ISBN 978-981-11-3671-9
Proceedings of 2017 the 7th International Workshop on Computer Science and Engineering (WCSE 2017)
Beijing, 25-27 June, 2017, pp. 233-238
doi: 10.18178/wcse.2017.06.040

Software Requirement Feature Selection Based On Tabu Search Algorithm

Tong Wang, Ying Shang [+]

College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029 China

Abstract. Software requirement classification plays a key role in test case reuse, and the most important part for requirement classification is the feature selection algorithm. In the classification process, the quality of the feature that been selected can seriously impact on the result of classification. In order to improve the feature selection, the tabu search had been used to apply different requirement feature combinations, which can effectively avoid getting in the local best solution, and finally obtain the optimal attribute subsets. In experiment, Chinese Corpus had been used to verify our algorithm and strategy, and results showed that, comparing with genetic algorithm, our method can effectively remove the invalid characters and improve classification accuracy, which means this method can ensure the accuracy of software requirement classification.

Keywords: software requirement classification, feature selection, tabu search

1. Introduction

Software requirements are the criteria for guiding the development and testing of software systems. To cover all software requirements, testers need to spend a lot of time designing complete test cases. Because of similar requirements between test cases, the software requirements can be classified, the same category of requirements are similar, which facilitate testing staff efficiently reuse test cases.

The description of software requirement is generally natural language, so the classification of software requirements is actually the classification of text. The key techniques of automatic classification of Chinese text include text preprocessing, text representation, feature reduction and text categorization algorithm [1]. The feature reduction is the key step of text categorization, which can improve the efficiency of text categorization and avoid the catastrophe of dimensionality. In this paper, the feature reduction is realized by studying the method of feature selection.

The main contributions of this paper include: (1) using TS algorithm to feature selection and improve the accuracy of text categorization; (2) The accuracy rate of text categorization can be improved after verifying the TS algorithm by experiment, and (3) verifying TS algorithm can be applied to software requirement feature selection through case study.

2. Related Work

Recently, some feature weight-based selection methods are as follows:

(1) Information gain (Information Gain, IG) [2]

Information gain is to calculate the existence or not of a characteristic word to the change of information entropy, it embodies the influence of the existence of a certain character word on the document category forecasting, so as to indicate the importance of the characteristic word.

[+] Corresponding author. Tel.: 15901115607;
E-mail address:shangy@mail.buct.edu.cn.

(2) Chi-square

Chi-square[3] is often applied to determine whether there is a correlation between two attributes or phenomena. When a text feature selection, a feature word is higher for a category Chi-square results, indicating the greater relevance of the feature word to that class, the more it represents the characteristics of the class. Chi-square are premised on the lack of independence between the feature word and the document category and the degree of freedom of 1 Chi-square distributions.

Considering that the feature selection process is the process of obtaining the optimal solution, in recent years, experts and scholars have begun to search the feature space by using heuristic search algorithm, thus obtaining the optimum solution [4]. Because of the traditional genetic algorithm (Genetic Algorithms, GA) has the versatility and high efficiency in obtaining the optimal solution problem, so it is originally applied to the feature selection [5], however, the traditional GA has a large amount of computational storage and cannot guarantee convergence to the global optimum solution [6], thus the characteristic subset of GA is sometimes not the optimal solution. In this paper, Tabu search (Tabu search, TS) algorithm is used to select feature, which avoid the search process into the local optimum, to obtain the characteristics of the optimal subset of the space, and improve the accuracy of software requirements classification.

3. Software requirement feature selection based on TS algorithm

3.1. Software requirements vectorization

The most commonly used software requirement representation model is a vector space model (VSM). Each requirement text can have the following representations [7]:

$$D = \{(t_1, w_1),\ (t_2, w_2),\ (t_3, w_3) \ldots\ (t_n, w_n),\} \tag{1}$$

Among them, t1, t2, t3...tn as the representative of the content of the requirement text features, W1, W2, W3 ...Wn is the weighting of the corresponding feature item, which is characterized by the preprocessing of the requriement text. Weights are computed by the TF-IDF algorithm [8].

3.2. The requirement text feature selection based on TS algorithm

(1) Encoding method

In this paper, binary encoding is used to show the selected feature in 1, and the characteristics are discarded by 0. For n-dimensional feature spaces, a feature combination is represented as a feature solution with 0 and 1 strings of length n.

(2) The initial solution

According to the encoding method mentioned above, the initial solution of the algorithm is a random generation of 0, 1 strings of length n.

(3) The fitness function

Considering the classification accuracy and the number of features, the following adaptation function is adopted [9]:

$$Fitness = \alpha \times M + \beta \times AVG + \gamma \times N \tag{2}$$

In formula(2): M is the number of error classifications; AVG is the average classification error rate obtained by using 10 crossover authentication; N is the feature selection after the number of characteristics, in addition α, β, γ is a constant parameter, the experiment sets them to 10,5,1.

(4) The generation of neighborhood solution

The generation of neighborhood solution is dependent on the current solution, for the current solution, by adding or decreasing a feature to produce its corresponding neighbor solution. Using formulas (2) to calculate the fitness of each neighborhood solution, select the fitness smallest solution as the best candidate solution.

(5) Taboo table and its length

This experiment adopts the fixed length taboo table; the taboo table length is set reasonable according to the solution space size. In the experiment, the data structure of taboo table selects the circular queue, which is to manage the taboo table with advanced first.

(6) Amnesty guidelines

The amnesty guidelines set out in this article are: If the best candidate solution is better than the current solution, then the best candidate solution chooses to accept and replaces the current solution, whether or not in the taboo table.

(7) Stop criterion

This article sets the algorithm maximum iteration number is 200 generations, simultaneously if the successive 10 generation solution does not change, the algorithm will also terminate.

3.3. The algorithm description

The algorithm for the above steps, and gives the method description of the TS algorithm for the feature selection of requirement text.

Algorithm 1 Requirement text feature selection based on TS algorithm

```
Input：Imax // maximum iteration
       Feather //feature number

Output：result // Termination solution
Set TabuList= Null//Init taboo table
curSolu = init(Feather)// Random Generation Initial Solution
fitness = fitFunction(curSolu)
do
  ++iterNum
  Neighbours = getNei(curSolu)
  fitness = fitFunction( Neighbours)
  bestNei= best( Neighbours，fitness)
  if bestNei is best so far
    curSolu = bestNei
  else
    curSolu = bestNotInTabuTable(Neighbours,fitness)
  update TabuList
until fitness is stable more than 10 times or iterNum >Imax
result = curSolu
return result
```

4. Experimental verification

4.1. Experimental corpus

In this study, we used Fudan University Chinese text categorization corpus, selected a total of 2716 documents in 10 categories of corpus, divided into training documents and test documents. The training documents and test documents were basically divided into 4: 1 ratio, each training category and the test document number is shown in table 1.

Table 1 each kind of documents

Text category	Number of training documents	Number of test documents
Environment	80	21
Computer	160	40
Traffic	162	52
Education	175	45
Economy	260	65
Military	199	50
Sports	360	90
Medicine	160	44
Artic	191	57
Political	400	105

4.2. Experimental parameter settings

In order to validate the validity of TS algorithm, the results of TS algorithm and GA are compared in the experiment process. The two algorithms set the parameters as follows, the TS algorithm set the taboo table length is 20, the maximum number of iterations is 200; In GA, the population size is 50, the maximum iteration number is 200, the crossover probability is 0.75, and the mutation probability is 0.05.

4.3. Classification algorithm selected by experiment and its evaluation criteria

The purpose of this experiment is to compare the effect of TS algorithm and GA on the accuracy of text classification after selecting feature. This experiment selects KNN as classifier, through the experiment to determine when the K value is 20 o'clock the classification effect is best, the spatial distance between the text is usually used to calculate the cosine similarity:

$$\cos\theta = \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{3}$$

In the formula, A and B are the two text vectors respectively, by calculating the angle cosine of two text vectors, when the cosine value is closer to 1, the smaller the distance between the two text, the greater the likelihood that the two texts belong to the same class, when the more the cosine value is closer to 0, the greater the distance between the two text, that is, the less likely that the two texts belong to the same class.

In order to evaluate the classification effect, in the experiment, the Precision (P), Recall (R) and F1 evaluation value are used as the evaluation criteria.

In addition, the above given P, R, and F1 evaluation values are for a category of classification results, and therefore also needs a global sense of evaluation methods. Because there is no difference between the weights of each category, the results are evaluated using macro averages.

4.4. Experimental results and analysis

The detailed results obtained by using TS algorithm and GA for feature selection are shown in the following figures.

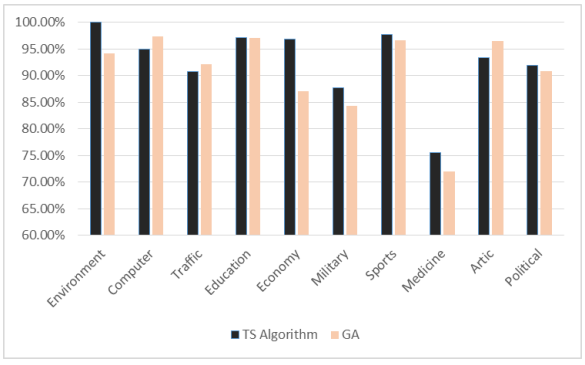

Fig. 1: Two kinds of algorithm contracts in precision

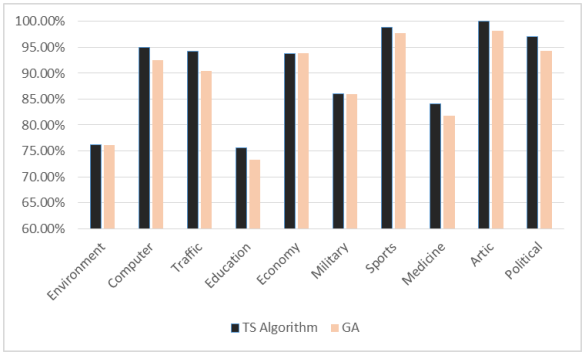

Fig. 2: Two kinds of algorithm contracts in recall

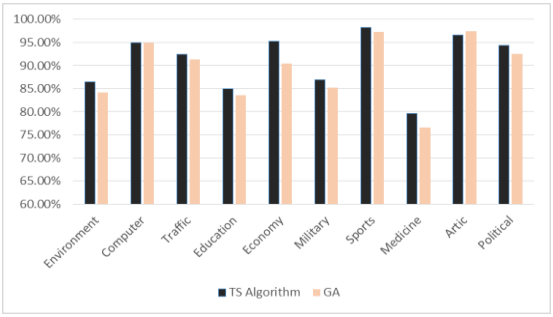

Fig. 3: Two kinds of algorithm contracts in F1 evaluation

By analyzing table 2, figure 1-3 can be drawn:

For the accuracy, compared with GA, TS algorithm has greatly improved in the environment, economy, military, medicine and four categories, the computer, transportation and art of these three categories have declined, and other categories have been raised less. For the recall rate, compared with GA, TS algorithm in addition to the environment, economy, military in three categories were flat, the other categories are ranging from the increase. For F1 valuation, the TS algorithm is improved in addition to the small range of the art categories, compared with GA.

The following is a general analysis of the impact of these two feature selection algorithms on the classification. The resulting data is shown below:

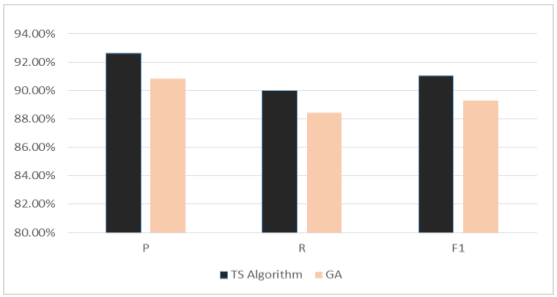

Fig. 4: Two kinds of algorithm contracts in F1 evaluation

From Fig. 4, the TS algorithm improves on the macro average accuracy, the macro-average recall rate and the macro average F1 evaluation, compared with GA.

To sum up, consider from the category, compared with GA, TS algorithm is basically improved on accuracy rate, recall rate and F1 valuation, and only decreases in individual categories. As a whole, the TS algorithm is superior to GA on accuracy rate, recall rate and F1 valuation.

5. Case study

In order to verify that the feature selection method based on Tabu search algorithm can be applied to software requirement classification; this paper selects some software requirements of a shopping site for case study.

5.1. Case study object

Case study object is a shopping web site software requirement, the software requirement needs to collate, to the functional requirements of the functional categories, will contain too little requirement for the category of removal, and finally constitute the case study object. The case study dataset is shown in table 2.

Table2: Case study data set

Requirement category	Number of training documents	Number of test documents
Shopping process	4	1
Customer member	4	1
Background manage	4	1
Merchandise show	4	1

In order to facilitate the software requirements description of the intuitive impression, enumerating the customer membership category of a requirement as follows: "Information into the database member, login to

the mall, through the page above the navigation bar ' personal information maintenance ' into the Information maintenance page, testing can modify personal information." ”

5.2. Case study process

The case study process is as follows: (1) The software requirement is preprocessed, the feature space is arranged. (2) Calculating the TF-IDF value of each software requirement, and each requirement is quantified. (3) The TS algorithm is used to select the software requirement, and the optimum collection is obtained. (4) The software requirement is re-quantified according to the feature selection result and the test set is used to test classification effect.

5.3. Case study Results and analysis

Table 3 results from the research results, which can be found that the TS algorithm is used to select the feature, the accuracy rate, the recall rate and the F1 evaluation values are up to 100%, which explains the feasibility of the method applied to the software requirement feature selection.

Table3: The result of case study

Requirement category	**TS algorithm**		
	P/%	R/%	F_1/%
Shopping process	100	100	100
Customer member	100	100	100
Background manage	100	100	100
Merchandise show	100	100	100

6. Conclusion

In this paper, a method for selecting the software requirement feature based on TS algorithm is proposed, and the optimal subset is obtained by using TS algorithm to evolve the given initial solution. Through the experiment, the TS algorithm compares with the GA, besides the individual text category, the text classification accuracy rate obtains the effective enhancement, through the example research, adopts the TS algorithm to carry on the software requirement characteristic choice, the classification effect is comparatively ideal. The next step is to apply this method to a large number of enterprise requirements, and to guide test case reuse based on the requirements classification.

7. References

[1] Javed K, Maruf S, Babri H A. A two-stage Markov blanket based feature selection algorithm for text classification[J]. Neurocomputing, 2015, 157: 91-104.

[2] Lee C, Lee G G. Information gain and divergence -based feature selection for machine learning -based text categorization[J]. Information processing & management, 2006, 42(1): 155-165.

[3] Moh'd A Mesleh A. Chi square feature extraction based SVMs Arabic language text categorization system[J]. Journal of Computer Science, 2007, 3(6): 430-435.

[4] Ghamisi P, Benediktsson J A. Feature selection based on hybridization of genetic algorithm and particle swarm optimization[J]. IEEE Geoscience and Remote Sensing Letters, 2015, 12(2): 309-313.

[5] Leung Y W, Wang Y. An orthogonal genetic algorithm with quantization for global numerical optimization[J]. IEEE Transactions on Evolutionary computation, 2001, 5(1): 41-53.

[6] Han K H, Park K H, Lee C H, et al. Parallel quantum-inspired genetic algorithm for combinatorial optimization problem[C]. Evolutionary Computation, 2001. Proceedings of the 2001 Congress on. IEEE, 2001, 2: 1422-1429.

[7] Shon T, Kim Y, Lee C, et al. A machine learning framework for network anomaly detection using SVM and GA[C]. Proceedings from the Sixth Annual IEEE SMC Information Assurance Workshop. IEEE, 2005: 176-183.

[8] Domeniconi G, Moro G, Pasolini R, et al. A study on term weighting for text categorization: a novel supervised variant of TFIDF[C]. Proceedings of the 4th international conference on data management technologies and applications (DATA). Candidate to the best conference paper award. 2015: 26-37.

[9] Tahir M A, Bouridane A, Kurugollu F, et al. Feature selection using tabu search for improving the classification rate prostate needle biopsies[C]. Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. IEEE, 2004, 2: 335-338.