Implementation of Fuzzy K-Means In Multi-Type Feature Coselection For Clustering

K.Parimala, V.Palanisamy

Abstract— Feature Selection is a preprocessing technique in supervised learning for improving predictive accuracy while reducing dimension in clustering and categorization. Multitype Feature Coselection for Clustering (MFCC) with hard k means is the algorithm which uses intermediate results in one type of feature space enhancing feature selection in other spaces, better feature set is co selected by heterogeneous features to produce better cluster in each space. Soft Clustering is an optimization technique of data analysis and pattern recognition which allocates a set of observations to cluster in a fuzzy way, constructing a membership-function matrix whose (i, j)th element represents the "the degree of belonging" of the ith observations to the jth cluster. This paper presents the empirical results of the MFCC algorithm with soft clustering and also gives the comparison results of MFCC with hard and soft k means. Fuzzy k-means clustering is proposed for getting the robustness against the outliers.

Index Terms— Feature Selection, MFCC, Fuzzy k-means.

I. INTRODUCTION

Information or knowledge can be conceptualized as data. It reflects in the data norm, the size and dimensions have improved high and more. The feature selection plays a vital role in machine learning, data mining, information retrieval, etc. the goal of feature selection is to identify those features relevant to achieve a predefined task. Many researchers have been to find how to search feature subset space and evaluate them.

In supervised methods [1], the correlation of each feature with the class label is computed by distance, information dependence or consistency measures [2]. In unsupervised method the feature selection does not need the class of information such as document frequency and term strength [3]. The newly proposed methods namely Entropy based feature ranking method (En) proposed by Dash and Liu [4] in which feature importance is measured by the contribution to an entropy index based on the data similarity; the individual 'feature saliency' is estimated and an Expectation Maximization (EM) algorithm using Minimum message length is derived to select the feature subset and the number of clusters [5].

While the methods above are not directly targeted to clustering text documents, [6] proposes two other feature selection methods for text clustering. One is Term Contribution (TC) which ranks the feature by it overall contribution to the document similarity in the data set. The other is Iterative feature selection (IF), which utilizes some successful feature selection methods such as Information Gain (IG) and CHI-Square (χ^2) text to iteratively select features and performs text clustering at the same time.

Manuscript received on November, 2012.

Mrs. K.Parimala, Assistant Professor, MCA Department, NMS S.Vellaichamy Nadar College, Madurai-625019, TamilNadu, India.

Dr. V.PalaniSamy, Professor & Head In-Charge, Department of Computer Science & Engineering, Alagappa University, Karaikudi, TamilNadu, India.

[7] Combines information about document contents and hyper link structures to cluster documents. The hypertext documents in a certain information space were clustered into a hierarchical form based on contents as well as link structure of each hyper text documents.

From the ideas of [8] & [9] co-training algorithms learn through classifiers over each of the feature set and combine their predictions to decrease classification error. Cot raining algorithm can learn from unlabelled data starting from a weak predictor.

Clustering helps users, tackle the information overload problem in several ways: explore the contents of a document collection; group duplicate and near duplicate documents. Unsupervised method can hardly achieve a good performance when evaluated using labeled data.

Data fusion [10] is well suited to-problems involving massive amounts of data where each subsystem may not have entire data set, problems with many possible approaches, allows for natural and flexible distribution of resources aim to provide better performance than best input system. Voting procedures are examples of data fusion – results from identical data sets are merged.

This paper is devised to show the results of MFCC algorithm using soft k means. This paper is organized as follows: Next we describe prior related work describing MFCC and soft clustering. Section 3 describes the learning of MFCC with soft k means. Then in section 4, the experiments and evaluation results are explained and discussed finally, section 6 describes the conclusion and future works.

II. RELATED WORK

A. Multitype Features Coselection for Clustering (MFCC):

In this section we briefly discuss about MFCC. It is made clear that the selection of each type feature and the clustering is an iterative one. After one iteration of clustering, each data object will be assigned to a cluster. In [6], Liu et al. assumed each cluster corresponded to a real class. Using such information, they did supervise feature selection, such as Information Gain (IG) and $\chi 2$ statistic (CHI) [2] during k-means clustering. MFCC tries to fully exploit heterogeneous features of a web page like URL, anchor text, hyperlink, etc., and to find more discriminative features for unsupervised learning. We first use different types of features to do clustering independently. Then, we get different sets of pseudoclass, which are all used to conduct iterative feature selection (IF) for each feature space.

After normal selection, some data fusion methods are used to conduct iterative feature selection (IF) for each feature space, i.e., feature coselection. In each iteration of clustering, the coselections in several spaces are conducted one by one after clustering results in different feature spaces have been achieved before any coselection. Thus, the sequence of coselection will not affect the final performance. The general

idea of coselection for k-means clustering is described in fig-1.

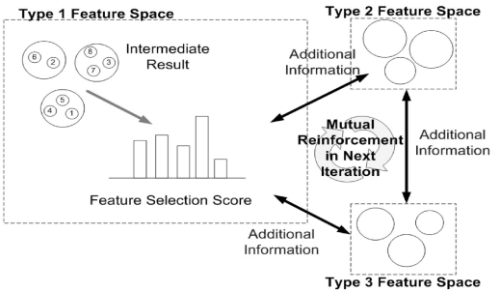

Fig.1. The basic idea of Multitype feature coselection.

Suppose that we categorize data objects with M heterogeneous features into L clusters. Let fv_n be one dimension of the feature vector, icr_i be the intermediate clustering results in the i^{th} feature space, SF be the fusion function.

The pseudo algorithm is listed as follows:

Loop for N iterations of k-means clustering

{

Loop for m feature spaces

{

Do clustering in feature space m

}

Loop for M feature spaces

{

For feature space m, do feature selection using results in all feature spaces.

For (fv_n) one dimension of the feature vector in space m, a feature selection

score fss (fv_n, icr_i) is obtained by using intermediate clustering results icr_i in feature space i.

Then a combined score fss (fv_n) is achieved by fusing the scores based on

different result sets.

$$fss(fv_n) = SF(fss(fv_n, icr_i)) \quad (1)$$

}

}

In the equation (1), $fss(fv_n, icr_i)$ can be the value calculated by the selection function or rank among all features. The feature selection criteria, the six commonly used feature selection function mentioned in [2]:

Function	Mathematical form
$IG(t_k, c_i)$	$p(t_k,c_i)\cdot\log\frac{p(t_k,c_i)}{p(c_i)\cdot p(t_k)} + p(\overline{t_k},c_i)\cdot\log\frac{p(\overline{t_k},c_i)}{p(c_i)\cdot p(\overline{t_k})}$
$\chi^2(t_k, c_i)$	$\frac{N\cdot(p(t_k,c_i)\cdot p(\overline{t_k},\overline{c_i}) - p(t_k,\overline{c_i})\cdot p(\overline{t_k},c_i))^2}{p(t_k)\cdot p(\overline{t_k})\cdot p(c_i)\cdot p(\overline{c_i})}$
$CC(t_k, c_i)$	$\frac{\sqrt{N}\cdot(p(t_k,c_i)\cdot p(\overline{t_k},\overline{c_i}) - p(t_k,\overline{c_i})\cdot p(\overline{t_k},c_i))}{\sqrt{p(t_k)\cdot p(\overline{t_k})\cdot p(c_i)\cdot p(\overline{c_i})}}$
$RS(t_k, c_i)$	$\log\frac{p(t_k \mid c_i)+d}{p(t_k \mid \overline{c_i})+d}$
$OR(t_k, c_i)$	$\frac{p(t_k \mid c_i)\cdot(1-p(t_k \mid \overline{c_i}))}{(1-p(t_k \mid c_i))\cdot p(t_k \mid \overline{c_i})}$
$GSS(t_k, c_i)$	$p(t_k,c_i)\cdot p(\overline{t_k},\overline{c_i}) - p(t_k,\overline{c_i})\cdot p(\overline{t_k},c_i)$

Table-1 Feature Selection Functions.

Depending on the choices of fss and SF, we obtain five fusion models including voting, average value, max value, average rank, and max rank. The equations are listed as follows:

$MaxRank(Rank(f_{vn})) = \arg\max(Rank(fvn,icr_i))$
$AverageRank(Rank(fvn))=(\Sigma Rank(fvn,icr_i))/M$
$Voting(val(fvn)) = \Sigma vote(fvn,icri)$
$Vote(fvn,icri)=\{0\ val(fvn,icri) < st$
$1\ val(fvn,icri) >= st$
$Average(val(fvn))= \Sigma\ val(fvn,icri)/M$
$Max(val(fvn)) = \arg\max(val(fvn,icri))$

Table-2 Fusion Models

In the above equation, $val(fv_n,icr_i)$ is the value calculated by selection function, $RANK(fv_n,icr_i)$ is the rank of fv_n in the whole feature list ordered by $val(fv_n,icr_i)$, and st is the threshold of feature selection. After feature coselection, objects will be reassigned, features will be reselected, and the pseudoclass-based selection score will be recombined in the next iteration. Finally, the iterative clustering and feature coselection are well integrated.

In each of the iterations, the whole feature space should be reconsidered. The reason is that our method can help in finding more effective features through a mutual reinforcement process. Properly selected features will help clustering and vice-versa. That is to say, some discriminative features will not be found until late in the clustering phase. This can be proved by empirical results.

B. Soft Clustering

The fuzzy or soft clustering is a method of data analysis and pattern recognition which allocates a set of observations to clusters in a "fuzzy way", more formally, constructs a membership function matrix whose $(i,j)^{th}$ element represents "the degree of belonging or membership" of the i^{th} observation to the j^{th} cluster:

1. Make initial guesses for the means m_1, m_2, m_3,..,m_k.
2. Until there are no changes in any mean:
   - Use the estimated means to find the degree of membership U(i,j) of x_i in cluster j; for example, if $A(i,j)=\exp(-|x_i - m_j|^2)$ one might use $U(i,j)=a(i,j)/\text{sum}_j\ a(i,j)$.
   - For j = 1 to k

     Repeat m_j with fuzzy mean of all examples for cluster j...

$$m_j = \frac{\sum u(i,j)^2 . x_i}{\sum u(i,j)^2}$$

     end-For.

end-Until.

We can define many fuzzy rules to verify or classify the existence of term or query in each document. One such generalized if-then-form of rule is,

R^k : IF x_i^k is X_i^k AND x_n^k is X_n^k

THEN y is Y^k

where in X_i^k , i = 1..n and k = 1..N, n is clusters and N is terms.

We define a fuzzy distance which allow us to classify the set of vectors{x_1,..,x_m} into n classes a_i, i = 1..n.

Each vector means the number of occurrences of a term in each document, i.e., there are vectors $x_i=\{x_{ij}\}$, j = 1..n, where x_{ij} means number of times that terms appears on document j.

Fuzzy generalization of F(V,C) is obtained by involvement of the membership with an exponent Φ, hence

$$F_x(V,C) = \sum_{i=1}^{n} \sum_{c=1}^{k} V_{ic}^{\Phi} d_{ic}^2 \qquad (2)$$

To be minimized under conditions-mutually exclusive, jointly exhaustive and non-empty continuous class instead of all-or-none membership. The exponent Φ is chosen in advance from [1,α). It determines the degree of fuzziness of the solution with the lowest meaningful value, Φ=1, the solution of

$$F_x(V,C) = \sum_{i=1}^{n} \sum_{c=1}^{k} V_{ic}^{\Phi} d_{ic}^2$$

is a hard partition, i.e., the result is not fuzzy at all. As Φ approaches α, the solution approaches its maximum degree of fuzziness, with V_{ic} = 1/k for each pair of i and c.

By their nature, continuous classes should provide better representation of outliers or typical individuals than discontinuous class. Fuzzy k means for instance, will indeed give intermediate memberships to outliers.

III. PROPOSED WORK

A. *FUZZY k means in MFCC*

In this paper we present a method to build a clustering system that merges MFCC with fuzziness. The general idea for modification is based on the coselection and fuzziness "the degree of belongingness", of fuzzy k-means clustering algorithm; MFCC reduces the noise feature effectively by and improved further performance. The modified MFCC got the idea from the fuzzy generalization, where we get intermediate membership to the noise features. So that the selection score for the modified MFCC will be as,

$$\text{fss}(V_n, icr_j) = \mathbf{SF}\left(\text{fss}\left(\sum_{i=1}^{n} \sum_{c=1}^{k} \mathbf{V_{ic}^{\Phi}\, d_{ic}^2, icr_j}\right)\right) \qquad (3)$$

where as,

SF – selection function to fuse the feature space selection (or, the intermediate clustering)

FSS – feature selection score to select best center point (or, mean) from the specified feature space.

V_{ic} - the membership of class, x_i document present with c centre in jth feature space.

d_{ic} - distance of the x_i to the centre of feature space.

icr_j - intermediate clustering of jth iteration.

Φ - Fuzzy ratio variance.

In the equation (3) fss (V_n,icr_j) can be the value calculated by the selection function or the rank among all features. Depending on the values of V_{ic}^{Φ} d_{ic}^2, the documents are clustered and features to different classes. And according to the values of fss & SF, five fusion models are obtained. After feature coselection, the objects are rearranged, features will be reselected and the pseudoclass –based selection score will be recombined for next iteration.

B. *Experiments & Results:*

The soft MFCC proposed in the paper has \been fully implement and evaluated with extensive experimentation; this section presents the details of implementation, data set and text results

Evaluation metrics:

A number of metrics used in feature selection and clustering are evaluated and measures for categorization effectiveness. We use the best recall k precision metrics. Such measures are F-measure and time precision in each fss criteria.

F-measure is calculated by the harmonic mean of vocabulary terms (P) and total terms(R). Each fss criteria define the P & R terms.

We also use accuracy in the paper as a measure. Accuracy is computed as the ration of correctly classified testing documents to the total number of testing documents. Of course, all these performance metrics are computed for each category separately (i.e.) we apply all the testing documents to each fss criteria to compute P, R, f1, and accuracy for each fss criteria.

C. *Experiment Results:*

The experimental evaluation was performed on testdata data set. Here we can explain and results on testdata dataset. The testdata contains almost 255 articles, evenly distributed on 10 categories. Further each article can be assigned to one/more category. In our experiments, following the MFCC, we ran a test on categories, which are the categories having highest number of documents.

Fuzzy k-means/soft clustering with MFCC is verified with test data database (Table – 1). It contains feature classes of HTML, text files, word documents, jpeg files, user logs, etc.

Classes	No of documents	Related terms	Total term frequency
ASP	2	22	23
CSS	10	1439	6771
Gif	144	975	976
Html	25	14210	63392
Jpeg	18	3554	3659
Js	19	4935	38415
Pdf	5	249229	398036
Php	10	1670	4644
Png	9	245	245
Ppt	13	193505	208541

Table – 1. Feature Classes of test database.

MFCC algorithm clusters the dataset according to the query term. TF-IDF is calculated and the following result is got for CHI-square, correlation coefficient, GSS coefficient, Information gain for each feature class.

The fuzzy K-means or soft clustering works in the concept of "degree of belongingness". In soft clustering each document gets into single cluster with its maximum degree of fuzziness (refer fig – 2).

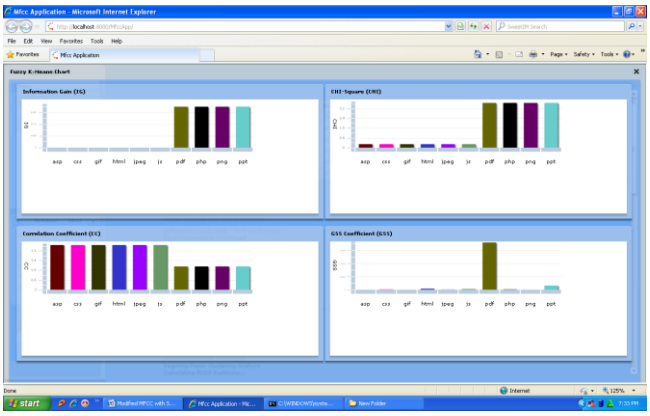

Fig – 2 Fuzzy K-means.

The following (table – 2) lists not the selection function – such as IG, CHI-square, correlation coefficient and GSS coefficient results for the fuzzy value and accuracy value and the time to calculate fss. As the accuracy value lineates or increases the time gets reduced.

	Selection Function	Time
Fuzzy Value = 0 Accuracy = 0.1	IG	16ms
	CHI	156 ms
	CC	0 ms
	GSS	0ms
Fuzzy Value = 1 Accuracy = 0.1	IG	15ms
	CHI	0ms
	CC	16ms
	GSS	0ms
Fuzzy Value = 10 Accuracy = 0.05	IG	0ms
	CHI	0ms
	CC	0ms
	GSS	0ms

Table – 2. Fuzzy k-means MFCC

The testdata database is verified with hard k-means MFCC. The result is shown in fig-3; the hard k-means clusters the classification into two clusters according to the query.

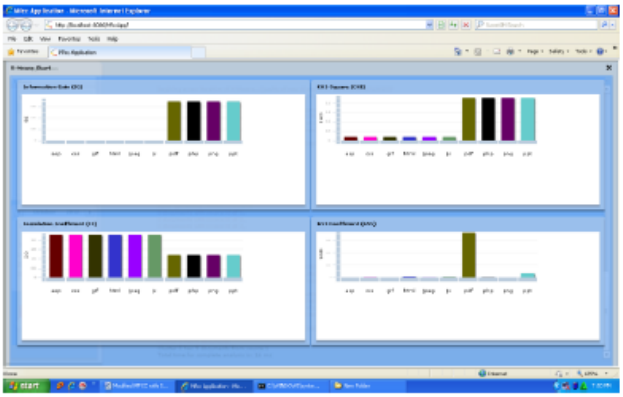

Fig – 3. Hard k-means

The hard k-means and fuzzy k-means shows the results more or less similar, they differ in time factor, soft k-means works in less timing than hard k-means (table - 3).

Fuzzy k-means	Hard K-means	
1	2	No. of clusters
Ig = 0.3, 0ms Chi = 3.6, 0ms Cc = 3.2, 0ms Gss = 0.2, 0ms	Ig = 7.27073772, 31ms Chi = 7.2707372, 16ms Cc = 11.693763, 31ms Gss = 3.6349692, 16ms	K value & time for selection function

Table – 3. Comparison of fuzzy k-means and hard k-means

The soft clustering shows better result than hard k-means clustering. Even though the two clusters show the same result, they differ in time factor (fig – 4)

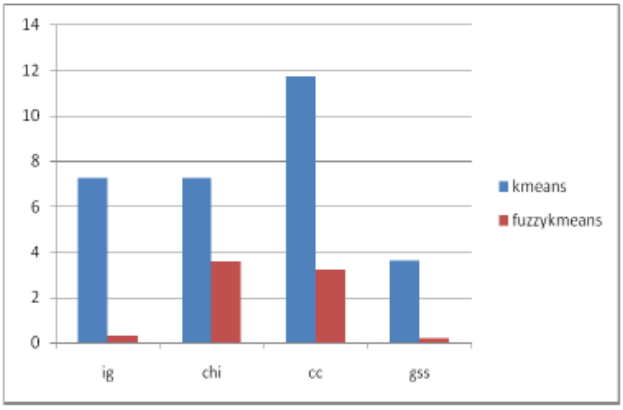

Fig – 4. Comparison of Fuzzy k-means & Hard k-means

IV. CONCLUSION

The MFCC algorithm implemented in fuzzy k-means by in feature selection score (fss) measures. It have been proved that fuzzy k-means shows better results than hard k-means. The outliers are reduced in MFCC hard k-means algorithm. The outliers are further reduced or totally removed in fuzzy k-means, since the accuracy value depends on the 'degree of belongingness'. The MFCC with k-means algorithm can be implemented or further extended to other data sets and applications.

REFERENCES

[1] M. Dash and H. Liu, "Feature Selection for Classification," Int'l J. Intelligent Data Analysis, vol. 1, no. 3, pp. 131-156, 1997.

[2] Y. Yang and J.O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," Proc. Int'l Conf. Machine Learning (ICML '97), pp. 412-420, 1997.

[3] Shen Huang, Zheng Chen, Yong YU & Wei_Ying Ma, "Multi type Features Coselection for Web document Clusering", IEEE Transactions on Knowledge and Data Engineering; vol-18,no.4,April 2006.

[4] M. Dash and H. Liu, "Feature Selection for Clustering," Proc. 2000 Pacific-Asia Conf. Knowledge Discovery and Data Mining, pp. 110 121,2000.

[5] H.C.L. Martin, A.T.F. Mario, and A.K. Jain, "Feature Saliency in Unsupervised Learning," Technical Report, Michigan State Univ., 2002

[6] T. Liu, S. Liu, Z. Chen, and W.-Y. Ma, "An Evaluation on Feature Selection for Text Clustering," Proc. Int'l Conf. Machine Learning (ICML'03), pp. 488-495, 2003.

[7] R. Weiss, B. Velez, M.A. Sheldon, C. Namprempre, P. Szilagyi, A. Duda, and D.K. Gifford, "HyPursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering,"Proc. Seventh ACM Conf. Hypertext, pp. 180-193, 1996.

[8] K. Nigam and R. Ghani, "Analyzing the Effectiveness and Applicability of Co-Training," Proc. Information and Knowledge Management, pp. 86-93, 2000.
[9] A. Blum and T. Mitchell, "Combining Labeled and Unlabeled Data with Co-Training," Proc. Conf. Computational Learning Theory, pp. 92-100, 1998.
[10] M. Montague, "Metasearch: Data Fusion for Document Retrieval," PhD Thesis, Dartmouth College, 2002.

Mrs.K.Parimala.,MCA, Research Scholar in Computer Science in Alagappa University, Karaikudi, INDIA, under the guidance of **Dr.V.Palanisamy**, Professor & Head In-Charge, Department of Computer Science & Engineering, Alagappa University. Currently working as Assistant Professor, in NMS SVN College with a teaching experience of 10 years.

Dr. V.PalaniSamy, MCA, MTech (Adv.IT), Ph.D, Professor & Head In-Charge, Department of Computer Science & Engineering, Alagappa University, Karaikudi, TamilNadu, INDIA, specialized in Algorithms, Wireless Networks & Network Security. He has 20 years of teaching experience and 15 years of Research experience.