Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 4, April 2016, pg.114 – 117

Classification of Words: Using PFCM Clustering

Ritika Singhal, N. Deepika
CSE Dept, NHCE, VTU, India
CSE Dept, NHCE, VTU, India
rsinghal1987@yahoo.co.in; deepikajvijay@gmail.com

Abstract--- There are various clustering models introduced for unsupervised learning. PFCM or the possibilistic c-means model was proposed in 2005. PFCM produces mainly three values: the typicality values, membership values and the centres of the clusters. It is a hybrid model of PCM and FCM. We propose an extension to PFCM so that it can be used to cluster the text files.

Keywords— possibilistic model, fuzzy clustering, c-means clustering, preprocessing, Euclidean norm, Text classification

I. INTRODUCTION

Clustering or Classification of any data set is the process of partitioning the data set into subgroups. Let the number of data points in a data set X be n, then the number of subgroups c is such that $1 < c < n$. There are many clustering models used for classification of the data points. The output of the models is generally a $(c * n)$ matrix, known as the membership matrix M. Each element u_{ik} tells the membership of the data point in the particular cluster, where i ranges between 1 to c and k ranges between 1 to n. The clustering algorithms use a distance norm to calculate the membership values. Generally, the Euclidean Distance Norm is used. Although, several advancements have been done about which norm to use.

The most widely used clustering technique is the fuzzy c-means (FCM) [1]. The main constraint of FCM is that each row in membership matrix M must sum to 1. FCM produces the memberships of the data points that are related to the distance of that data point from the centers of the clusters. If a data point is equidistant from the clusters then it will have the same membership value in each cluster. The main problem with FCM is that the noise points or the outliers are also accounted in the membership values.

To prevent the noise points from being counted in, another clustering technique was introduced by Krishnapuram and Keller [2], named possibilistic c-means (PCM). PCM relaxes the row sum constraint of the FCM. The main constraint of the PCM model is that each membership value in M can be anything between 0 and 1 or equal to any one of them, i.e. $0 \leq u_{ik} \leq 1$. So these values were called the typicalities of the data points in each cluster. The problem with PCM is that sometimes it produces coincident clusters.

In 1997, Nikhil and James[4] proposed a fuzzy possibilistic c-means (FPCM) model and algorithm that generated both the membership and typicality values. FPCM used a scaling technique to overcome very small typicality values in case of large data

sets. To get a stronger candidate for fuzzy clustering, Nikhil and James proposed a possibilistic fuzzy c-means (PFCM) algorithm[3] in 2005. PFCM can avoid coincident clusters and also is less sensitive to outliers.

This paper is distributed in following sections. Section II discuses about the word classification. Section III gives the detailed proposed model. The experiment details are given in Section IV and the Section V gives the conclusions drawn.

II. WORDS CLASSIFICATION

Text or words are the basic components of any sort of communication or documentation. These words in any single document sometimes need to be grouped too. Words classification deals around this problem. It basically means that the words in the documents are grouped in some clusters according to some criteria.

This sort of classification can be used in many aspects. Some uses of such word-classification is:

- It can be used to detect any fraud data.
- It can also be used to fetch keywords from a huge document.
- It can further be used for preparing word frequency reports.

All these uses hint that word classification can be helpful in many scenarios. Till now, the Bag-of-Words model[5] was being used for words classification and preprocessing. This paper uses PFCM model to classify words in text files.

III. PROPOSED MODEL

The model proposed by us uses an implementation of the PFCM model. A text file is taken as an input. The output is the centers of the clusters that gives the groups of the words. Also the membership and typicality matrix are output too. The model proposed in this paper goes through following steps, which are also depicted in Fig. 1.

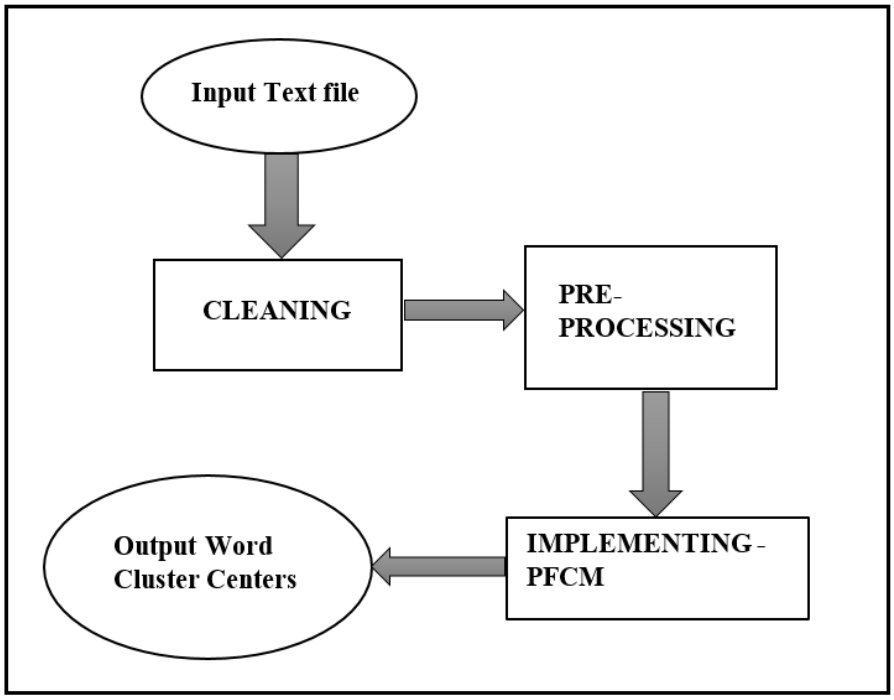

Fig 1. The Proposed Model Workflow

A. Text Cleaning

This step involves cleaning the input text file. Basically this revolves around removing the common words from the input document. This step is crucial because the common words like “if”, “hi”, “yes”, “we” etc. can create noise points for the clustering algorithm. This step outputs a clean document, which mainly has words of user’s interest.

B. Pre-Processing

This step involves using the conversion of words into the data points. This step is also mandatory as the input to PFCM model are the data points and not the textual words. The method to convert word to the data points can be invented according to the experimental needs. The idea used by us is to incorporate the frequency of each word in the mapping of it to a data point. The term frequency is used because the number of times a word appears in the document is important factor to be considered.

C. Implementing PFCM

When the words are converted to the data points and the input is ready, then PFCM model is implemented on the data points. The output of PFCM are the cluster centers. The key idea here is that each data point represents a word. So the clusters formed by PFCM gives the groups or classification of the words. The membership matrix and the typicality matrix are also output that can be further used for analyzation.

As proposed by Nikhil and James[3] in 2005, the primary optimization problem of PFCM is given as:

$$J_{m,\eta}(M,T,V;X) = \sum_{k=1}^{n}\sum_{i=1}^{c}(au_{ik}^{m} + bt_{ik}^{n}) \times \left||x_k - v_i|\right|_A^2 + \sum_{i=1}^{c}\gamma_i \sum_{k=1}^{n}(1-t_{ik})^{\eta}$$

Here the problem is to minimize J. Also, a, b, γ, η are the user-constants. The minimization problem can be achieved by calculating the Membership matrix M, the typicality matrix T, and the centroids matrix V, using the equations given in Fig. 2.

$$u_{ik} = \left(\sum_{j=1}^{c}\left(\frac{D_{ikA}}{D_{jkA}}\right)^{2/(m-1)}\right)^{-1}$$
$$1 \le i \le c; \;\; 1 \le k \le n$$
$$t_{ik} = \frac{1}{1 + \left(\frac{b}{\gamma_i} D_{ikA}^2\right)^{1/(\eta-1)}}$$
$$1 \le i \le c; \;\; 1 \le k \le n$$
$$\mathbf{v}_i = \frac{\sum_{k=1}^{n}(au_{ik}^{m} + bt_{ik}^{\eta})\,\mathbf{x}_k}{\sum_{k=1}^{n}(au_{ik}^{m} + bt_{ik}^{\eta})}$$
$$1 \le i \le c.$$

Fig 2. Equations for calculation of matrices in PFCM

The parameter D_{ikA} gives the Euclidean Norm here. It represents the distance between the data point at i and k. When these equations are used and the matrices are populated, then the output centers are quite close to the ideal ones. The respective matrices give the participation and the typicality of a word (or a data point) in each cluster. The output can then be manipulated and displayed as per user needs. Reports can be generated that can display in a user-friendly view about what are the findings of the algorithm run.

IV. EXPERIMENTS DONE

The actual experiments of the proposed model were done in Visual Studio. Net. The outputs were similar to expected values when PFCM was implemented. The input text file was taken to be a user-generated file which contained many words with different frequencies. The output clusters separated the important and the un-important words.

V. CONCLUSIONS

The conclusion drawn is that PFCM is a better candidate for fuzzy clustering. This clustering when used for text files, requires an exclusive preprocessing step. If that step is properly implemented then PFCM clusters the words in a text file very accurately. The clusters can depict some useful information too.

ACKNOWLEDGEMENT

The authors wish to acknowledge all the readers and the reviewers for this paper. This paper is written as a part of research on fuzzy clustering techniques. The authors also wish to acknowledge the department and the university who offered us this academic research and all other facilities.

REFERENCES

[1] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms.* New York: Plenum, 1981.
[2] R. Krishnapuram and J. Keller, "A possibilistic approach to clustering," IEEE Trans. Fuzzy Syst., vol. 1, no. 2, pp. 98-110, Apr. 1993.
[3] Nikhil R. Pal, Kuhu Pal, James M. Keller, and James C. Bezdek, *"A possibilistic fuzzy c-means Clustering Algorithm",* IEEE Trans. Fuzzy Systems., vol. 13, no. 4, Aug. 2005.
[4] N. R. Pal, K. Pal, and J. C. Bezdek, "A mixed c-means clustering model," in IEEE Int. Conf. Fuzzy Systems, Spain, 1997, pp. 11-21.
[5] https://en.wikipedia.org/wiki/Bag-of-words_model

