Fuzzy Logic for Document Clustering

Bhushan Talekar M.E – Computer Engineering VIVA Institute of Technology

Saniket Kudoo B.E – Computer Engineering VIVA Institute of Technology

Pragati Patil M.Tech – Information Technology VIVA Institute of Technology

Pallavi Vartak M.E – Information Technology VIVA Institute of Technology

ABSTRACT

This paper shows document clustering by applying fuzzy logic. The method involves cleaning up the text and stemming of words. Then, chose 'm' features which differ significantly in their word frequencies (WF), normalized by document length, between documents belonging to these two clusters. The documents to be clustered are represented as a collection of 'm' normalized WF values. Then use Fuzzy c-means (FCM) algorithm to cluster these documents into two clusters. After the FCM execution finishes, the documents in the two clusters are analyzed for the values of their respective 'm' features. By using fuzzy logic, we not only get the cluster name, but also the degree to which a document belongs to a cluster.

Keywords

Fuzzy c-means algorithm, fuzzy logic , Document clustering

1. INTRODUCTION

This paper proposes fuzzy logic to text mining to perform clustering of documents into a number of pre- specified clusters. In an example it is shown how to classify given documents into two categories with the fuzzy system that is sports and politics. Initially, the documents are cleaned. After that for every word we carry out word stemming. Each documents will now be treated as a bag of words. The calculation of the word frequency(weight of the word on basis of their significance in the document)is done as :

WF = (Word Count/(Total Words in the Document)) x10000 (i)

Where 'm' no of words are selected[3][4][5]. After this we them perform FCM to cluster the documents into required number of categories. To name these particular cluster pre known knowledge is used.

2. FUZZY LOGIC

Fuzzy logic is nothing but mathematical logic model in which truth can be partial i.e. it can have values between 0 and 1 that may be completely false or completely true which is based on approximate reasoning instead of exact reasoning. Fuzzy logic is an approach to computing based on "degrees of truth

"rather than usual true or false on which modern computer is based.

3. CLUSTERING

Fuzzy logic helps us to cluster similar documents together.

In this algorithm every observation has a membership value associated with each of the cluster which is related inversely to the distance of that observation from the cluster centre.

4. ALGORITHM FOR FUZZY C-MEANS

1. Initialize the fuzzy partition matrix, U=[uij] matrix, U(0)

2. At k-step: calculate the centre vectors C (k) = [cj] with U (k) using equation (3).

3. Update U (k),U (k+1) using equation (4).

4. If || U (k+1) - U (k) ||< then STOP; otherwise return to 2

The main aim of FCM algorithm is to minimize the objective function given by [6]:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^m \left\| x_i - c_j \right\|^2 , \quad 1 \le m < \infty \qquad \text{...(ii)}$$

Where

m = any real number greater than 1

uij= membership degree of xi in Jth cluster xi = ith dimension of the measured data

ci = ith dimension of the cluster centre.

The formula for updating the cluster centre ci is:

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m} \qquad \text{...(iii)}$$

The values of partitions matrix are updated by using formula

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

When the maximum change in the values of fuzzy c partitions matrix is less than E the FCM algorithm iteration stops. Where E is termination criteria with value between 0 and 1. The "Bag of words" is a representation based on which the documents are clustered.

5. "BAG OF WORDS" REPRESENTATION

Every word contains some weight according to significance in the document which can be no of occurrence of word in that document.

Consider the following paragraph:

"It is a variation of the Hard C-Means clustering algorithm. Each observation here has a membership value associated with each of the clusters which is related inversely to the distance of that observation from the center of the cluster." The Bag of Words representation looks like this:

Table 1: Bag of Words

Word	Occurrences
Variation	1
Hard	1
Cluster	2
Observation	2
⋮	⋮
Centre	1

In this approach we have assigned the weights as the number of occurrences of a particular word in the document normalized by document length and multiplied by 10,000.

6. FUZZY LOGIC TO TEXT MINING

Following are the steps required for converting fuzzy logic to text mining

1. Text preprocessing

It involves cleaning up the text like removing advertisements , removing extra tags, hyphens, etc. Removal of this un-wanted text helps improve the efficiency of our algorithm.

2. Feature Generation

The documents are represented in the "Bag of Words" method. Stop words like "the", "a", "an" etc. are removed. Word Stemming [2] is carried out. Word Stemming refers to representing the word by its root, example the words - writing, wrote and written should be represented by write.

3. Feature Selection

The features to be used are selected. This selection of features can either be done before use or based on use. We choose only the features which will help us in our process [3][4][5] .

In order to cluster separately, the documents relating to sports and politics; presence of names of people in a document will not help much. However, presence of words like "ground" , "bat" ,"commentator" etc. will help relate the documents to the field of "sports". Similarly, words like "vote", "dictatorship", "president" will relate the documents to the category "politics".

7. CLUSTERING

We use FCM clustering algorithm. Each document to be clustered has already been represented as a "bag of words", and from that "bag" only the essential important "words" (features) have been assigned. The aim now is to cluster the documents into categories (called classes in FCM). Using FCM, the documents having similar frequencies (normalized by document length) of various selected features are clustered together.

Once the FCM execution finishes, because of max-change in

the Fuzzy Partition Matrix being lesser than the threshold, we get the required number of clusters and all the documents have been clustered among them.

8. EVALUATION AND INTERPRETATION OF RESULTS

From the Fuzzy logic Partition Matrix, one can find to what degree a given document contains belongs to each cluster. If the membership value of a document for a given cluster is high, the document can be said to strongly belong to that cluster. However, if all the membership values corresponding to a specific document are almost same (Example .35, .35, .30 among three clusters), it would imply that the document does not quite strongly belong to any of those clusters.

Example

Suppose we are given some documents which we have to cluster into two categories -"sports" and "politics". We will start from the third – feature selection, as first two steps are simple. We had no idea about which words (features) we could use to cluster our documents. Hence, we had to find such features.

We took some documents that we knew were related to Sports and Politics respectively. We applied step1 and step 2 on them and then counted the word frequency of various words using eq(i). Then we analyzed the differences in these WF values for same words between documents relating to Sports category and the documents relating to Politics category.

Table 2. Word Frequencies for Sports and Politics-related document

Word	WF (Sports)	WF (Politics)
Victory	10.0213	8.9012
Ground	203.2321	7.1214
Dictatorship	1.1213	140.1213
Bat	501.6553	30.2121
Group	250.6312	80.8452
Member	38.7658	40.2313
Publicity	8.8350	9.4213

V11 = (180+200+210+7)/4=149.25

Similarly, Calculating all V1j we get

V1= {149.25, 300, 162.5, 7.5} is the centre of cluster 1. Similarly, V2= {50,110.75, 67.75, 33.25} is the centre of

cluster 2.

By looking at the Table 2, it was apparent that words like "victory", "publicity" are used in similar amounts in documents belonging to both groups. However, words like "Ground" (203 v/s 7), "Bat" (501 v/s 30), "Group" (250 v/s 80) and "Dictatorship" (1 v/s 140) could be used to differentiate among documents belonging to Sports and Politics respectively. For the sake of simplicity and an easy example, we chose only 4 words – "Ground", "Bat", "Group" and "Dictatorship". On basis of these 4 words, we clustered the given documents into Politics and Sports types of categories.

Now we could start our clustering process.

Let D= {d1, d2, d3…dn} represent our "n" documents to be

clustered.

Now each of these documents, DI, is defined by the "M"selected features i.e. di= {di1, di2, di3...dim}

Now we have to initialize the fuzzy partition matrix. Keeping in mind the constraints that all entries should be in the interval [0, 1] and the columns should add to a total of 1 and considering that we are given 4 documents to cluster:

Table 3. Fuzzy Partition Matrix

	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Doc 7	Doc 8
Cluster 1	1	1	0	0	1	1	0	0
Cluster 2	0	0	1	1	0	0	1	1

Our initial assumption is that Doc1, Doc2, Doc5 and Doc6 belong to cluster 1 and Doc3, Doc4, Doc7 and Doc8 belong to cluster 2. Now we have to calculate the initial cluster centres.

For c=1, that is cluster1, the centre (V1) can be calculated using eq (iii). Since, the membership value of Doc3, Doc4, Doc7 and Doc8 in Cluster 1 is 0, we have

Now calculating the Euclidian distances of each Document vector from both centre clusters:

D11 = ((180-149.25)2 + (400-300)2 + (200-162.5)2 + (1-7.75)2)1/2 = 111.32

Similarly,

D12 = 149.5, D13 = 339.5, D14 = 352.5, D15 = 102.2, D16 = 353.4, D17 = 109.2, D18 = 351.1

Similarly, from the cluster two:

D21 = ((180-50) 2+ (400-110.75) 2+ (200-67.75) 2+ (1-33.25)
2) 1/2 = 345.10

D22 = 382.2, D23 = 105.1, D24 = 118.3, D25 = 334.4, D26 =

119.6, D27 = 339.7, D28 = 116.9

Now that we have distance of each vector from both cluster centres, we will now update the

Fuzzy Partition Matrix, by using eq(iv).

Note that we have already decided the value of the „m‟ in this formula, earlier in the example.

U11= [1+ (111.32 / 345.10)2]-1=.90, U12= .867, U23= .913, U24= .898, U15=.9146, U26=.897, U17=.906, U28=.901

Say our threshold change (stopping condition) was 0.001. Our max change here is 0.133 which is greater than 0.001. Hence we will continue. Now we will repeat the process by again calculating the new cluster centres. But now we will use the updated fuzzy partition matrix values. So, now our centre for cluster 1 will be calculated using of words like democracy (which are related to politics). Hence Cluster2 represents Politics.

Table 5. Updated fuzzy partition matrix after one iteration

	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Doc 7	Doc 8
Cluster 1	0.900	0.867	0.087	0.102	0.915	0.103	0.906	0.099
Cluster 2	0.100	0.133	0.913	0.898	0.085	0.897	0.094	0.901

V1j = (0.900x1j + 0.867 x2j + 0.087 x3j + 0.102 x4j + 0.915x5j + 0.103x6j + 0.906x7j + 0.099x8j)/ (0.9002 + 0.8672 + 0.0872 + 0.1022 + 0.9152 + 0.1032 + 0.9062 + 0.0992)

9. RESULT INTERPRETATION

Suppose that our final fuzzy partition matrix for 8 documents looks something like this.

Table 6. Fuzzy Partition Matrix

	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Doc 7	Doc 8
Cluster 1	0.890	0.804	0.149	0.155	0.865	0.159	0.810	0.168
Cluster 2	0.110	0.196	0.851	0.845	0.135	0.841	0.190	0.832

We see that Doc1, Doc2, Doc5, Doc7 belong to cluster 1 whereas Doc3, Doc4, Doc6, Doc8 belong to cluster 2 on basis of high membership values in those respective clusters.

In the example we took, the documents Doc1, Doc2, Doc5, and Doc7 had higher frequency of words like stadium, ball and team (which are related to sports). Since these documents have been clustered together, one can say that Cluster 1 is Sports. At the same time, Cluster2 will contain documents with higher frequency

Here, Doc1 relates to Sports to a large degree (due to its very high membership value). If we set the criteria that membership values greater than 0.85 with respect to a given cluster can be called "strong membership", then Doc1 and Doc5 can be said to "strongly" belong to Cluster1 which is sports. Moreover, we can say that Doc1 relates to Sports more "strongly" than Doc5 does. This interpretation of results in linguistic form [8] is what gives advantage to usage of Fuzzy Logic over Probability models like Bayesian in Text Mining.

10. CONCLUSION

This paper shows how one can use fuzzy logic in text mining to cluster documents by taking an example where the documents were clustered into two categories :- "sports" and "politics". The advantage of using fuzzy logic is that, it could calculate the degree to which a given document belonged to either categories- "sports" as well as "politics". By doing this for all documents in the data-set, we could alsocompare two documents and tell which one belongs "more"to which topic. Thus,for the scenario in the given example, Fuzzy C-means proves to be a better technique for document clustering.

11. REFERENCES

[1] Vishal Gupta, Gurpreet S. Lehal; "A Survey of Text Mining Techniques and Applications";Journal of Emerging Technologies in Web Intelligence, Vol.1, No.1, August 2009

[2] K.Sathiyakumari, V.Preamsudha, G.Manimekalai; "Unsupervised Approach for Document Clustering Using Modified Fuzzy C mean Algorithm";

International Journal of Computer & Organization Trends –Volume 11 Issue3-2011.

[3] R. Rajendra Prasath, Sudeshna Sarkar: Unsupervised Feature Generation using Knowledge Repositories for Effective Text Categorization. ECAI 2010: 1101-1102

[4] Sumit Goswami, Sudeshna Sarkar, Mayur Rustagi: Stylometric Analysis of Bloggers' Age and Gender. ICWSM 2009

[5] Sumit Goswami, Mayank Singh Shishodia; "A fuzzy based approach to stylometric analysis of blogger‟s age and gender"; HIS 2012: 47-51

[6] Ross, T. J. (2010); "Fuzzy Logic with Engineering Applications", Third Edition, John Wiley & Sons, Ltd, Chichester, UK

[7] Fuzzy logic vs Proobability (Good Math, BadMath);http://scientopia.org/blogs /goodmath/2011 /02/02 / fuzzy-logic-vs-probability/, last checked on 28thJuly 2012.

[8] Nogueira, T.M. ; " On The Use of Fuzzy Rules to Text Document Classification "; 2010 10th International Conference on Hybrid Intelligent Systems (HIS),; 23-25Aug 2010 Atlanta, US.