Chinese Text Classification Based on Ant Colony Optimization

LUO Xin [1, a]

[1]College of Business Administration

South China University of Technology, SCUT

Guangzhou, China

[a]Email:luoxin@scut.edu.cn

Keywords: Text processing; classification; Artificial intelligence; Ant colony optimization

Abstract. It's significance for us to study Chinese Text Classification, when we face so much dynamic information. The development of Text Classification has a close connection with Pattern Recognition. However, some peculiarity of Chinese Text Classification, such as it has many classes, much noise, and excessive samples, make Pattern Recognition difficult to classify texts. Recently, Artificial Intelligence provides a new intellectualized method to Text Classification. This paper tentatively leads Ant Colony Optimization, a ripe algorithm of Swarm Intelligence, into Text Classification. We construct a Text ACO-Miner Classification Model based on Ant Colony Optimization, and test it. The result shows the model can accurately be used to classify Chinese texts.

Introduction

More and more information appears on the Internet, in our life work. Text classification is a key technology to deal with a large amount of text data. It can solve the problem of information explosion in a large extent.

In 1990s, the text classification method based on artificial intelligence gradually, more attention to classification model of automatic mining and generation and dynamic optimization ability, the classification results and the flexibility of the text classification model based on knowledge engineering and expert system has been a breakthrough, become the relevant fields of research and application of the classic example[1]. Almost all important artificial intelligence algorithms have been introduced into the text classification field.

In 1991, Italy scholars Dorigo M. were inspired by the ant colony foraging behavior, proposing Ant Colony Optimization (ACO)[2]. As the ant colony algorithm has strong robustness, adaptive, positive feedback and excellent distributed computer system, easy to combine with other algorithms, it has become a hot research in the field of artificial intelligence[3]. Based on ant colony algorithm for rule digging initially by Brazilian scholar Parpinelli R.S. in 2002 is proposed to extract classification rules from the database, and used to predict future data types[4]. In recent years, the algorithm caused widespread concern in the academic circles, has been used in many fields[5][6]. The theory of the classification rule mining algorithm also provides a new intelligent method for Chinese text classification, but there are few researches in this field. In this paper, we will try to introduce the ant colony algorithm to text classification.

Text classification model based on ACO

Definition rules.

Defines the connection of the ant colony search path as the feature node and the class node, that is, the path corresponding to each ant is a rule. In a rule, each feature node is only one time and must have a class node, which corresponds to the text feature discrete values, as shown in Fig. 1.

© 2016. The authors - Published by Atlantis Press

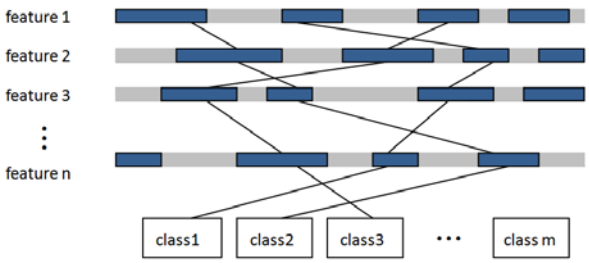

Fig.1 Classification Rules of Text ACO-Miner

Initialization, randomly generating a rule, the rule of the form is: IF ＜ $term_1$AND $term_2$ AND… ＞ THEN ＜ class ＞

Where, $term_i$ is the condition item, term=<feature number =feature value>. Since the Chinese text vector data is a continuous value, at the beginning, data need to be discredited.

Classification process.

The text ACO-Miner classification model can be divided into two parts, training and testing, as shown in Fig.2. Where, ⇦ said the training process, ⇦ said the test process.

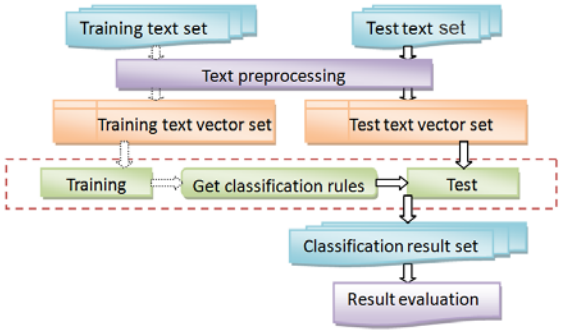

Fig.2 Text Classification Process

There are 3 stages in the training process: rule structure, fitness calculation, rule to cover the training data. The classification rules are obtained by the training process, and the testing process is classified by the classification rules. Using classification rules to classify the text is very simple, a training process is an important part, its pseudo code is as follows:

```
Procedure：Text ACO-MinerTrain
RS= ∅                                  %%Initialization rule set RS is empty
FOR（i=1,i++,i<=M）                    %%The training set consists of M classes
    TS=Train Set                       %%Set TS as training text vector set
    WHILE (TSi>threshold)              %%When the number of text vector of class i is greater than the threshold
   BestRule＝ACORuleConstructer()      %%Run the program, Fig. 3
      RS=RS＋BestRule                  %%Update rule set RS
      TS=TS–CTR                        %%CTR is a text vector that has  been covered by the found rule
    END WHILE
END FOR
```

Rule structure.

In the pseudo code, Rule structure flow chart is shown in Fig.3.

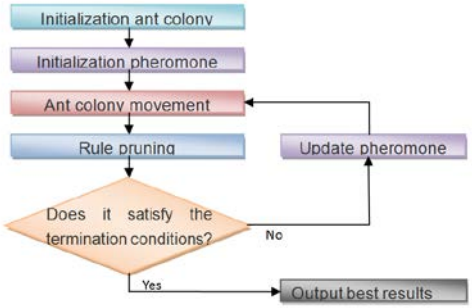

Fig.3 Rule Structure Flow Chart

The operations to be carried out are:

Initialize the ant colony. And M ants are randomly distributed on a node of the first attribute.

Initialize the pheromone. The pheromone concentration of all path nodes is initialized to the same value.

$$\tau_{ij}(0) = \frac{1}{\sum_{i=1}^{a} b_i} \tag{1}$$

where,τ_{ij} is the pheromone concentration of term_{ij}, a is the total number of attributes in the database(not including the category attributes), b_i is all the possible values for the attribute i.

Ant colony movement. Ant according to Eq.2 to choose the next node.

$$P_{ij}(t) = \frac{\tau_{ij}(t)^{\alpha} \cdot \eta_{ij}(t)^{\beta}}{\sum_{i=1}^{a}\sum_{j=1}^{b_i} \tau_{ij}(t) \cdot \eta_{ij}(t)} \tag{2}$$

Roulette wheel mechanism used to select the attribute node. For each attribute, the probability that the node term_{ij} is chosen is $P_{ij}(t)$. $\tau_{ij}(t)$ is the pheromone concentration of term_{ij}, and η_{ij} is a heuristic function of term_{ij}. $\alpha$and $\beta$are two parameters, which reflect the relative importance of the pheromone concentration and the heuristic function in the path selection process. Heuristic function value formulaη_{ij} is as follows:

$$\eta_{ij} = \frac{\max(\sum_{n} freqT_{ij}^{1}, \sum_{n} freqT_{ij}^{2}, \cdots \sum_{n} freqT_{ij}^{k})}{\sum_{n} T_{ij}} \tag{3}$$

Where, T_{ij} is the text number meeting the conditional term term_{ij}; $freqT_{ij}^{w}$ for the frequency, the category is w in the T_{ij}. In the process of rule mining, each rule is obtained, to remove the records that are part of the rules, Thus, the value of $\max(\sum_{n} freqT_{ij}^{1}, \sum_{n} freqT_{ij}^{2}, \cdots \sum_{n} freqT_{ij}^{k})$ and $\sum_{n} T_{ij}$ after a final rule is changed, it needs to be updated dynamically.

Rule pruning. The validity of the rule is calculated by the Eq.4.

$$Q = R_i \times P_i \tag{4}$$

Where, R_i is recall, P_i reflects Precision of classification results.

Pruning method is that turning removal feature nodes, that can make the rules more effective. It is to remove the redundant feature nodes, until either the feature node removal will reduce effectiveness of the rules.

If the end condition is reached. Get good enough rule or maximum number of iterations, then the end, otherwise returns to step 3.

When several consecutive ant search to the same path, that search convergence, the path of the pruning rules become a final rule. Or when the number of iterations reaches a threshold, the best quality rule will be used as the final rule in all iterations. In the iterative process, the better rule, because of the increasing of the concentration of pheromone, can be preserved, and is regarded as the final classification rule. Other rules for poor quality are discarded.

Update the pheromone. The pheromone concentration of the node is updated, according to the Eq.5:

$$\tau_{ij}(t+1) = (1-\rho)\tau_{ij}(t) + \Delta\tau_{ij} \tag{5}$$

Where, ρ is the attenuation coefficient of the pheromone concentration, which is usually located in $\rho<1$, to avoid the infinite accumulation of the pheromone on the path. Q_k is the quality of the classification rule, found by the k ant in the process of this iteration. $\Delta\tau_{ij}^{k}(t)$ is the amount of information, which is left on the node term_{ij} by the k ant in this iteration. All pheromone concentration in the attribute node try to be updated.

Model Validation

Verification steps.

In this paper, the Chinese text data set is provided by Ronglu Lee, working at the natural language processing group of the international database center, the computer information and Technology Department of Fudan University. From this data set, we randomly select 3240 documents as the experimental data. Firstly, the experimental data set is divided into the training set A and the test set B according to the ratio of 1:1.Then, the A and B are input to the preprocessing program, and the information gain, the χ^2 statistics, the mutual information, the text evidence right, the expected cross entropy were selected as feature selection method. 5 sets of vector space matrix corresponding to the feature selection method are obtained, Training set vector space matrix, denoted as $A_j(j=1,2,3,4,5)$, Test set vector space matrix, denoted as $B_j(j=1,2,3,4,5)$. Then $A_j(j=1,2,3,4,5)$ and $B_j(j=1,2,3,4,5)$ are input to classification program. Classification results $C_j(j=1,2,3,4,5)$ of SVM, KNN and NB are obtained. Choose the best results of C_k as a benchmark for evaluating the effect of Text ACO-Miner, then, input A_k (corresponding to C_k) to Text ACO-Miner to get classification rules. Use this rule to classify B_k (also corresponding to C_k), get the classification results D_k. Compare C_k and D_k.

Experimental parameters is set to: feature dimensions: 50; KNN adjacent values: 35; kernel function of SVM: radial basis function (RBF); training volume datasets: 1620; test set data: 1620; class number: 4; The number of ants: 350: Maximum speed: $\alpha=\beta=1$; attenuation coefficient: 0.1;the maximum number of iterations: 1500; the remaining the small sample number: 5.

Evaluation criteria.

Classification results are evaluated by macro average recall rate (MR), macro average precision (MP), micro averaged recall (mR), micro precision average (mP), macro average measure F value (MF) and micro average measure F value (mF). Classification results of Naive Bayesian (NB), k nearest neighbor algorithm (KNN), support vector machine (SVM) will be used as a benchmark for evaluating the effect of Text ACO-Miner.

Experimental results.

SVM, KNN, and NB classification models were selected to obtain the mean value of MF and mF. By the feature selection method, such as information gain, χ^2 statistics, mutual information, text evidence right, expected cross entropy, etc..In comparison, the effect of using the information gain on the data set is the best. Therefore, the feature selection method for the following classification models will adopt the information gain. See Table 1.

Table 1 MF and Mf mean values of each model

	information gain	χ^2 statistics	mutual information	text evidence right	expected cross entropy
SVM	0.930	0.920	0.366	0.860	0.924
KNN	0.922	0.909	0.618	0.854	0.905
NB	0.742	0.693	0.647	0.656	0.699

After training, Text ACO-Miner found a total of 30 classification rules. The confusion matrix of Text ACO-Miner classification results is shown in Table 2. ACO-Miner Text classification performance index values, as shown in Table 3.

Table 2 The confusion matrix of Text ACO-Miner

	Military	Sports	Transportation	Economy
Military	428	9	3	5
Sports	27	440	4	3
Transportation	14	14	293	8
Economy	22	18	10	322

Table 3 Classification performance index values of each model

	MR	MP	mR	mP	MF	mF
KNN	0.917	0.927	0.929	0.914	0.922	0.921
NB	0.757	0.814	0.768	0.642	0.784	0.699
SVM	0.932	0.929	0.937	0.922	0.930	0.930
ACO	0.927	0.916	0.926	0.917	0.921	0.922

Conclusion

In this paper, we develop a more mature ant colony algorithm in the field of text classification, and construct a text classification model based on ant colony intelligence. The results show that the Text ACO-Miner text classification model can be applied to text classification. Limited to the author's ability and time, the work of this paper is to introduce ant colony algorithm to text classification. The research is just a start, and a lot of work needs to be further studied, for example, the selection of the target function and the formulation of information update and group collaboration mechanism.

Acknowledgment

The research is financially supported by the National Natural Science Foundation of China (Grant No. 51209095, 51210013,51579105), the Fundamental Research Funds for the Central Universities (2014ZZ0027).

References

[1] Sebastiani F. "Machine learning in automated text categorization". ACM Computing Surveys, Vol.34,No.1,pp.1-47,March 2002.

[2] Colorni A, Dorigo M, Maniezzo V."Distributed optimization by ant colonies",Proceedings of European Conference on Artificial Life.Paris: Elsevier,pp. 134-142,1991.

[3] Dorigo M, Maniezzo V, Colorni A,"Ant system: optimization by a colony of coorperating agents",IEEE Trans on SMC,Vol.26,No.1, pp.29-41,1996.

[4] Parepinelli R S, Lopes H S, Freitas A A,"Data Mining with an Ant Colony Optimization Algorithm",IEEE Transaction in Evolutionary Computation,Vol.6,No.4,pp. 321-332,2002.

[5] Amioy Kumar, Ajay Kumar,"Adaptive Management of Multimodal Biometrics Fusion Using Ant Colony Optimization",Information Fusion, In Press, Available online 25 September 2015.

[6] B. Biswal, P.K. Dash, S. Mishra,"A hybrid ant colony optimization technique for power signal pattern classification",Expert Systems with Applications, Vol.38, Issue 5, pp. 6368-6375,May 2011.