TEXT AND IMAGE CLASSIFICATION USING FUZZY SIMILARITY BASED SELF CONSTRUCTING ALGORITHM

Dinesh Kavuri [1], Pallikonda Anil Kumar [2], Doddapaneni Venkata Subba Rao [3]

[1] *Student, in CSE Dept of PVP Siddhartha Institute of Technology, Kanuru, Vijayawada.*
dinesh.kavuri@gmail.com
[2] *Asst. Professor, in CSE Dept of PVP Siddhartha Institute of Technology, Kanuru, Vijayawada.*
anilkumar_pallikonda@yahoo.co.in
[3] *Assoc. Professor, in CSE Dept of SRK Institute of Technology, Vijayawada.*
doddapanenivenkat@gmail.com

Abstract

In this new and current era of technology, advancements and techniques, efficient and effective text document classification and image segmentation is becoming a challenging and highly required area to capably categorize text documents and images into mutually exclusive categories. Fuzzy similarity provides a way to find the similarity of features among various documents. For text document classification, the words in the feature vector of a document set are grouped into clusters, based on similarity test. Each cluster is characterized by a membership function with statistical mean and deviation. We then have one extracted feature for each cluster and membership functions are derived to match closely with and describe properly the real distribution of the training data. For image segmentation, an improved fuzzy c-means (IFCM) clustering algorithm is presented. The originality of this algorithm is based on the fact that the conventional FCM-based algorithm considers no spatial context information, which makes it sensitive to noise. The new algorithm is formulated by incorporating the spatial neighbourhood information into the original FCM algorithm by a priori probability.

Index Terms: *Classification, Image Segmentation, Clustering, Fuzzy Similarity, Priori Probability etc.*

---------- *** ----------

I. INTRODUCTION

In text classification, the dimensionality of the feature vector is usually huge. Such high dimensionality can be a severe obstacle for classification algorithms. To alleviate this difficulty, feature reduction approaches are applied before document classification tasks are performed. Two major approaches, feature selection and feature extraction have been proposed for feature reduction. In general, feature extraction approaches are more effective than feature selection techniques, but are more computationally expensive. Therefore, developing scalable and efficient feature extraction algorithms is highly demanded for dealing with high-dimensional document data sets.

Image segmentation is a key step toward image analysis and serves in the variety of applications including pattern recognition, object detection, and medical imaging. The task of image segmentation can be stated as the partition of an image into different meaningful regions with homogeneous characteristics using discontinuities or similarities of the image such as intensity, colour, tone or texture, and so on. Among the fuzzy clustering methods, fuzzy c-means (FCM) algorithm is the most popular method used in image segmentation because it has robust characteristics for ambiguity and can retain much more information than hard segmentation methods. Although the conventional FCM algorithm works well on most noise-free images, it has a serious limitation: it does not incorporate any information about spatial context, which cause it to be sensitive to noise and imaging artifacts.

In this paper we propose a fuzzy similarity-based self constructing feature clustering algorithm, which is an incremental feature clustering approach to reduce the number of features for the text classification task and a novel improved FCM (IFCM) clustering algorithm for image segmentation is proposed. In fuzzy similarity-based self constructing feature clustering algorithm, Words that are similar to each other are grouped into the same cluster. Each cluster is characterized by a membership function with statistical mean and deviation. If a word is not similar to any existing cluster, a new cluster is

created for this word. When all the words have been fed in, a desired number of clusters are formed automatically. We then have one extracted feature for each cluster. The extracted feature corresponding to a cluster is a weighted combination of the words contained in the cluster. For image segmentation, improved FCM (IFCM) clustering algorithm is developed by incorporating the spatial neighbourhood information into the standard FCM clustering algorithm by a priori probability. The probability is given to indicate the spatial influence of the neighbouring pixels on the centre pixel. The new fuzzy membership of the current centre pixel is then recalculated with this probability obtained from above. The algorithm is initialized by a given histogram based FCM algorithm, which helps to speed up the convergence of the algorithm.

II. REALTED WORK

Over the last decades, fuzzy similarity based text document classification has got attention very much and considered as an important research area. Different techniques, models and ways are searched to design a best categorization system. Such field is not only used in the small level organizations, industries and corporate, but also covers a vast community all around the world.

The new techniques, their collaboration and research always open a new paradigm towards the Advancements. Current research studies show that fuzzy logic and its area of concerns provide efficient base for text categorization, dimensionality reduction, feature selection and extraction, and similarity analyzer related issues. Fuzzy logic is considered as a branch of logic especially designed for representing knowledge and human reasoning in such a way that it is amenable to processing by a computer. The major concepts of fuzzy logic are fuzzy sets, linguistic variable, possibility distributions, and fuzzy if – then rules. Fuzziness or Degree of Uncertainty pertains to the uncertainty associated with a system, i.e., the fact that nothing can be predicted with exact precision. Practically, the values of variables are not always precise; rather approximate values are more likely to be known. The vagueness can adequately be handled using fuzzy set theory.
This theory provides a strict mathematical framework using which vague conceptual phenomena can be studied rigorously. It is also called the property of language. Its main source is the imprecision involved in defining and using symbols. It is a property of models, computational procedures, and languages. Hence, a fuzzy set is a collection of distinct elements with a varying degree of relevance or inclusion.

Feature Clustering

The concept of feature clustering enhances the provision of text dimension criticality solution. It is an efficient way to compress the collected feature sets more, so that the resultant data can be handled and used properly without any loss. These clusters are represented either by the term of maximum frequency in a group (or cluster) or can be found by self constructing feature clustering algorithm. Feature clustering is also done with the use of the pseudo-thesaurus by identifying each term [6] as noun, pronoun, adverb, adjective, delimiters etc. Researchers have shown that it helps to reduce the high dimensional data into smaller one adequately.

Fuzzy Association

Fuzzy sets pay an important and vital role in text categorization. They are widely recognized as many real world relations are intrinsically fuzzy. Fuzzy association is used to discover important associations between different sets of attribute values.

Fuzzy Production Rules

The novel method of rule-base construction and a rule weighting mechanism can result in a rule-base containing rules of different lengths, which is much more useful when dealing with high dimensional data sets.

Fuzzy Clustering and C-Means

In fuzzy clustering, each point has a degree of belonging to clusters, as in fuzzy logic, rather than belonging completely to one cluster. Thus, points on the edge of a cluster may be in the cluster to a lesser degree than points in the centre of cluster.

Fuzzy Signatures

Fuzzy signatures are used in those applications and key areas which require the handling of complex structured data and interdependent feature problems. They can also used in special concerns where data is missing. So, this depicts many areas where objects with very complex and sometimes interdependent features are to be classified along with the evaluation of similarities and dissimilarities. This leads a complex decision model hard to construct effectively. Due to the very nature of fuzzy signatures of flexibility, it can be used for many text mining tasks, with the benefit of the hierarchical structuring; therefore, the text document classification models can be constructed.

III PROPOSED METHOD

A fuzzy self-constructing feature clustering (FFC) algorithm for Text Document:

A fuzzy self-constructing feature clustering (FFC) algorithm which is an incremental clustering approach to reduce the dimensionality of the features in text classification.

Feature clustering is an efficient approach for feature reduction, which groups all features into some clusters, where features in a cluster are similar to each other. The feature clustering methods proposed in are "hard" features belongs to exactly one word cluster. Therefore each word contributes to the synthesis of only one new feature. Each new feature is obtained by summing up the words belonging to one cluster. The matrix consisting of all the original documents with m features and the matrix consisting of the converted documents with new k features. The new feature set corresponds to a partition clustering methods, where each word of the original features belongs to exactly one word cluster. Therefore each word contributes to the synthesis of only one new feature. Each new feature is obtained by summing up the words belonging to one cluster. Each cluster is characterized by a membership function with statistical mean and deviation. If a word is not similar to any existing cluster, a new cluster is created for this word. Similarity between a word and a cluster is defined by considering both the mean and the variance of the cluster. When all the words have been fed in, a desired number of clusters are formed automatically. We then have one extracted feature for each cluster. The extracted feature corresponding to a cluster is a weighted combination of the words contained in the cluster. Three ways of weighting, hard, soft, and mixed, are introduced. By this algorithm, the derived membership functions match closely with and describe properly the real distribution of the training data.

Improve fuzzy c-means algorithm for Image segmentation:

The objective function of the traditional FCM algorithm does not take into account any spatial information, which means the clustering process is related only to gray levels independently of the pixels of image in segmentation. However, according to the theory of Markov random field (MRF) that most pixels belong to the same class as their neighbours, which means the class probability of a pixel depends on class memberships of its (spatial) neighbour clusters, in this way it can reduce the possible influence of noise and overlapping clusters. Therefore, the limitation of the standard FCM algorithm makes it very sensitive to noise. The general principle of the technique presented in this paper is to incorporate the neighbourhood information into the FCM algorithm during classification. Here, the shape of the neighbourhood is selected as an 8-neighborhood. Therefore, the a priori probability, pik is determined and updated during the clustering by converting the fuzzy partition matrix to a crisp partition matrix in this 8-neighborhood.

In this paper, the defuzzification is carried out with the maximum membership procedure. To prevent that our algorithm gets trapped in a local minima, the IFCM algorithm is initialized with the fast FCM algorithm. Once the fast FCM is stopped, the IFCM algorithm continues with the values for the prototypes and membership values obtained from the fast FCM algorithm. The IFCM algorithm then iteratively updates its a priori probability, membership and centroids with these values. When the IFCM algorithm has converged, another defuzzification process takes place in order to convert the fuzzy partition matrix to a crisp partition matrix that is segmentation. Thus the IFCM algorithm is presented as follows:

Notations: uik is the degree of membership of xk in the i-th cluster, c is the number of clusters, q is a weighting exponent on each fuzzy membership, vi is the prototype of the centre of cluster i, d2(xk, vi) is a distance measure between object xk and cluster centre vi, g is gray level.

Step 1: Set the cluster centroids vi according to the histogram of the image, fuzzification parameter q, the values of c and ε > 0.
Step 2: Compute the membership function using

$$u_{ig}^{(b)} = \frac{1}{\sum_{j=1}^{c}\left[\frac{d(g, v_i)}{d(g, v_j)}\right]^{2/(q-1)}}, \forall i, g$$

Step 3: Compute the cluster centroids using

$$v_i^{(b+1)} = \frac{\sum_{g=L_{\min}}^{L_{\max}} \left(u_{ig}^{(b)}\right)^q \mathrm{His}(g) g}{\sum_{g=L_{\min}}^{L_{\max}} \left(u_{ig}^{(b)}\right)^q \mathrm{His}(g)}, \forall i$$

$$\mathrm{His}(g) = \sum_{s=0}^{S-1}\sum_{t=0}^{T-1} \delta\left(f(s,t) - g\right)$$

Step 4: Go to step 2 and repeat until convergence.
Step 5: Compute the *a priori* probability using

$$p_{ik} = \frac{\#N_k^i}{\#N_k}$$

with the obtained results of membership function and centroids.

Step 6: Recompute the membership function and cluster centroids using

$$u_{ik}^{*(b)} = \frac{p_{ik}}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}} \right)^{2/(q-1)}}$$

and

$$v_i^{*(b+1)} = \frac{\sum_{k=1}^{n} \left(u_{ik}^{*(b)} \right)^q x_k}{\sum_{k=1}^{n} \left(u_{ik}^{*(b)} \right)^q}$$

with the probabilities.

Step 7: If the algorithm is convergent, go to step 8; otherwise, go to step 5.

Step 8: Image segmentation after defuzzification using

$$C_k = \arg_i \{ \max(u_{ik}) \}, \quad i = 1, 2, \ldots, c$$

and then a region labelling procedure is performed.

IV PERFORMANCE

It is important to note that the proposed method performs the best for the segmentation with more homogeneous regions and with least spurious components and noises particularly. The segmentation accuracy of applying these algorithms to the images with different levels of noise. Here, the segmentation accuracy (SA) is defined as:

SA= (Number of correctly classified pixels/ Total number of pixels)* 100%

Figure 1: (a) original image

Figure 1: (b) FCM image

Figure 1: (c) IFCM image

It can be seen that as the noise level increases, the SA of FCM degrades rapidly. The proposed IFCM algorithms can all handle the problem caused by noise and can get higher SA even under the noise of 7%. Overall, the KFC and IFCM algorithms produce almost identical results, which are a little better than that of the SFCM algorithm. However, from the above analysis, it should be noted that the KFC algorithm needs more computational time than IFCM algorithm. The proposed IFCM algorithm gets the highest uniformity value, while the original FCM algorithm gets the lowest uniformity value.

CONCLUSION

In this paper we propose a fuzzy similarity-based self constructing feature clustering algorithm, which is an incremental feature clustering approach to reduce the number of features for the text classification task and a novel improved FCM (IFCM) clustering algorithm for image segmentation is proposed. In fuzzy similarity-based self constructing feature

clustering algorithm, Words that are similar to each other are grouped into the same cluster. Each cluster is characterized by a membership function with statistical mean and deviation. If a word is not similar to any existing cluster, a new cluster is created for this word. When all the words have been fed in, a desired number of clusters are formed automatically. We then have one extracted feature for each cluster. The extracted feature corresponding to a cluster is a weighted combination of the words contained in the cluster. For image segmentation, improved FCM (IFCM) clustering algorithm is developed by incorporating the spatial neighbourhood information into the standard FCM clustering algorithm by a priori probability. The probability is given to indicate the spatial influence of the neighbouring pixels on the centre pixel. The new fuzzy membership of the current centre pixel is then recalculated with this probability obtained from above. The algorithm is initialized by a given histogram based FCM algorithm, which helps to speed up the convergence of the algorithm.

REFERENCES

[1] KIM J., FISHER J.W., YEZZI A., CETIN M., WILLSKY A.S., A nonparametric statistical method for image segmentation using information theory and curve evolution, IEEE Transactions on Image Processing 14(10), 2005, pp. 1486–1502.
[2] DONG G., XIE M., Color clustering and learning for image segmentation based on neural networks,
IEEE Transactions on Neural Networks 16(4), 2005, pp. 925–936.
[3] HARALICK R.M., SHAPIRO L.G., Image segmentation techniques, Computer Vision, Graphics and Image Processing 29(1), 1985, pp. 100–132.
[4] PAL N.R., PAL S.K., A review on image segmentation techniques, Pattern Recognition 26(9), 1993, pp. 1277–1294.
[5] H. Kim, P. Howland, and H. Park, “Dimension Reduction in Text Classification with Support Vector Machines,” J. Machine Learning Research, vol. 6, pp. 37-53, 2005.
[6] F. Sebastiani, “Machine Learning in Automated Text Categorization,” ACM Computing Surveys, vol. 34, no. 1, pp. 1-47, 2002.
[7] BEZDEK J.C., Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York 1981.
[8] BEZDEK J.C., HALL L.O., CLARKE L.P., Review of MR image segmentation techniques using pattern recognition, Medical Physics 20(4), 1993, pp. 1033–1048.

BIOGRAPHIES

Dinesh Kavuri received the B.tech (IT) Degree from Nalanda Institute of Engg & Tech College, Kantepudi and affiliated to JNTU Kakinada University, India. He is currently pursuing M.Tech (CSE) Degree at the Dept of Computer Science Engineering, PVP Siddhartha Institute of Technology, Vijayawada and affiliated to JNTU Kakinada University, India.

Pallikonda Anil Kumar received the B.tech (IT) Degree from Gudlavalleru Engineering College, Gudlavalleru, Krishna District and affiliated to JNTU Kakinada University, India. He received the M.Tech (Software Engineering) Degree from Avanthi Institute of Engg & Tech, Narsipatnam, Vizag. He has more than 6 years teaching experience. He is currently working as an Asst Professor in the Dept of Computer Science Engineering, PVP Siddhartha Institute of Technology, Vijayawada and affiliated to JNTU Kakinada University, India. His interests are Sofware Engineering, Data Mining.

Doddapaneni V Subba Rao received the B.Tech (Civil) Degree from KSRM College of Engg and affiliated to SVU University, Tirupathi, India. He received M.C.A from Indira Gandhi National Open University, Guntur. He also received M.Tech. in Computer Science Engineering from JNTU University, Kakinada, India. He is currently pursuing Ph.D. on Image Processing from JNT University, Kakinada, India. He has 22 years of teaching experience. He is presently working as an Associate Professor in the Dept of Computer Science Engineering, SRK Institute of Technology; Vijayawada (affiliated to JNTU Kakinada University, Kakinada, India).