FRC: A New Dynamic Fuzzy Clustering Algorithm 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 fulfills 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 information 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, text 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 unknown, 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 targets 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 partition-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 semantic 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: mengxinqin@163.com).

DOI: 10.7763/IJCTE.2018.V10.1224

190
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 engenders many leaf node, which is used as input of FCM algorithm. The result of clustering is acquired with FCM algorithm.

II. CHARACTER PRESENTATION 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

\[ V(d) = (t_1, w_1(d); \ldots; t_n, w_n(d)) \]

where \( t_i \) is term, \( w_j(d) \) is the weight in \( d \). \( w_j(d) \) is function of appearing frequency \( f_j(d) \) of \( t_i \) in \( d \), namely, \( w_j(d) = \psi(t_j(d)) \).

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

\[ \psi = t_j(d) \times \log \left( \frac{N}{n_i} \right) \]

(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 \( m \) 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:

\[ J = \sum_{i=1}^{n} \sum_{j=1}^{k} [u_{ij}]^p \| x_i - c_j \|^2 \]

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:

Step 1: Initialize membership of data points based upon the initial centroid \( U=[u_{ij}] \) matrix \( U^{(0)} \).

Step 2: Calculate the centers vectors \( C^{(k)}=[c_j] \) with \( U^{(k)} \)

\[ c_j = \frac{1}{m} \sum_{i=1}^{n} [u_{ij}]^p x_j \]

(3)

Step 3: Update \( U^{(k)} \), the formula of \( U^{(k+1)} \) is as follow:

\[ u_{ij} = \frac{[x_i - c_j]^\frac{1}{p-1}}{\sum_{j=1}^{k} [x_i - c_j]^\frac{1}{p-1}} \]

(4)

When \( \max \{ |u_{ij}(k+1) - u_{ij}(k)| \} < \epsilon, 0 \leq \epsilon \leq 1 \), 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:

\[ H = -\sum_{i=1}^{n} \sum_{j=1}^{k} [u_{ij}] \log_2 [u_{ij}] + (1-u_{ij}) \log_2 (1-u_{ij})/N \]

(5)

where \( u_{ij} \) is membership of sample data \( i \) belonging to class \( j \), 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 \( k=2 \), iteration \( h=0 \), exponential weight \( m \), iterative stop threshold \( \epsilon \).

Step 2: Sample vectors are input of random forest algorithm. The codes between root node and leaf node are 0-1 sequence.

Step 3: The code matrix which the decision tree constructs is used as inout 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

FCR 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 algorithm 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 are 2160 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>
<thead>
<tr>
<th>TABLE I: CLUSTERING RESULT OF FRC ALGORITHM</th>
</tr>
</thead>
<tbody>
<tr>
<td>agriculture</td>
</tr>
<tr>
<td>Wrong statistical number</td>
</tr>
<tr>
<td>Right statistical number</td>
</tr>
<tr>
<td>Total number of segmented texts</td>
</tr>
<tr>
<td>Precision%</td>
</tr>
<tr>
<td>recall%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE II: CLUSTERING RESULT OF FCM ALGORITHM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agriculture</td>
</tr>
<tr>
<td>Wrong statistical number</td>
</tr>
<tr>
<td>Right statistical number</td>
</tr>
<tr>
<td>Total number of segmented texts</td>
</tr>
<tr>
<td>Precision%</td>
</tr>
<tr>
<td>recall%</td>
</tr>
</tbody>
</table>

Compare Table I with Table II, the precision of FRC algorithm is higher than that of FCM in documentation. Information entropy is used to determine the number of clustering. FRC algorithm 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

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.