INDEXING BANGLA NEWSPAPER ARTICLES USING FUZZY AND CRISP CLUSTERING ALGORITHMS

A. K. M. Zahiduzzaman, Mohammad Nahyan Quasem, Faiyaz Ahmed and Rashedur M. Rahman
Department of Electrical Engineering and Computer Science, North South University, Bashundhara, Dhaka, Bangladesh

Keywords: Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Factor Analysis (FA), Intensification, c-Means clustering, Fuzzy c-Means clustering.

Abstract: The paper presents two document clustering techniques to group Bangla newspaper articles. The first one is based on traditional c-means algorithm, and the later is based on its fuzzy counterpart, i.e., fuzzy c-means algorithm. The key principle for both of those techniques is to measure the frequency of keywords in a particular type of article to calculate the significance of those keywords. The articles are then clustered based on the significance of the keywords. We believe the findings from this research will help to index Bangla newspaper articles. Therefore, the information retrieval will be faster than before. However, one of the challenge is to find the salient features from hundred of features found in documents. Besides, both clustering algorithms work well on lower dimensions. To address this, we use three dimensionality reduction techniques, known as Principle Component Analysis (PCA), Factor Analysis (FA) and Linear Discriminant Analysis (LDA). We present and analyze the performance of traditional and fuzzy c-means algorithms with different dimensionality reduction techniques.

1 INTRODUCTION

Clustering is a technique that allows grouping of related data. The clustering is generally carried out in two dimensional spaces where each sample point is represented by an X and Y coordinate. The samples are then grouped together according to some inherent property that related points share. Among many clustering algorithms, we deal with traditional c–means (mostly known as k-means algorithm) and the fuzzy c-means algorithm (Han and Kamber, 2000). Both of those techniques use the idea of grouping points against a pre selected centroid within the points. The distance between the points and the centroid is calculated and they are assigned to relevant centroids according to some proximity. A new centroid is calculated using the mean of all points and the process repeats until no new centroids are found. The primary difference between traditional c–means and fuzzy c–means is the fact that the former uses Euclidean distance to measure proximity and determine corresponding clusters, where the later uses a degree of membership to a particular cluster.

Our research uses the idea of clustering to index Bangla newspaper articles, according to their topic. The frequency and significance of the keywords are the features that are used to make the clusters. The objective for this research is to make Bangla information retrieval faster and more meaningful.

The paper introduces data acquisition in section 2. Methodology conducted in this research is discussed in section 3. Section 4 presents and analyzes the performance of fuzzy and traditional c-means algorithms. Finally, section 5 concludes and gives direction of future research.

2 DATA ACQUISITION

Data is collected manually in our research from a Bangla newspaper, ProthomAlo (Prothm Alo, 2011) which is available online. Figure 1 shows the home page of that newspaper. We look for different type of articles in this research and we will deal with 9 different types of articles.

The articles are collected for various days. An example of such an article is shown in Figure 2. We collect 332 articles of 9 different types. The articles type and their frequencies are depicted in Table 1.

K. M. Zahiduzzaman A., Nahyan Quasem M., Ahmed F. and M. Rahman R..
INDEXING BANGLA NEWSPAPER ARTICLES USING FUZZY AND CRISP CLUSTERING ALGORITHMS.
DOI: 10.5220/0003492603610364
In *Proceedings of the 13th International Conference on Enterprise Information Systems* (ICEIS-2011), pages 361-364
ISBN: 978-989-8425-53-9

Figure 1: Prothom Alo home page.

Table1: Type and number of articles taken.

Number	Classes of Article	
Assigned	Name	Freq
1	Crime	24
2	Politics	35
3	Business	50
4	Development	26
5	Education	16
6	Sports	50
7	Entertainment	50
8	General	31
9	International	50
	Total	**332**

Figure 2: A news on Share Market.

3 METHODOLOGY

We have designed a very structured pre-processing stage where we tokenize each word and prepare the document readable to the system. Then we intensify the features and fed it to the clustering algorithms to group documents. The clustering algorithms perform better in lower dimensions, therefore from hundreds of features we only keep salient features by dimensionality reduction techniques, namely LDA (Fisher, 1936, McLachlan, 2004), PCA(Pearson, 1901) and FA (MacCallum,1983).

3.1 Pre-processing: Unicode Arrays as Words

The first part of the extraction is to get the article text from the Prothom Alo web page and keep it in Unicode (big endian) encoded text document. We use Matlab (Maaten, 2007) as the platform where we test our algorithm. It is not possible to manipulate Unicode data in Matlab. Therefore, we gather a set of delimiters for the strings including the white space and then tokenize every word as array of Unicode unsigned 16 bit numbers.

For example the word “tumi” in Bengali is represented like this:

2468 2497 2478 2495

3.2 Keyword Extraction

We use mainly the frequency of the keyword as feature value. Initially collection of words from all articles is considered as features. Later we pass those words in two phases, namely keyword intensification and rejection phase.

3.2.1 Intensification

We have built a database where we store some words which are the major keys to a particular article, for example, Bangla word for murder will mostly occur in the crime articles rather than a sports article. If we get those words we will increase the frequency count of the feature by multiplying with a factor. This multiplying factor may be learnt by machine learning algorithm. However, we find the value by trial and error method and a good approximation found is 10.

Imposing this intensification we have a weighted frequency feature extraction. This technique will make the articles biased towards main keys which will help to get scattered in group with the class itself.

3.2.2 Rejection

We also include in the database a list of bad words that commonly occur to articles but are meaningless and irrelevant to the article. If we have those words in an article, we exclude those. There are two ways we could deal extracting these keywords:

1. Select some known previous keywords
2. Learn from the article.

We preferred the later one to use in our system. This way we could make our system flexible that allows us to categorize almost any type of crime article that could be cyber-crime or neighbourhood crime. Both will be categorized as crime articles.

3.3 Dimensionality Reduction

The dimension that is found after the key word extraction is huge to handle so we had to reduce the dimension. As there are almost 10,000 features in a document, we got the vectors with 10,000 dimensions. We use PCA first to reduce this huge dimension. Then we use the Factor Analysis (FA) and LDA. The PCA first reduce the dimension to 250. Then two other techniques, one is supervised (LDA) and other is unsupervised (FA) to make it down to 4 and 5. We have tested with other dimensions and find that the clustering algorithms do not perform well in higher dimensions. It performs best in 4th and 5th dimensions.

3.4 Clustering

We use both traditional and fuzzy c-means clustering to group articles. We analyze the performance through confusion matrix. The samples in the fuzzy clustering algorithm if had a membership to a particular cluster as maximum membership for that sample *m,* then, we checked whether the sample gives *k*m* membership to other clusters. This will introduce the concept that we explained earlier that which article goes to which cluster can sometimes vary on the context. This way we could create the opportunity to classify one article as crime and say, politics if necessary. The crisp c-means doesn't have that special feature.

4 RESULT ANALYSIS AND DISCUSSION

The performance of the two clustering methods were not much apart in terms of the accuracy. Although our expectation was that the fuzzy clustering would perform better, the traditional c-means algorithm also perfom well. The variaion of the results is due to different feature reduction technique.The accuracy of the clustering techniques is presented in Table 2.

The blue shades in Figure 3 are the accuracy rates for Factor Analysis feature reduction technique and others are for LDA feature reduction technique. The graph clearly visualizes the differences between the

Table 2: Accuracy Rates.

	3 Feat	4 Feat	5 Feat
Fuzzy LDA	91.5663	87.9518	93.3735
Fuzzy FA	31.3253	31.3253	30.7229
Crisp LDA	88.5542	91.5663	92.7711
Crisp FA	43.9759	43.3735	33.7349

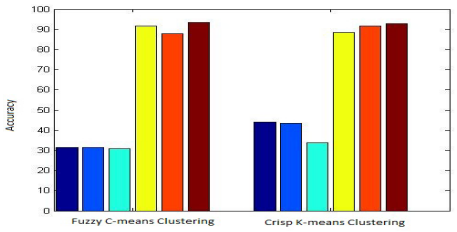

Figure 3: Accuracy plots.

two feature reductions techniques although, the algorithm for the clustering does not much differ.

Both of the algorithms are iterative and they recalculate the centroids for each cluster using some data for the particular algorithm. The traditional c-means algorithm calculates the Euclidian distance to calculate the centroids in each iteration and the fuzzy c-means calculates the centroids using the membership values which again are calculated by the Euclidian distance. So, getting the similar result is more likely.

LDA is a supervised dimensionality reduction technique. Therefore, the extracted features get biased towards a particular class. This helps the samples of the same class to converge to a particular cluster. Since FA is unsupervised, i.e., it does not use any class information, hence samples are grouped more sparsely than LDA in dimension space.

The features for the data are extracted in two different techniques. First we use LDA to reduce to 5 features. Each row in the Table 3 represents a single class of article that was collected into the database. Each column represents the cluster into which a particular sample has been included. However, the cluster number does not necessarily correspond to the class number. In this table, what is important to note is, how samples of a particular class exists in a particular cluster. For example, most of the samples of class 4 have been clustered in cluster 4 (cl4). On the other hand, most samples of class 7 belong to cluster 1, the rest belong to cluster 8. The point that we want to stress is that the cluster number is arbitrary and has nothing to do with a particular class. In this case 92% of class 7 is clustered in cluster 1; we consider this as the accuracy of the clustering for class 7. Using the above theory, the average accuracy of this system, is 93.3735%. Now let us take a look at the membership values for all of one class into all clusters. Here the following example shows the membership values for article class 6 for the current system.

Table 3: Confusion matrix for 5 feature LDA Fuzzy c-means algorithm.

	Cl1	Cl2	Cl3	Cl4	Cl5	Cl6	Cl7	Cl8	Cl9
C1	0	0	0	0	23	0	0	1	0
C2	0	0	29	0	4	0	0	2	0
C3	0	0	0	0	0	0	0	2	48
C4	0	0	0	24	0	0	0	2	0
C5	0	16	0	0	0	0	0	0	0
C6	0	0	0	0	0	50	0	0	0
C7	46	0	0	0	0	0	0	4	0
C8	1	0	0	0	0	0	0	29	1
C9	0	0	0	0	0	0	45	5	0

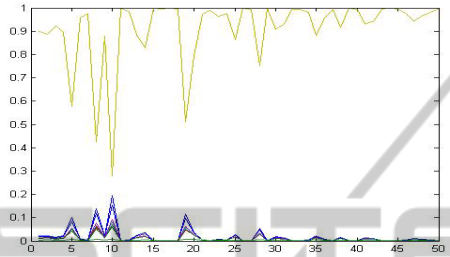

Figure 4: Membership of the Articles in class 6.

In the Figure 4 we plot the membership value of the samples on Y axis and sample numbers on X axis. The olive line is the maximum membership value for documents in class 6 (cl6 cluster membership) where other are other line represent membership to others. The Cluster 6 is the most dominant cluster for this class. So we say documents in class 6 are clustered in 6th cluster. On the other hand, if we observe the class 2 articles we will see a very different result.

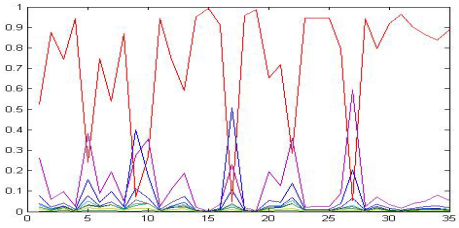

Figure 5: Membership of the articles in class 2 using LDA.

The class contains 35 articles where the cluster 3, which is represented as a magenta line in the plot above, has the most articles in it. But around the 10th, in between 15th and 20th and in between 25th and 30th samples, we observe some blue lines and purple lines which exceed the value of membership for that of the magenta line. This enumerates some documents are wrongly clustered in other clusters, i.e., cluster 5 (blue line) and cluster 8 (purple line). Table 3 contains all these data in matrix form.

It is possible that documents overlap to multiple clusters. A particular sample might have same or very (90%) close membership to multiple clusters.

Figure 6 presents a situation where very low performance is achieved by unsupervised FA feature extraction technique. There are 4 features extracted and the membership of the samples in class 2 is shown here:

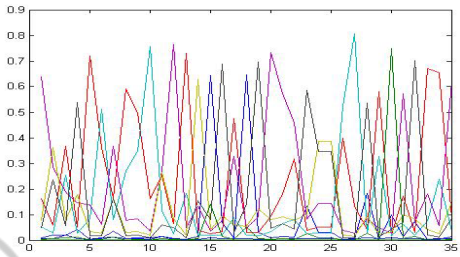

Figure 6: Membership of the articles in class2 using FA.

It is really hard to predict from the graph how articles are clustered to a particular cluster, rather distributed randomly to multiple clusters.

5 CONCLUSIONS AND FUTURE WORK

Experimental results show that reducing number of features using LDA prevail over FA in terms of accuracy. However, choosing the number of features has a significant impact on accuracy. Future work will include supervised data collection that could result in better clustering. By improving the weighted frequency calculation we could get better results. Lastly, neuro-fuzzy clustering could be exploited to learn parameters by the system itself and perform optimally.

REFERENCES

Fisher, R, 1936. *The Use of Multiple Measurements in Taxonomic Problems* In: Annals of Eugenics, 7, p. 179—188.

Han J., Kamber, M., 2000. *Data Mining Concept and Techniques*, Morgan Kaufmann Publishers.

Maaten, L. J. P. van der, 2007 *An Introduction to Dimensionality Reduction Using Matlab*, Technical Report MICC 07-07. Maastricht University, Maastricht, The Netherlands.

MacCallum, R, 1983. *A comparison of factor analysis programs in SPSS, BMDP, and SAS*. Psychometrika 48 (48).

McLachlan, 2004. *Discriminant Analysis and Statistical Pattern Recognition* In: Wiley Interscience.

Pearson, K. 1901. *On Lines and Planes of Closest Fit to Systems of Points in Space*. Philosophical Magazine 2 (6): 559–572.

Prothom Alo website, 2011. http://www.prothom-alo.com