Extraction of Text Classification Rules Based on Multi-Population Collaborative Optimization

He Liu[1,2] and Dayou Liu[1,2]
[1]*College of Computer Science and Technology*
[2]*Key Laboratory for Symbolic Computation and Knowledge Engineering of Ministry of Education*
Jilin University
Changchun, China
liuhe1980@163.com and dyliu@jlu.edu.cn(Corresponding Author)

***Abstract*—Most text classification methods are highly complicated on computation and can not be used on the occasion of classifying a large number of texts. A novel approach based on multi-population collaborative optimization was proposed for the extraction of text classification rules. The mutual information was applied to generate the initial populations and the multi-population collaborative optimization method was adopted to evolve the current population. Experimental results show that the number of classification rules is small, the accuracies of classification rules are high and the time of computation is short using this approach. And this approach is competent for processing the large-scale text datasets.**

***Keywords*—text classification, rule extraction, collaborative optimization, mutual information, genetic algorithm**

I. INTRODUCTION

With the rapid growth of the World Wide Web (WWW), a mass of text and multimedia information can be acquired through the internet expediently. The abundant and heterogeneous information contains great and valuable knowledge. And it is need to find valuable knowledge from the vast information. Therefore, some new techniques need to be research which can find, extract and filtrate out the valuable knowledge from the internet automatically. Text classification (TC, also known as text categorization) is a fundamental technique for information organization and management, and it has become one of the most important topics in the field of data mining [1].

Text classification is the task of assigning predefined category labels to new texts based on the rule suggested by a training text set. Text classification is often used as a basis for applications in text processing and visualization, Web mining, technology watch, patent analysis, etc. So far, a growing number of machine learning algorithms have been proposed for text classification, including naive Bayes classifier [2], decision tree [3], kNN classifier [4], online classifier [5], Rocchio classifier [6], support vector machine [7] and boosting method [8] etc. However, these approaches are highly complicated on computation, and can not be used to classify a large number of texts. For this reason, a novel approach based on multi-population collaborative optimization is proposed for the extraction of text classification rules in this paper. This approach applies mutual information to generate the initial populations and adopts the multi-population collaborative optimization method to evolve the current population. The optimization performance of this approach has been improved largely by the competition and sharing mechanism among populations. Experimental results show that the number of classification rules is small, the accuracies of classification rules are high and the time of computation is short using the approach proposed in this paper. And this approach is competent for processing the large-scale text datasets.

The rest of this paper is organized as follows: Section II introduces the information classification based on text mining briefly. Section III describes a method for generating the initial populations based on mutual information in detail. Section IV depicts the framework of multi-population collaborative optimization. The experimental results and analyses are presented in Section V. Finally, Section VI concludes this paper.

II. INFORMATION CLASSIFICATION BASED ON TEXT MINING

Text mining is a new branch in the field of data mining. Firstly, it extracts terms from texts by text cut technique and changes text data to structured data which can describe contents of texts. Afterward, the structured tree comes into being through using the techniques of data mining such as classification, clustering, relation analysis and so on. Furthermore, the new conceptions are found and the relevant relations are obtained in this structure.

In the flow of text classification based on text mining, it is the most central step to extract text classification rules and it is the emphases of this paper. This paper uses the multi-population collaborative optimization approach to extract text classification rules. And then it applies mutual information to generate the initial populations and adopts the multi-population collaborative optimization approach to evolve the current population. And the optimization performance of this approach is improved largely by the competition and sharing mechanism among populations. According the method of ordinal overlay, it tries to mine a list of text classification rules which is able to cover with most training texts. The executive flow of the approach proposed in this paper can be described as follows:

1) Initialization of Variables: Set the list of rules empty which have been found and add all training texts into the

training text set.

2) Evolvement of Multi-population Collaborative Optimization: A text classification rule can be found in each evolvement of multi-population optimization method. The rule which is found in this time is added to the list of rules which have been found after the evolvement of multi-population optimization method and the texts covered by this rule are deleted from the training text set.

3) Condition of End: The multi-population optimization method stops evolving when the number of texts which are not covered is smaller than the default set by the user.

III. The Method for Generating the Initial Populations Based on Mutual Information

In the process of the evolvement of multi-population collaborative optimization, the individual whose fitness is low is eliminated gradually and the individual whose fitness is high is reserved at the same time. The point whose objective function obtains the optimal value is the perfect solution of optimization. The problem of optimization in this paper is mining classification rules contained in the training texts. An array is used for expressing a chromosome corresponding to a single classification rule. And the encoding mode of each classification rule's chromosome is as follows:

t_1	t_2	⋯	t_n	c

.

Where t_i is the value of a certain attribute in the classification rules, c is the value of the class to which the classification rule belongs.

Before describing the method proposed in this paper, several conceptions are introduced as follows:

1) Term: The term $term_{ij}$ expresses the condition $A_i=V_{ij}$, where A_i is the ith attribute and V_{ij} is the jth value of A_i. For example, let *age* is the second attribute which has four values, namely, *juvenile*, *youth*, *midlife* and *senile* in turn, then $term_{21}$ expresses the condition *age= juvenile*, $term_{22}$ expresses the condition *age= youth*, $term_{23}$ expresses the condition *age= midlife*, and $term_{24}$ expresses the condition *age= senile*.

2) Selection Probability of Term: Each term is selected to join the current rule according to a certain probability. The selection probability is based on mutual information [9].

Definition 1 (Mutual Information of Term) Let D is the training text set, $C=\{c_1,c_2,...,c_k,...,c_l\}$ is the class set, l is the number of classes in D, $P(term_{ij}|c_k)$ is the empirical probability to which $term_{ij}$ appears according in the text which belongs to the class c_k, $P(c_k)$ is the empirical probability to which a text appears in the class c_k according, $P(term_{ij})$ is the empirical probability to which $term_{ij}$ appears according. Then the mutual information of $term_{ij}$ is defined as follows:

$$MI(term_{ij}) = \sum_{k=1}^{l} P(c_k) \log \frac{P(term_{ij} \mid c_k)}{P(term_{ij})}. \quad (1)$$

The mutual information of a term represents the value distribution of the attribute in the condition which this term expresses. The smaller the value of $MI(term_{ij})$ is, the evener the value distribution of the attribute A_i is.

Definition 2 (Selection Probability of Term) Let n is the number of attributes, m_i is the number of values of the attribute A_i, x_i is a Boolean variable, if the attribute A_i has not been added into the current rule, the value of x_i is 1, and conversely the value of x_i is 0. Then the selection probability of $term_{ij}$ is defined as follows:

$$P_{ij} = \frac{MI(term_{ij})}{\sum_{i=1}^{n} x_i \cdot \sum_{j=1}^{m_i} MI(term_{ij})}. \quad (2)$$

The selection probability of a term expresses the average information amount of this term. The larger the value of P_{ij} is, the more the relevance of this term and the current classification is. Here it is a great probability for $term_{ij}$ to be added into the current classification rule.

3) Insertion Condition of Term: If the term $term_{ij}$ is able to meet the following two conditions at the same time, this term can be added into the current rule.

i. The attribute A_i has not been added into the current rule.

ii. After $term_{ij}$ is added into the current rule, the number of the training texts with which this rule covers is larger than the threshold preset by user, that is, the minimal number of texts with which each rule covers.

4) Insertion Termination Condition of Term: If the current state meets one of the following two conditions, stop inserting terms into the current rule.

i. In the current classification rule, the number of terms is larger than or equal to the number of attributes.

ii. No terms can meet the insertion condition of term.

According to the conceptions mentioned above, this paper proposed a method based on mutual information for generating the initial populations. And this method is described as follows:

Algorithm 1 An Algorithm for Generating the Initial Populations

Step1. Generate an empty rule;

Step2. Select a term marked as $term_{ij}$ at random by the selection probability marked as P_{ij};

Step3. If $term_{ij}$ meets the insertion condition of term, insert $term_{ij}$ into the current rule;

Step4. If the current state does not meet the insertion termination condition of term, go to *Step 2*, or else go to *Step 5*;

Step5. Select the text class which is able to cover with the maximum training texts for the current rule;

Step6. Repeat from *Step 1* to *Step 5* until the number of individuals in the initial populations is equal to the threshold preset by user.

IV. MULTI-POPULATION COLLABORATIVE OPTIMIZATION (MPCO)

This section describes the implementation details of multi-population collaborative optimization. The main contents include: 1) Framework of Multi-population Collaborative Optimization. 2) Fitness Evaluation Function of Rule. 3) Competition Method among Populations. 4) Sharing Mechanism among Populations.

A. Framework of Multi-population Collaborative Optimization

The template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization). This template was designed for two affiliations.

In nature, the biology in different regions has different characteristics and evolution degree, and they contest resource from nature for their own use. In addition, these biology learn from other's strong points to offset own weakness through information exchange for collective advancement. The approach of multi-population collaborative optimization proposed in this paper is designed referring to this phenomenon in nature. The whole method is composed of some ordinary populations and an excellent population in computing environment. Each ordinary population gets computing resource through the competition from computing environment. Once an ordinary population gets hold of computing resources, it will carry out an own evolution. Then each ordinary population contributes excellent individuals obtained by own evolution to compose an excellent population from which ordinary populations can get excellent individuals to improve their own quality.

In the approach of multi-population collaborative optimization, the evolution of each ordinary population is accomplished through various genetic algorithms mainly. The difference among various genetic algorithms is embodied in the phases: evolution mechanisms, optimizing operator, control parameters, etc. In view of limitations space, this paper will not give unnecessary details of various genetic algorithms.

B. Fitness Evaluation Function of Rule

This paper designs the fitness evaluation function of the genetic algorithm mainly in the two ways of support and confidence.

1) Support: The support represents the universality of a rule in the training text set, namely, the proportion of texts which support this rule in the training text set. The larger the support is, the larger the proportion of texts which support this rule in the training text set is.

2) Confidence: The confidence denotes the credibility that the premise of the rule (the term attributes) leads to the conclusion (the class attribute) and reflects the reasoning uncertainty in the instance of the incomplete knowledge. The larger the confidence is, the easier the relevant conclusion is educed as long as the premise (terms) is true. Because the rules express the uncertain or incomplete knowledge, the fore item and the rear item are assigned by a certain value of the confidence. In the reasoning, this value is processed in a predetermined manner and each conclusion has a confidence which denotes the credibility that this conclusion comes into existence.

The design of the fitness evaluation function should consider the part of term attributes, the part of class attributes, and their matching synthetically. This paper combines the support with the confidence to design the fitness evaluation function. During the process of evolution, each rule competes with others through own fitness evaluation function and the rule can exist whose support and confidence are all large.

Definition 3 (Fitness Evaluation Function of Rule) Let N is the number of training texts, $N_1(r)$ is the number of texts which meet the whole rule r, $N_2(r)$ is the number of texts which meet the premise of the rule r, then the support of the rule r is $SI(r)= N_1(r)/ N$, the confidence of the rule r is $CI(r)= N_1(r)/ N_2(r)$, and the fitness evaluation function of the rule r is defined as follows:

$$Fitness(r) = SI(r) + CI(r) \tag{3}$$

C. Competition Method among Populations

In the approach of multi-population collaborative optimization, the competitive power of population lies on the value of the maximum fitness. It is a great probability for the population whose competitive power is strong to obtain computing resource. In the process of finding the maximum fitness, if the maximum fitness value of a population is little, this population still is in the initial phase of the evolution and is far from the optimal solution. And the resource should be assigned to the populations which are near to the optimal solution. In this way, the population whose maximum fitness is large will obtain computing resource first.

Definition 4 (Competitive Power Index of populations) Let N_p is the number of ordinary populations, $F_{\max}^{i}$ is the maximum fitness value of the ith population. Then the competitive power index of the ith population is defined as follows:

$$CP_i = \frac{F_{\max}^{i}}{\sum_{i=1}^{N_p} F_{\max}^{i}} \tag{4}$$

The competitive power index of a population denotes the competitive power of this population for the resource. The larger the competitive power index of a population is, the earlier this population obtains computing resource.

D. Sharing Mechanism among Populations

The excellent population composed of the some individuals whose fitness values are large in various ordinary populations is an excellent seed warehouse shared by evolutions of various populations. There are two sharing mechanism: excellent seed migration and excellent seed cross. The excellent seed migration indicates that ordinary populations introduce some excellent individuals from the excellent population to replace own bad individuals in their own evolution. And the excellent seed cross indicates that ordinary populations select some excellent individuals from

the excellent population to cross with the own individuals in their own evolution and the original individuals in the ordinary populations are replaced by the excellent individuals generated through the cross.

V. EXPERIMENTS

A. Datasets

To test the performance of the approach proposed in this paper, the two normal text datasets are used to experiment.

1) Reuters-21578(*Reuters*)*:* Reuters is one of the most usual and normal text datasets in the field of text mining that includes 21,578 texts in 135classes [10]. In the experiments, the texts which belong to one class at least and have the Lewis split mark are selected as the experimental data. At the same time, the first class mark of the text is regarded as the normal class mark of this text.

2) 20 Newsgroups (*20NG*)*:* 20NG is also a usual text dataset that includes 20,000 pieces of news in 20 newsgroups [11]. 20 NG has two versions, and the second version is used in the experiments. The most of heads in the texts and the repetitive texts in the first version are deleted to form the second version which includes 18,828 texts.

B. Evaluation Criterion

The performance of the method for extraction of text classification rules is evaluated in the following three ways in this paper mainly.

1) Number of Rules: After the extraction of classification rules from the testing dataset, the number of classification rules can be obtained. Generally users hope that the number of classification rules is as small as possible.

2) Accuracy of Rule: When the extracted classification rules are applied to classify the testing dataset, the accuracies of classification rules can be obtained. Apparently users hope that the accuracy of a classification rule is as high as possible.

3) Time of Computation: Firstly classification rules are extracted, and then the testing dataset is classified by these rules, finally the result of classification is obtained. It is no doubt that users hope that the time of this process is as short as possible.

C. Experimental results and analysis

To test the performance of the approach proposed in this paper, we make experiments on the two text datasets mentioned above with CN-2 [12], Fuzzy Decision Tree (FDT) [13] and the approach proposed in this paper (MPCO) respectively. The basic configuration of the personal computer in the experiments is as follows: CPU is Pentium IV, frequency is 2.4GHz, and memory is 512MB. Ten fold cross validation method is applied in the experiments. The results of the experiments on the two text datasets with the three different methods are shown in the TABLE I.

TABLE I. RESULTS OF THE EXPERIMENTS

Dataset	Evaluation Criterion	CN-2	FDT	MPCO
Reuters	Rules Number	174.9	156.6	139.5
	Rule Accuracy	85.6%	86.8%	91.3%
	Computation Time (s)	1360.4	1196.5	908.7
20 NG	Rules Number	131.2	109.1	95.6
	Rule Accuracy	80.5%	85.1%	89.2%
	Computation Time (s)	1201.9	1103.3	850.4

TABLE I shows that: *1) Number of Rules:* For Reuters, the number of rules which are extracted by MPCO is smaller than the number of rules which are extracted by the other two methods evidently; for 20NG, the number of rules which are extracted by MPCO is also obviously smaller than the number of rules which are extracted by CN-2 and is approaching to the number of rules which are extracted by FDT. *2) Accuracy of Rule:* For Reuters, the accuracy of MPCO is higher than the accuracies of the other two methods markedly; for 20NG, the accuracy of MPCO is also remarkably higher than the accuracies of the other two methods. *3) Time of Computation:* For the two datasets, the time of MPCO is shorter than the time of the other two methods clearly.

The reasons that the performance of the method proposed in this paper is high are as follow: *1) The method for generating the initial populations based on mutual information:* For each term which can be added to the current classification rule, the method proposed in this paper applies selection probability to estimate average information amount of a term and selects terms to join the current rule according to the selection probabilities of these terms. The method proposed in this paper can improve the accuracies of classification rules effectively. The experimental results in TABLE I indicate that the accuracies of classification rules extracted by the method proposed in this paper are higher than the accuracies of the other two methods. *2) The extraction mechanism of classification rules based on multi-population collaborative optimization:* The approach of multi-population collaborative optimization proposed in this paper is designed referring to the phenomenon in nature that the biology learn from other's strong points to offset own weakness through information exchange for collective advancement. The whole method is composed of some ordinary populations and an excellent population in computing environment. Each ordinary population gets computing resource from computing environment through the competition. Once a population gets hold of computing resources, it will carry out an own evolution. Then each ordinary population contributes excellent individuals obtained by own evolution to compose an excellent population from which ordinary populations can get excellent individuals to improve their own quality. The optimization performance of the multi-population collaborative optimization approach is improved largely by the competition and sharing mechanism among populations. The experimental results in TABLE I indicate that the number of classification rules is small, the accuracies of classification rules are high and the time of computation is short using the approach proposed in this paper. So this approach is competent for processing the large-scale text datasets.

VI. Conclusion and Future Work

In this paper, a novel approach based on multi-population collaborative optimization is proposed for the extraction of text classification rules. And this approach applies mutual information to generate the initial populations and adopts the multi-population collaborative optimization method to evolve the current population. The experimental results show that the number of rules extracted by the approach proposed in this paper is small and the accuracies of these rules are high. At the same time, the computational time of this approach is short. So this approach is competent for processing the large-scale text datasets.

The future works are as follows: *1) Set the parameters in the approach proposed in this paper:* It makes this approach resolve the optimization problems rapidly and efficiently through setting each parameter in this approach. *2) Improve the competition and sharing mechanism of multi-population collaborative optimization:* It makes the optimization performance of the multi-population collaborative optimization approach improved largely through improving of competition and sharing mechanism among populations. *3) Simplify the classification rules:* It makes the number of terms in the current classification rule as small as possible through simplifying the classification rules.

Acknowledgment

This work was supported in part by the NSFC Major Research Program 60496321, National Natural Science Foundation of China under Grant Nos. 60573073, 60503016, 60603030, 60773099, 60703022, the National High-Tech Research and Development Plan of China under Grant No. 2006AA10Z245,2006AA10A309, the Major Program of Science and Technology Development Plan of Jilin Province under Grant No. 20020303, the Key Program of Science and Technology Development Plan of Jilin Province under Grant No. 20060213, the Science and Technology Development Plan of Jilin Province under Grant No. 20030523, European Commission under Grant No. TH/Asia Link/010 (111084).

References

[1] S. C. Hoi, R. Jin, and M. Lyu, "Large-scale text categorization by batch mode active learning," World Wide Web Conference, 2006, pp. 633-642.

[2] D. D. Lewis, "An evaluation of phrasal and clustered representations on a text categorization task," SIGIR-92, 15th ACM International Conference on Research and Development in Information Retrieval, 1992, pp. 37-50.

[3] D. D. Lewis and M. Ringuette, "A comparison of two learning algorithms for text categorization," SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, US, 1994, pp. 81-93.

[4] Y. Yang, "Expert network: effective and efficient learning from human decisions in text categorization and retrieval," SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval, Dublin, IE, 1994, pp. 13-22.

[5] E. Wiener, J. O. Pedersen and A. S. Weigend, "A neural network approach to topic spotting," SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, US, 1995, pp. 317-332.

[6] D. A. Hull, "Improving text retrieval for the routing problem using latent semantic indexing," SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval, Dublin, IE, 1994, pp. 282-289.

[7] T. Joachims, "Text categorization with support vector machines: learning with many relevant features," ECML-98, 10th European Conference on Machine Learning, Chemnitz, DE, 1998, pp. 137-142.

[8] R. E. Schapire, Y. Singer and A. Singhal, "Boosting and rocchio applied to text filtering," SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval, Melbourne, AU, 1998, pp. 215-223.

[9] K. W. Church and P. Hanks, "Word association norms, mutual information and lexicography," The 27th Annual Meeting of the Association for Computational Linguistics (ACL), Vancouver, BC, Canada, 1989, pp. 76-83.

[10] R. Bekkerman, R. EI-Yaniv, N. Tishby, et al., "On feature distributional clustering for text categorization," The 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, New Orleans, Louisiana, 2001, pp. 146-153.

[11] S. Zelikovitz and H. Hirsh, "Using LSI for text classification in the presence of background text," The Tenth International Conference on Information and Knowledge Management, ACM Press, Atlanta, Georgia, 2001, pp. 113-118.

[12] P. Clapk and R. Boswell, "Rule induction with CN2: some recent improvements," Artificial Intelligence, Y. Kodratoff (eds.), Springer-Verlag, 1991, pp. 151-163.

[13] Y. Wang and Z. O. Wang, "Text categorization rule extraction based on fuzzy decision tree," Computer Applications, vol. 25, no.7, 2005, pp. 1634-1637.