.

Algorithm 1 clustering procedure

Input: (data points: X, # of classes: k)
Output: cluster assignment;
Begin

1. **Initialization:**
   Put all data points into one cluster
   Compute initial criterion H_o
2. **Iteration:**
   Repeat until no more changes in cluster assignment
   Randomly pick a point *x* from a cluster A
   Randomly pick another cluster B
   Put *x* into B
   Compute the new entropy *H*
   If $H > H_o$
   Put x back into A
   H = H_o
   end
   H_o= H
   Goto Step 2.1
   end
3. **Return** the cluster assignment
   End

Fig 3: Clustering Algorithm

VIII. EVALUATION MEASURES USING PURITY

Purity is another method of the extent to which a cluster objects of a single class. Purity of cluster i is $p_i = \max_j p_{ij}$, the overall purity of as clustering is $purity = \sum_{i=1}^{k} \frac{m_i}{m} p_i$

Twenty Newsgroup Dataset

The three News groups data set is a subset of 20 newsgroups, which is a popular data set for experiments in text applications of machine learning techniques, such as text classification and text clustering. Each newsgroup represents one class in the hierarchy structure. Each article is designated to one or more.

The data is organized into three different newsgroups, each corresponding to a different topic. Some of the newsgroups are very closely related to each other (e.g. comp.sys.ibm.pc.hardware/ comp.sys.mac.hardware), while others are highly unrelated (e.g misc.forsale / soc.religion.christian).semantic categories and the total number of categories is 20. We choose 6 categories including 250 documents for the first experiment and 1000 documents for the second. All the documents selected randomly from the 20-newsgroup data sets.

The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. The 20 newsgroups collection has become a popular data set for experiments in text applications of machine learning techniques, such as text classification and text clustering.

From the dataset 1000 documents which belongs to talk politics, alt.athesim, comp.windows x and rec.motorcycles group are randomly chosen for the experiment here is a list of the 20 newsgroups, partitioned (more or less) according to subject matter:

TABLE 1: PURITY OF CLUSTERING

Clustering Method	talk. politics.guns	alt. atheism	comp. windows.x
Modified FCM	0.975	0.930	0.992

The measures used to evaluate the proposed techniques are described below:

Table 1 and figure 1 shows comparison of the purity of classification for the proposed method with the existing methods. From the table 1, it can be observed that for comp. Graphics category, the purity of classification using MFCM algorithm is 0.899 and for the proposed method the purity is higher i.e. 0.975. When the talk politics category is considered, the higher purity i.e. 0.930 is achieved by the proposed technique. When sci.electronics is considered, the better purity is achieved using the proposed technique i.e. 0.992 and 0.987 respectively.

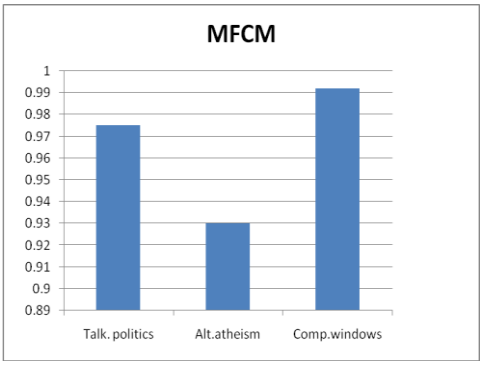

Figure 1: Purity for the Result

The resulted entropy is provided in table 2 and figure 2. It can be seen that the resulted entropy is minimum for using MFCM, whereas it is higher for the other existing method. This clearly shows the

improvement of the clustering when compared to the existing clustering techniques.

TABLE 2: ENTROPY FOR DIFFERENT CLUSTERING METHODS

Clustering Method	talk.politics. guns	alt.atheism	comp. windo ws.x
Modified FCM	0.113	0.252	0.041

Table 3 and Figure 3 shows the classification time resulted for the proposed and exiting technique. It can observed that the time required for classification using the proposed technique for talk.politics.guns, alt.atheism and comp.windows.x are 0.58, 0.51 and 0.41 seconds respectively, whereas, more time is needed by existing techniques for classification i.e., 0.71, 0.79 and 0.59 respectively .

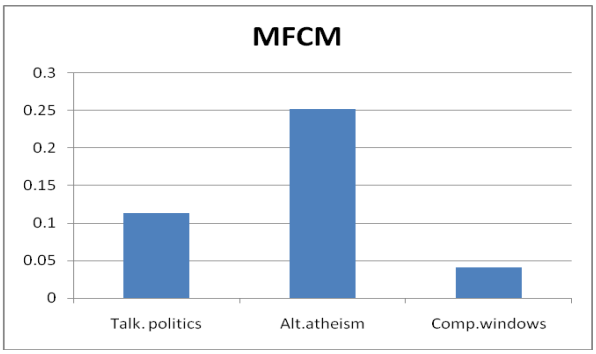

Figure 2: Entropy for the Result

Table 3: Classification Time for Different Clustering Methods

Clustering Method	talk.politics. guns	alt.atheism	comp. windows.x
Modified FCM	0.58	0.51	0.41

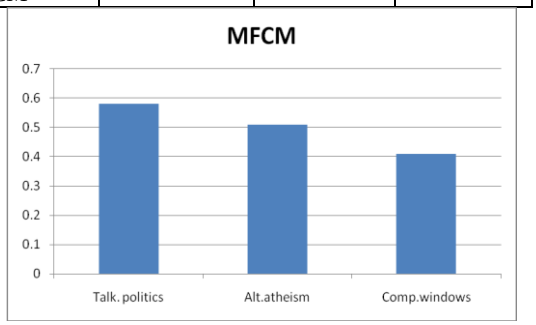

Figure 3: Clustering Time Comparison for the Proposed Technique and Existing Technique

V.CONCLUSION AND FURURE WORK

In this paper, we proposed a new method for clustering documents using the relationship between the existing documents and other documents. To evaluate the efficiency of this system, we make experiments on clustering newsgroup documents by using our method and by using mfcm algorithm. As the results of these experiments, we found that it is effective for document clustering to combine the confusion matrix. Moreover, the proposed method is more effective for the document clustering in comparison with the clustering purity and the entropy accuracy. We believe that these results are encouraging to consider future research on unsupervised clustering approaches as highly reliable.

Further work would be required to compare the Modified Fuzzy C mean (MFCM) and the Fuzzy C means (FCM) clustering methods by the many kinds of document data.

REFERENCES

[1] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2000.

[2] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Kluwer, Norwell, MA, 1981.

[3] W. Pedrycz, Conditional fuzzy C-means, Pattern Recognition Letters 17 (1996) 625–632.

[4] J.Li, X.B.Gao, L.C.Jiao, *A novel typical-sample-weighting clustering algorithm for large datasets*, LANI, vol. 3801,2005.

[5] X. Wang. A Course in Fuzzy Systems and Control. Prentice Hall, Inc, Upper Saddle River, NJ, 1997.

[6] M. F. Porter,"An algorithm for suffix stripping", Program; automated library and information systems, 14(3), 130- 137, 1980.

[7] Aggarwal, C. C., Wolf, J. L., Yu, P. S., Procopiuc, C., & Park, J. S. (1999). Fast algorithms for projected clustering. ACM SIGMOD Conference (pp. 61–72).

[8] Zhao, Y., & Karypis, G. (2001). Criterion functions for document clustering: Experiments and analysis (Technical Report). Department of Computer Science, University of Minnesota.

[9] R. Baeza-Yates and B. Ribeiro-Neto (1999). Modern Information Retrieval. New York: Addison Wesley, ACM Press, 1999.

[10] Nikravesh, L. A. Zadeh, B. Azvin and R. Yager (editors). Enhancing the Power of the Internet - Studies in Fuzziness and Soft Computing, Springer, vol. 139, pp. 255-278, January 2004

[11] Pallav Roxy, and Durga Toshniwal, "Clustering Unstructured Text Documents Using Fading Function", International Journal of Information and Mathematical Sciences, Vol 5, NO. 3 2009.

[12] Shady Shehata, Fakhri Karray and Mohamed S. Kamel, "An Efficient Model For Enhancing Text Categorization Using Sentence Semantics", International Journal of Computational Intelligence, 2010.

[13] Jun Zhai, Yan Chen, Qinglian Wang and Miao Lv "Fuzzy ontology models using intuitionistic fuzzy set for knowledge sharing on the semantic web", 12th International Conference on Computer Supported Cooperative Work in Design, 2008.

[14] A. Hinneburg and D.A. Keim. Optimal gridclustering: Towards breaking the curse of dimensionality in high-dimensional clustering. In Proc. of VLDB-1999, Edinburgh, Scotland, September 2000. Morgan Kaufmann, 1999.

[15] H. Schuetze and C. Silverstein. Projections for efficient document clustering. In Proc. of SIGIR-1997, Philadelphia, PA, July 1997, pages 74–81. Morgan Kaufmann, 1997.

[16] Liping Jing," Survey of Text Clustering", Department of Mathematics, The University of Hong Kong, HongKong, China, , ISBN: 7695-1754-4/02

[17] G. Stumme, R. Taouil, Y. Bastide, N. Pasquier and L. Lakhan, "Computing iceberg concept lattice with Titanic", Journal on Knowledge and Data Engineering, Vol. 42, No. 2, 2002, pp. 189-222.

[18] S. Pollandt, Fuzzy-Begriffe: Formale Begriffsanalyse unscharfer Daten, Springer Verlag, Berlin- Heidelberg, 1996.