2009 International Conference on Machine Learning and Computing
IPCSIT vol.3 (2011) © (2011) IACSIT Press, Singapore

Clustering suggestion for Chinese news web pages from multi-media sources

Deng-Yiv Chiu [+], Ya-Chen Pan

Department of Information Management, Chung Hua University

Abstract. There exist some news obviously classified into incorrect categories on Chinese web pages portal. The main reasons could be that it is difficult to automatically classify Chinese news and the news appearing on web pages portal are retrieved from many media sources. In this study, we integrate genetic algorithm and multi-class support vector machine (SVM) classifier to construct a Chinese news classification method. In addition, we find that some similar documents are scattered in different categories. The main reason could be that the categories of original media sources are different from those of news web pages portal. Those similar news should be collected to form a new category. We try to combine genetic algorithm and fuzzy c-means algorithm to propose a new approach to offer clustering suggestion for news web pages that are scattered in different categories and are from multi-media sources.

Keywords: Chinese web pages clustering, multi-class SVM, fuzzy c-means algorithm, genetic algorithm

1. Introduction

With the rapid development of the internet, many people browse news web pages through portal. Therefore, how to automatically classify a huge amount of news retrieved from many media sources efficiently and correctly becomes very important. We find that there exist some news obviously classified into incorrect categories on Chinese web pages portal of Taiwan Yahoo!. According to our investigation, the news web pages are classified through an automatic Chinese classification information system developed in Taiwan Yahoo Co. This is a reasonable way to deal with this kind of classification. The main reasons for the news classified incorrectly are that it is intrinsically difficult to classify Chinese news and the news documents are retrieved from 31 media sources.

In this study, we explore the classification problem of Chinese portal news web pages retrieved from many media sources. In addition, we find that some similar documents are scattered in different categories. The news web pages of Taiwan Yahoo! in August, 2008 are used as the targets. We aim to construct an automatic classification approach for Chinese news classification and to offer clustering suggestion for news web pages that are scattered in different categories and are from multi-media sources.

2. Related techniques

2.1. Support vector machine (SVM)

Support vector machine (SVM) is an efficient learning machine. It has been used to solve binary classification problem, but it is still an ongoing research issue for multi-class problems. Nowadays, some methods have been proposed to solve multi-class problems, such as one-against-one and one-against-all methods.

One-against-all SVM strategy was proposed in 1994 and it is used to solve multi-class classification problem [4]. For the classification problem with k classes, k binary SVM classifiers are built to optimize

[+] Corresponding author. Tel.: +1 3518 6524; fax: +1 3518 6543.
E-mail address: chiuden@chu.edu.tw (D. -Y. Chiu).

hyper-plane separation for each class. Each binary classifier is trained by training samples with all samples belonging to class C_j and all samples not belonging to class C_j. In order to classify new example x, new example x would be sent to k SVM classifiers to evaluate. “Winner-take-all” rule will be adopted to determine which class the new example should be classified into. The class labeled with the highest decision function value would be selected to become the class label of the new sample.

One-against-one SVM strategy uses a pair of binary classifiers [6]. For the classification problem with k classes, $k(k\text{-}1)/2$ binary classifiers are built. In order to class new example x, new example x would be sent to $k(k\text{-}1)/2$ SVM classifiers to vote. “Max wins” rule is adopted and voting is used to decide which class the new example should be classified into. The class labeled with the best votes is the winner.

2.2. Fuzzy c-means algorithm (FCM)

Fuzzy c-means algorithm (FCM) is one of the fuzzy clustering algorithms introduced in 1981 [3]. It can improve the efficiency of clustering algorithm. In the traditional clustering algorithm, such as K-means, each data is classified to one category. In FCM, fuzzy theorem concept is employed to evaluate the membership degree of each data belonging to each cluster. The range of membership degree is between zero and one (zero denotes low membership degree and one denotes high membership degree) and the sum of membership degrees of one data belonging to every cluster is one. Later, it keeps updating the centroid of each cluster and the membership degree of each data belonging to each cluster repeatedly until the minimum of object function is obtained.

3. The proposed hybrid clustering approach

In this study, first we apply genetic algorithm with optimal parameter character to select four feature thresholds used to get representative features of each class and build the vector space model of each document. Then we use multi-class support vector machine with optimal hyper-plane character to produce appropriate classifier used to reach significant classification performance and efficiency. In addition, we employ genetic algorithm to select the membership degree of news belonging to each cluster. Hopefully, the similar news can be classified into the same cluster. Then, FCM clustering method is used to promote the clustering efficiency. Finally, the new news cluster suggested can be proposed.

3.1. Classification with multi-class GA-SVM

The purpose of multi-class GA-SVM method is to construct multi-class GA-SVM classifier using a combination of genetic algorithm and multi-class support vector machine. In order to get better classification performance, One-against-one SVM method is used to train. Fitness function of genetic algorithm is used to evaluate the classification performance.

3.1.1 Representative feature selection for each class

The purpose of feature selection is to obtain representative feature set and reduce noise in a specific field. Here, four thresholds including term frequency, document frequency, uniformity and conformity are used for selecting representative features [1].

(1) Term Frequency: Term frequency (*TF*) denotes the weight of occurrence probability of feature in a class, the feature with higher TF value means that the feature can represent the class better.
(2) Document Frequency: Document Frequency (*DF*) denotes the weight of occurrence probability of documents with feature in a class. The feature with higher *DF* values can represent the class better since it appears more in documents in the class than in documents in other classes.
(3) Uniformity: Uniformity denotes the occurrence weight of feature appearing in all documents in a class. The feature with higher uniformity value can represent the class better than other features.
(4) Conformity: Conformity denotes the occurrence weight of documents with feature appearing in all classes, the feature with smaller conformity value can represent the class better since the feature appears in less classes.

3.1.2 Fitness function of GA-SVM

In order to select representative features of each class, the fitness function of Chinese news for feature selection in this research mainly considers three important factors, precision, recall and F-measure.

Precision denotes the percentage of count of documents classified correctly into a class to count of documents classified into the class. Recall denotes the percentage of count of documents classified correctly into a class to count of documents belonging to the class. Because the higher the precision is, the lower the recall will be. In order to get the balance, the F-measure value is computed to consider the precision and recall simultaneously. If the values of precision and recall are higher, the value of F-measure is higher. The formula of F-measure is as below.

$$F-measure = \left(2Precision * Recall\right) \Big/ \left(Precision + Recall\right)$$

Therefore, the fitness function for our multi-class GA-SVM is shown as follow, where *I* is the number of class.

$$fitness\ \ funciton = \left(\sum_{i=1}^{I} F-measure_{s,p,C_i} \right) \Big/ I$$

3.2. Clustering with GA-FCM algorithm

The purpose of GA-FCM is to combine genetic algorithm and fuzzy c-means algorithm to suggest new category of similar documents scattered in various classes. We apply genetic algorithm with optimal parameter character to determine the membership degree of each data belonging to clusters and try to reduce the discrete degree of each cluster. We compute the centroid of each cluster by FCM with fuzzy logical membership degree character.

3.2.1 Fitness function for GA-FCM

In order to improve the efficiency of GA-FCM, we apply genetic algorithm to determine the membership degree of each document belonging to clusters to find the better membership degree. The objective function of FCM clustering algorithm emphasizes the weight u_{ij}. Because there are many undetermined parameters in traditional FCM algorithm, different values of weight or weighted index *m* could result in different result. In order to solve this problem and obtain better clustering performance, our fitness function of GA-FCM considers disjoint function to measure the cluster cohesion degree.

Disjoint function can be used to measure the cluster quality. If documents in a cluster belong to one concept, these documents are similar and distances among documents are relatively small; that is, the cohesion of this cluster is high [2]. Disjoint function considers the document similarity in one cluster. If the value of document similarity in the cluster is high and the distance among documents is small, the disjoint value is small. Therefore, disjoint function could measure if documents in the same cluster have high similarity. The disjoint function is shown as follow.

$$Disjoint_c = \sum_{n=1}^{l_c} \sum_{n'=1}^{l_c-1} d\left(o_n, o_{n'}\right) \Big/ \left[I_c \left(I_c - 1 \right) / 2 \right]$$

where o_n denotes document n in cluster c, and $o_{n'}$ denotes document n' in cluster c ($n \neq n'$), and l_c denotes the number of documents in cluster c, $d\left(o_n, o_{n'}\right)$ denotes the dissimilarity between document n and document n'. Therefore, the fitness function for the GA-FCM is shown as follow. This equation is expected to evaluate discrete degree among documents in clusters. The lower of the discrete degree in a cluster is, the better of cluster quality is.

$$fitness\ \ function = \sum_{c=1}^{C} Disjoint_c$$

where C is the number of clusters, and j is the generation of chromosomes, $Disjoint_j$ denotes disjoint value in jth generation.

4. The Architecture of the proposed method

The detailed explanation of the proposed method is as follows.

(1) *Data preprocess*: We collect documents as data from news webpage of Taiwan, Yahoo! http://tw.yahoo.com. The experimental data is electronic documents of Chinese news collected from 31 different media sources. The collected documents are segmented by CKIP Chinese Word Segmentation System and we select features belonging to general noun (Na), place noun (Nc) and terminology (Nb) to form candidate features. In addition, we also delete the candidate features with the length of string of 1 to reduce document noise.

(2) *Presenting training documents with vector space model (VSM):* In order to form the vector of each

document, we select features of each class that satisfy the thresholds of the chromosome.

(3) *Initialization of membership degrees for GA process:* We produce the chromosomes of the first generation randomly. A generation includes 20 chromosomes and each chromosome stands for the four thresholds used to obtain representative feature.

(4) *Training SVM classifier with training documents:* The vector space models of training data are used to train the SVM classifier. One-against-one SVM classifier is trained with documents of 6 classes.

(5) *Evaluation of fitness values of classifier*: The fitness value of classification performance is calculated. Larger value indicates higher classification performance.

(6) *Termination criterion of GA for classifier training*: The termination criterion is evolution of 100 generations. If the criterion is not met, produce chromosome of next generation and then go to step (4).

(7) *Necessary examination of clustering:* After get the optimal SVM training model, testing documents will be sent to the trained classifier to class testing documents. After classifying testing data with SVM classifier, we check if there exists any unclassified document. If so, we cluster those documents as followings. Otherwise, go to step (13).

(8) *Vector space model construction for unclassified documents:* From unclassified documents, we extract features appearing more than one time to form base for vector space model for the unclassified documents. In order to perform document clustering efficiently, we adopt binary method to build document vector base.

(9) *Initialization of membership degrees for GA process:* We produce the chromosomes of the first generation randomly. A generation includes 20 chromosomes and each chromosome stands for the membership degrees of documents belonging to clusters.

(10) *Fuzzy C-means Clustering:* FCM clustering algorithm is used to cluster documents. It clusters unclassified documents to form clusters of similar documents.

(11) *Evaluation of fitness values of cluster performance*: The fitness value of clustering performance is calculated. Larger value indicates higher accuracy of cluster algorithm and documents properly clustered into the similar clusters.

(12) *Termination criterion of GA for fuzzy c-means*: The termination criterion is evolution of 100 generations. If the criterion is not met, produce chromosome of next generation and then go to step (10).

(13) *Result evaluation*: Finally, the evaluation of proposed method is performed by means of precision, recall, and F-measure.

5. EXPERIMENTS

Here, we introduce experimental data and performance evaluation in this section.

The experimental data and existing document classification structure in yahoo.com.tw are collected from internet between August 13, 2008 and August 20, 2008. The empirical documents are Chinese news.

The document classification structure consists of 9 classes including policy, finance, health, education, sport, film, technology, art and travelling.

The distribution of collected documents and class title are shown in Table 1. There are totally 9 classes and 6578 documents of which 4469 documents are used as training data of SVM classifier and 2109 documents are used for testing. The training data excludes the news classified incorrectly originally. The testing data includes some of news classified correctly from class1 to class 6 and those belonging to class 7 to class 9.

In order to testify each phase of the proposed approach, the data collection is designed as below.

(1) For training data, there are no training data collected from class 7 to class 9 for verification purpose.

(2) Another 2109 documents are collected from class 1 to class 9 for testing data. 1188 of the 1488 documents are collected from class 1 to class 6. Those documents are mainly used to testify GA-SVM classifier built in the first phase. In other hand, the remaining 300 documents are collected from class 7 to class 9. Most of those documents are supposed to become unclassified documents after first phase and to be clustered by the GA-FCM algorithm in the second phase.

Table 1. Datasets from the Yahoo! web site

No.	Class	Training data	Testing data	No.	Class	Training data	Testing data	No.	Class	Testing data
1	Policy	996	279	4	Education	417	91	7	Technology	100
2	Finance	1025	286	5	Sport	1027	291	8	Art	100
3	Health	342	64	6	Film	662	177	9	Travelling	100

Table 2. The precision, recall, and F-measure of GA-SVM classifier in training process

Class	Precision	Recall	F-measure
Policy	90.61%	79.57%	84.73%
Finance	76.65%	89.51%	82.58%
Health	83.93%	73.44%	78.33%
Education	72.15%	62.64%	67.06%
Sport	86.65%	95.88%	91.03%
Film	92.11%	79.10%	85.11%
Average	83.68%	80.02%	81.47%

Table 3. The precision, recall, and F-measure of GA-FCM process

Class	Precision	Recall	F-measure
Technology	60.00%	84.00%	70.00%
Art	68.79%	97.00%	80.50%
Travelling	53.19%	75.00%	62.24%
Average	60.66%	85.33%	70.91%

The trained GA-SVM classifier is used to construct a Chinese news classification. The precision, recall, and F-measure are shown as Table 2. The precision of each class is 90.61%, 76.65%, 83.93%, 72.15%, 86.65%, and 92.11%, respectively. And the average precision is 83.68%. The recall of each class is 79.57%, 89.51%, 73.44%, 62.64%, 95.88%, and 79.10%, respectively. And the average recall is 80.02%. The F-measure of each class is 84.73%, 82.58%, 78.33%, 67.06%, 91.03%, and 85.11%, respectively. And the average F-measure is 81.47%.

The GA-FCM classifier is used to propose new approach to offer clustering suggestion for news web pages. The precision, recall, and F-measure are shown as Table 3. The precision of each class is 60.00%, 68.79%, and 53.19%, respectively. And the average precision is 60.66%. The recall of each class is 84.00%, 97.00%, and 75.00%, respectively. And the average recall is 85.33%. The F-measure of each class is 70.00%, 80.50%, and 62.24%, respectively. And the average F-measure is 70.91%.

6. CONCLUSIONS

In this paper, information retrieval, multi-class support vector machine, fuzzy c-means and genetic algorithm are used to propose an appropriate approach to construct a Chinese news classification and to offer clustering suggestion for news web pages that are scattered in different categories and are from multi-media sources. The Taiwan Yahoo! News web pages are selected to test the proposed method.

For future studies, the proposed method can be incorporated with other semantic analysis method to retrieve representative features. Also, the number of clusters setting to improve empirical performance can be another issue to explore.

7. References

[1] C. H. Chou, C. C. Han, and Y. H. Chen. GA based Optimal Keyword Extraction in an Automatic Chinese Web Document Classification System. *Lecture notes in Computer Science*. 2007, **4743**, pp. 224-234.

[2] F. R. Lin and C. M. Hsueh. Knowledge Map Creation and Maintenance for Virtual Communities of Practice. *Information Processing & Management*. 2006, **42**(2), pp. 551-568.

[3] J. C. Bezdek. *Pattern recognition with Fuzzy Objective Function Algorithms*. Plenum Press. New York, 1981.

[4] L. Bottou, C. Cortes, J. Denker, H. Drucker, I. Guyon, L. Jackel, Y. LeCun, U. Muller, E. Sackinger, P. Simard, and V. Vapnik. Comparison of Classifier Methods: A Case Study in Handwriting Digit Recognition. *Proc. of International Conference on Pattern Recognition*, 1994, pp. 77-87.

[5] L. X. Xie, and G. Beni. A Validity Measure for Fuzzy Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1991, **13**(9), pp. 841-847.

[6] S. Knerr, L. Personnaz, and G. Dreyfus. Single-Layer Learning Revisited: A Stepwise Procedure for Building and Training a Neural Network, *Neurocomputing: Algorithms, Architectures and Application, F68, Springer-Verlag*. 1990, pp. 41-50.