FRC: A New Dynamic Fuzzy Clustering Alogorithm Based on Random Forest Algorithm

Xinqing Geng

***Abstract*—A novel approach is presented in the paper. The main defect of traditional methods of fuzzy clustering is to know the number of clustering in advance. The text eigenvector is acquired based on the vector space model (VSM) and TF.IDF method. This paper applies information entropy to FRC algorithm. Random forest fulfils reduce dimension and initial classification result, which is used as input of FCM algorithm. In the iterative process, the number of clustering is ascertained until inforamtion entropy is minimum value. Due to FRC algorithm don't need reduce dimension, the present algorithm possess higher precision and efficiency. The fuzzy clustering algorithm is suitable for dealing with the semantic variety and complexity. The example demonstrates the effectiveness of the present algorithm.**

***Index Terms*—Fuzzy clustering, random forest, ext mining.**

I. Introduction

Recently a lot of data are obtained in industry, agriculture, medical, education areas and so on. KDD technology develops gradually with database's. Data mining is key step of KDD. Data mining extracts patterns from large data such that they are certain, previously unkown, and potentially useful for the specific application [1]-[3].

Data mining includes two learning algorithm, supervised learning and unsupervised learning [4], [5]. Clustering is unsupervised learning which characterizes the datasets into subparts based on observation. Datapoints which belong to the same clusters share common property. Most of the times distance measure is used for deciding the membership of the clusters [6]-[9].

Classification learning based on text mining usually generates a dataset from the texts and applies learning methods, such as decision tree and SVM, to generated data. However, in ordinary text data analytics, the performance of such methods is not good. because selection of keywords depends on frequencies of keywords, and connection of keywords and targests cannot be captured by frequencies [10]-[16].

The performance of k-mean clustering is affected by initial cluster center and number of cluster centroid. Zhang Chen et.al has proposed a new concept for selecting the number of clustering. Mark Junjie Li troids *et al.* has proposed an extension to the standard fuzzy K-Means algorithm by introducing a penalty term to the objective function to make the clustering process not sensitive to the initial cluster centers. Mrutyunjaya Panda *et al.* has used k-mean and fuzzy k-mean for intrusion detection. Sometimes k-mean clustering does not gives best results for large datasets. So for removing this problem, Yu Guan *et al*. have introduced a new method Y- mean which is variation of k-mean clustering it removes the dependency and degeneracy problem of k-mean clustering. Sometime single clustering algorithm does not gives best result so for removing this problem, Fangfei Weng *et al.* has used k-mean clustering with new concepts which is called Ensemble K-mean clustering. Cuixiao Zhang *et al.* have used K-M clustering for intrusion detection. Some of the authors have used k-mean clustering along with the other method for improving the detection rate of intrusion detection system. Authors have used k-mean clustering along with the other data mining techniques for intrusion detection [17], [18].

Text clustering has been widely discussed for many years so far, which is application of data mining. Text clustering algorithm comprises of pattition-based algorithm, hierarchical algorithm, density-based algorithm, grid-based algorithm, and model-based algorithm. Hard clustering algorithm can't express semantic relation of texts. In real world, Fuzzy clustering is more suitable for sematic expression [19], [20].

Random forest is a group of decision trees, each one corresponds to a training data. Random forest is high accuracy of prediction and clustering and good tolerance of outlier and noise, with which over-fitting is not easy to occur [21].

RF is the paradigm of Bagging algorithm in ensemble learning. In 2001, Breiman combined decision trees into a forest. Features (columns) and data (rows) are randomly sampled to generate multiple decision trees and their predictions are aggregated. RF improves the prediction accuracy without significant increase in computation, and it is insensitive to multicollinearity. Its performance is robust against missing and imbalanced data, and it can well measure the roles of thousands of variables. Compared with models in deep learning, it has the advantage of small computation, outstanding generalization ability and strong interpretability, which makes it ever appealing to researchers and practitioners in spite of its long existence [22], [23].

However, there are also limitations of RF. First of all, the RF model can only be extended horizontally (more decision trees) but not vertically since the decision trees exist in parallel and cannot be stacked in layers in the same fashion as neurons in neural networks. Secondly, these decision trees have the same weight in voting for the final prediction despite that some of these trees may perform poorly. Lastly, all points in training data have the same weight and are treated equally in the sampling and training process, despite that some of the data are easy to classify while others are hard.

Manuscript received September 10, 2018; revised November 3, 2018.

Xinqing Geng is with the College of Mathematics and Information Science, Anshan Normal University, Anshan, China (e-mail: gengxinqing@163.com).

DOI: 10.7763/IJCTE.2018.V10.1224

FRC algorithm is presented based on random forest algorithm in this paper. Random forest algorithm maps high-dimensional space to low-dimensional space. FRFC algorithm is united with random forest and fuzzy clustering, which overcomes traditional fuzzy clustering base on partition is to know the number of clustering in advance. Random forest algorithm engengders many leaf node, which is used as input of FCM algorithm. The result of clustering is acquired with FCM algorithm.

II. Character Presetation of Texts

Vector Space Model (VSM) is wide used in character representation of texts. In the model, the space of texts is considered as vectors space of orthogonal vectors [24]-[29]. Each text means one of feature vector

$$V(d) = (t_1, w_1(d); \cdots; t_i, w_i(d); \cdots; t_n, w_n(d))$$

where t_i is term, $w_i(d)$ is the weight in d. $w_i(d)$ is function of appearing frequency $tf_i(d)$ of t_i in d, namely, $w_i(d) = \psi(tf_i(d))$.

TF·IDF is a $\psi(tf_i(d))$ method of determining the weights of the terms. The formula is defined as

$$\psi = tf_i(d) \times \log(\frac{N}{n_i}) \tag{1}$$

where N is the number of texts, n_i is the number of t_i in texts.

III. FRC Model

A. Random Forest Algorithm

Random forest algorithm is to obtain k decision trees, which constructs k random vector $\theta_1, \theta_2, \ldots, \theta_k$ [30]-[32].

Algorithm:

Step 1: N is the number of training cases, M is the number of variable in classifier.

Step 2: m variables are selected at random out of the M and the best split on these m is used to split the node where m<<M. The value of m is held constant during the forest growing.

Step 3: Choose a training set for this tree by choosing N times with replacement from all N available training cases. Use the rest of the cases to estimate the error of the tree, by predicting their classes.

Step 4: For each node of the tree, randomly choose m variables on which to base the decision at that node. Calculate the best split based on these m variables in the training set.

Step 5: Each tree is grown to the largest extent possible. There is no pruning.

B. FCM Algorithm

FCM algorithm is that datapoint belongs to every cluster with some membership [33]-[35].

It is based on minimization of the following objective function:

$$J = \sum_{k=1}^{n} \sum_{j=1}^{k} [u_{ij}]^{\beta} \left\| x_i - c_j \right\|^2 \tag{2}$$

where m is any real number is the degree of membership of x_i in the cluster j, c_j is the d-dimension centre of the cluster.

Algorithm is as follow:

Input: Set of data points, number of cluster

Output: Set of datapoints in form of cluster along with their membership

Step1: Initialize membership of data points based upon the initial centroid $U=[u_{ij}]$ matrix $U^{(0)}$.

Setp2: Calculate the centers vectors $C^{(k)}=[c_j]$ with $U^{(k)}$

$$c_j = \frac{1}{\sum_{i=1}^{n} u_{ij}^m} \sum_{i=1}^{n} [u_{ij}]^{\beta} x_{ij} \tag{3}$$

Step3: Update $U^{(k)}$,the formula of $U^{(k+1)}$is as follow:

$$u_{ij} = \frac{[\frac{1}{\left| x_i - c_j \right|^2}]^{1/\beta-1}}{\sum_{j=1}^{k} [\frac{1}{\left| x_i - v_j \right|^2}]^{1/\beta-1}} \tag{4}$$

When max $\{|u_i(x^{k+1})-u_i(x^k)|\}<\varepsilon, 0 \le \varepsilon \le 1$,the iteration will stop. Where k is the iteration steps.

C. Determination of Number of Clustering

Entropy is used to describe disordered distribution atoms. The distribution of dataset is similar to atoms'. The more rational the partition of clustering is, the more positive which cluster the data points belongs to is. he average information entropy serves as the standard for the number of clustering. Suppose initial cluster k=2 ，the number of clustering is updated during iterative process. When the average information entropy reaches minimum value, correspond number of clustering is the best. The definition of average entropy is as follow:

$$H = -\sum_{j=1}^{C} \sum_{i=1}^{N} \{[u_{ij} \times \log_2(u_{ij}) + (1-u_{ij}) \times \log_2(1-u_{ij})]/N \tag{5}$$

where u_{ij} is membership of sample data j belonging to class i, the number of clustering with minimum value of H is the best number of clustering.

D. FRC Algorithm

Step 1: Suppose initial number of clustering k=2, teration b=0, exponential weight m, iterative stop threshold ε.

Step2: Sample vectors are input of random forest algorithm. The codes between root node and leaft node are 0-1 sequence.

Step 3: The code matrix which the decision tree constructs is used as inut of FCM algorithm.

Step 4: Compute $H^{(b)}$ according to formula(5) after the iteration of FCM algorithm terminates. If $H^{(b+1)}<H^{(b)}$,the number of clustering c=c+1 and turn to step 2;else When H(x) achieves minimum value, the number of clustering c=c+1,clustering process ends.

IV. The Application of FRC Model in Text Clustering

FRC algorithm don't need reduce dimension. The texts are segmented in the preprocess phase, and VSM model is adopted. Random forest algorithm achieves reducing dimension and the initial classification result are gained. The code matrix random forest algorithm engenders is used

as input of FCM algorithm. When FRC algortihm running is finished, the number of clustering is determined, which is applied to text mining.

V. Experiment

We collect 3000 documents that are obtained from www.nlp.org.cn, Chinese natural language open platform. The text vectors are2160 dimensionality after segmentation. Agriculture, economy, education, industry are selected. Each class is 750 pieces of texts. The right number of clustering is 4.

The initial number of clustering is 2, m is 2, ε=0.01, the right number of clustering is 5.

To demonstrate the superiority of FRC algorithm, FRC algorithm makes a contrast experiment with FCM algorithm. In the same experimental condition, FRC algorithm needs 3 training cycles and execution time is 2 minutes; FCM need 5 training cycles and execution time is 10 minutes. So the efficiency of FRC algorithm is higher than FCM algorithm.

TABLE I: Clustering Retult of FRC Algorithm

	agriculture	economy	education	industry
Wrong statistical number	18	13	25	16
Right statistical number	732	737	725	734
Total number of segmented texts	752	759	738	741
Precision%	97%	97%	98%	99%
recall%	98%	98%	97%	98%

TABLE II: Clustering Retult of FCM Algorithm

	Agriculture	Economy	education	industry
Wrong statistical number	29	37	48	31
Right statistical number	721	713	702	719
Total number of segmented texts	750	760	732	740
Precision%	96%	94%	96%	97%
recall%	96%	95%	94%	96%

Compare Table I with Table II, the precision of FRC algorithm is higher than that of FCM in documentation. Inforamtion entropy is used to determine the number of clustering. FRC algortihm obtain the number of clustering during dynamic iteration. Because random forest algorithm don't need reduce dimension, FRC algorithm improves the precision and convergent rate.

References

[1] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," *ACM Sigmod Record*, vol. 22, pp. 207-216, 1993.

[2] L. Breiman, "Stacked regressions," *Machine Learning*, vol. 24, no. 1, pp. 49-64, 1996.

[3] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[4] L. Breiman, "Classification and regression trees," *Routledge*, 2017.

[5] P. H. Giang, *A Machine Learning Approach to Create Blocking Criteria for Record Linkage*, Springer Science &Business Media New York, 2014.

[6] Z. Chen and X. Shixiong, "K-means clustering algorithm with improved initial center," in *Proc. Second International Workshop on Knowledge Discovery and Data Mining*, 2009.

[7] M. J. Li, K. N. Michael, Y. Cheung, and J. Z. Huang, "Agglomerative fuzzy k-means clustering algorithm with selection of number of clusters," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 11, 2008.

[8] M. Panda and M. R. Patra, "Some clustering intrusion detection system," *Journal of Theoretical and Applied Technology*, pp. 710-716, 2005-2008.

[9] F. Weng, Q. Jiang, L. Shi, and N. Wu, "An intrusion detection system based on the clustering ensemble," in *Proc. IEEE International Workshop on 16-18 April 2007*, p. 12.

[10] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 6, pp. 610-621, 1973.

[11] O. Chqelle, P. Hafiier, and V. N. Vapnik, "Support vector machines for histogram-based image classiflcation," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1055-1064, 1999.

[12] A. Bosch, A. Zisserman, and X. Muoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 712-727, 2008.

[13] R. E. Banfield, L. O. Hall, and K. W. Bowyer, "A comparison of decision tree ensemble creation tcchniques," I*EEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 173-180, 2007.

[14] H. Tamura, S. Mori, and T. Yamawaki, "Textural features corresponding to visual perception," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 8, no. 6, pp. 460-473, 1978.

[15] Y. Zhang, P. Jin, and R. Sun, "Chinese word sense induction based on improved k-means algorithm," *Journal of Computer Applications*, vol. 32, no. 5, pp. 1332-1334, 2012.

[16] H. Suo and Y. Wang, "An improved k-means algorithm for document clustering," *Journal of Shandong University*, vol. 43, no. 1, pp. 60-64, 2008.

[17] D. Amorim, B. Mirkin, and M. Metric, "Feature weighting and anomalous cluster initializing in k-means clustering," *Pattern Recognition*, vol. 45, no. 3, pp. 1061-1075, 2012.

[18] Q. Wu, X. Cai, and M. Cai, "A study of weighting exponent in expressway traffic state estimation based on fuzzy c-means," *Science Technology and Engineering*, vol. 17, no. 6, pp. 289-295, 2017.

[19] F. Zeng and X. Zhang, "Application of cluster analysis to preventive maintenance scheme design of pavement," *Journal of Harbin Institute of Technology*, vol. 16, no. 4, pp. 581-586, 2009.

[20] B. Krause, C. Altrock, and M. Pozybill, "Intelligent highway by fuzzy logic: Congestion detection and traffic control on multi-lane roads with variable road signs," in *Proc. of the 1996 5th IEEE International Conference on Fuzzy Systems, New Orleans*, 1996, pp. 1832-1837.

[21] S. Dong and Z. Huang, "A brief theoretical overview of Random Forests," *Journal of Integration Technology*, vol. 2, no. 1, pp. 1-7, 2013.

[22] L. Gao, Y. Liu, and Z. Sheng, "Application of random forest algorithm to traffic state identification," *Experiment Technology and Management*, vol. 34, no. 4, pp. 43-46, 2017.

[23] L. Gao, Y. Liu, and Z. Sheng, "Application of random forest algorithm to traffic state identification," *Experiment Technology and Management*, vol. 34, no. 4, pp. 43-46, 2017.

[24] L. Zhang, Y. Jia, and Z. Niu, "Traffic state classification based on parameter weighting and clustering method," *Journal of Transportation Systems Engineering and Information Technology*, vol. 14, no. 6, pp. 147-151, 2014.

[25] Y. Zhang, X. P. Wu, and Z. L. Xu, "High-dimensional fuzzy classification system design based on multi-objective genetic algorithm," in *Proc. Twenty-Seventh Chinese Control Conf.*, Beijing: Beijing University Press, 2008.

[26] K. N. Fang, J. B. Wu, and J. P. Zhu, "A review of technologies on ran-dom forest," *Statistics and Information Forum*, vol. 26, no. 3, pp. 32-38, 2011.

[27] J. S. Su, B. F. Zhang, and X. Xu, "Advances in machine learning based text categorization," *Journal of Software*, vol. 17, no. 9, pp. 145-147, 2006.

[28] Y. Zhuang, "An improved TFIDF algorithm in electronic information feature extraction based on document position," *Lecture Notes in Electrical Engineering*, vol. 177, no. 2, pp. 449-454, 2012.
[29] J. Kashif, A. B. Haroon, and S. Mehreen, "Feature selection based on class-dependent densities for high-dimensional binary data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 3, pp. 465-477, 2012.
[30] J. He, Y. Zhang, and X. Li, "Learning naive bayes classifiers from positive and unlabelled examples with uncertainty," *International Journal of Systems Science*, vol. 43, no. 10/12, pp. 1805-1825, 2012.
[31] A. Dehzangi, S. Phon-Amnuaisuk, and O. Dehzangi, "Using random forest for protein fold prediction problem: An empirical study," *Journal of Information Science and Engineering*, vol. 26, no. 6, pp. 1941-1956, 2010.
[32] "Bagging and Ada Boost Algorithms for vector quantization," *Neurocomputing*, vol. 73, no. 1/3, pp. 106-109, 2009.
[33] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in *Proc. of the 11th Conf. on Uncertainty in Artifi-Cial Intelligence*, San Francisco: Morgan Kaufmann Publishers Inc., 1995, pp. 338-345.
[34] S. L. Salzberg, "Programs for machine learning," *Machine Learning*, vol. 16, no. 3, pp. 235-240, 1994.
[35] C. J. C. Burges, "A tutorial on support vector machines for pattern recog-nition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121-167, 1998

Xinqing Geng was born in Anshan, Liaoning province on 4 November, 1973. She received the Ph.D degree in management science and engineering from Tianjin University, China, in 2006. She received the master's degree from Computer College of University of Science and Technology Liaoning in 2002. She received the bachelor's degree from Computer Department of University of Science and Technology Liaoning, China in 1998. Her research interests are in data mining, especially web mining, knowledge discovery from text data. She is interested in web content analysis and document clustering using machine intelligence technique based on efficient structures and algorithms.

From 1995 to 2002, she was with the Computer Department of University of Science and Technology Liaoning as a teaching assistant. She was promoted as a lecturer in 2002.

She is a member of Computer Society since 2010.