A Feature Selection Method Based on Genetic Algorithms

Mingyang Jiang
College of Computer Science and Technology
Inner Mongolia University for the Nationalities
Tongliao, China
e-mail: jiang_ming_yang@163.com

Yuxin Zhou
College of Computer Science and Technology
Inner Mongolia University for the Nationalities
Tongliao, China

Xiaojing Fan
College of Mechanican Engineering
Inner Mongolia University for the Nationalities
Tongliao, China

QiangHu Wang
College of Computer Science and Technology
Inner Mongolia University for the Nationalities
Tongliao, China

Xinhong Zhang
Department of Neurology
The Affiliated Hospital of Inner Mongolia University
for Nationalities
Tongliao, China

ZhiFeng Zhang
College of Computer Science and Technology
Inner Mongolia University for the Nationalities
Tongliao, China

Lian Jie
College of Computer Science and Technology
Inner Mongolia University for the Nationalities
Tongliao, China

Zhili Pei*
College of Computer Science and Technology
Inner Mongolia University for the Nationalities
Tongliao, China
e-mail: zhilipei@sina.com

Abstract—Feature extraction technology is a major factor in determining good classification results, the traditional feature extraction method has many deficiencies, such as when a high degree of imbalance in the distribution of the categories and characteristics, it can not effectively deal with low-frequency words; single feature for improper handling, leading to local optima generating solution. For traditional feature extraction methods can not fully and effectively examine the shortcomings of the candidate feature words, proposed a text feature extraction method based on genetic algorithm. In this method, a variety of heuristics word frequency, correlation, part of speech, and location to be elected to the comprehensive test features, and to optimize the weight parameter for each heuristic using genetic algorithms. By comparing the different test sets, the experimental results show that, compared with traditional methods, this method can effectively avoid the traditional feature extraction method produces bias, obtain a representative set of features, making this method has some practical value.

Keywords-feature extraction technology; classification; genetic algorithms; word frequency; feature set

I. INTRODUCTION

Complete text classification requires four steps: text preprocessing, text to vector space model to represent extracted feature words, the establishment of an appropriate classifier. As one of the most critical aspects of text classification, feature extraction, can play to reduce the vector dimension, eliminating noise and simplify the calculation and so on. So, it would have the effect of text categorization impact indicators such as accuracy. First construct an evaluation function, and then calculating the weight value of this function characterized in terms of the characteristic words and then re-right sorted values, select the rake preset characteristic feature of the final subset of the set of words, the above is the basic idea of feature extraction[1][2]. Feature extraction is to extract metrics extracted from the initial feature set in the relevant test set initial feature subset of the original feature vector space to reduce the dimension of purpose according to certain characteristics.

Not relevant in the process of feature extraction and redundancy features will be deleted. Feature Extraction Method for processing data as the learning algorithm can well improve the accuracy of the learning algorithm, reducing the time spent learning algorithm. If the learning algorithm can learn to use the full features of irrelevance, redundancy, and interference, then the learning algorithm result will be poor. In practical applications, feature extraction how to get an optimal feature subset is an NP problem.

II. FEATURE EXTRACTION

Feature extraction can be extracted in accordance with the policy features, feature drop-dimensional model distinction.

A. *Feature Extraction Strategy*

By feature extraction strategy, feature extraction methods can be divided into global and local features extracted feature extraction. Faced with different data sets, we should choose the appropriate feature extraction strategies.

Local feature extraction is based on the characteristics of a given evaluation function for all the features present in each category, respectively, at the statistical value of each category of each local feature feature space entry, will each feature items sorted by their eigenvalues and Select the appropriate threshold is set according to the size of the category feature subset, then each category feature set combined into a new feature space, the feature space is a collection of various categories of feature subset. Local feature extraction is based on a given feature selection metrics were selected optimal feature set for each specific category, the main consideration local characteristics of this class of common and internal characteristics[4-6].

Global feature extraction based on the characteristics of a given evaluation function for all the features present in each category is calculated, and then combined with the corresponding values for each category of each feature by the global value calculation method to calculate a unique global values for each feature, the various features of the item in its global eigenvalues sort and select the appropriate size of the global feature subset based on a set threshold[7]. Global feature extraction is related to the classification of all categories to choose a common optimal feature set, the main consideration differences between the categories.

References 8 noted in a balanced global data feature selection is better than the local feature selection, because the use of local feature extraction alone drawback is that local feature extraction can not effectively choose to be able to represent all categories of feature set, ignoring all the training samples and the overall category. References 9 noted skewed data to use in the local feature selection is better than the weighted global features, because a lot of the skew between data gap in the number of documents in each category, based on the global selection feature set does not represent a minority class, if you use non-weighted feature selection from Select the same number of each category will also affect the characteristics of feature selection results.

B. *Feature Dimension Reduction Model*

According to a feature dimension reduction model can be divided into extraction based Filter feature extraction model, Wrapper features of the model and hybrid methods. Filter-based feature extraction methods and differences between model-based feature extraction method Wrapper model is taken to evaluate whether specific classification criteria and relevant.

Filter model feature extraction based on the calculation method illustrated in Fig .1 is dependent on the sample data set itself and adopt a classification independent evaluation criteria, the evaluation criteria of the evaluation function is called characterized by all the features of the feature selection method was filtered, filter out redundant, irrelevant characteristics, retention and category-related characteristics, the correlation is generally considered a subset of a larger feature will benefit the accuracy of the learning algorithm, so the objective function related to the degree of choice and a large subset of features, the use of features feature selection method selected from the collection of said samples in order to carry out training of classifiers.

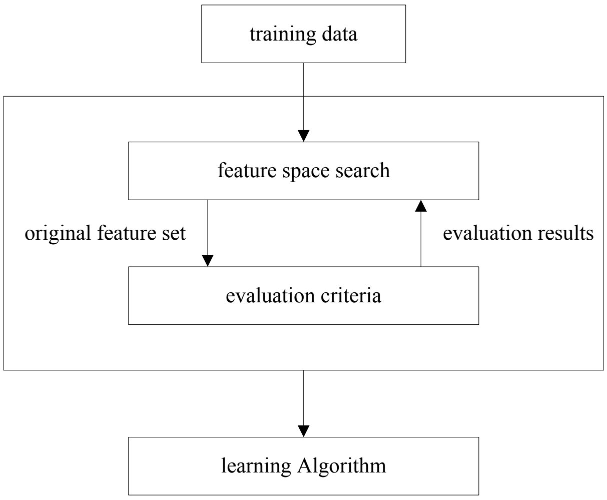

Figure 1. The Filter Model of Feature Selection

According to the evaluation function characteristic features of the pros and cons of different metrics, feature extraction methods can be divided into: Based on the distance measure, the correlation measure, measure the amount of information and consistency metric feature extraction method. Feature extraction method based on the distance metric aim is to find a subset of features can make the maximum distance between each class so that the largest category of separability, the lowest classification error rate. Distance measure used mainly Euclidean distance, all the theoretical basis of the evaluation criteria are based distance metric best.

III. GENETIC ALGORITHM

Genetic algorithms are stochastic, iterative and evolutionary process of natural selection and genetics of populations on the basis of its main features is to take groups of search strategies and between groups of individuals to exchange information, using simple coding techniques and propagation mechanisms to the performance of complex phenomena, not restrictive hypothesis search space constraints, without continuity, derivative, and the presence of a single peak and other assumptions. Genetic algorithm optimization is now in the areas of machine learning and parallel processing has been more widely used[10-12].

A. *Principles of Genetic Algorithms*

Genetic algorithm is an iterative algorithm that from a solution or a specific set of randomly generated initial departure, according to certain operating rules, such as selection, reproduction, crossover and mutation, constantly iterative calculations to get the next generation solution set , and according to the fitness of the individual, in accordance with the principles of survival of the fittest and survival of the fittest, and guide the search process to the "most adapt to the environment," the individual approach, each generation evolved better and better approximate solution, eventually converge to the optimal solution or satisfactory solution[13].

B. The Basic Steps of Genetic Algorithm

1. N individuals constitute a randomly generated initial population.

2. Calculate the fitness of each individual of the current population according to the fitness function.

3. Algorithm to determine whether the termination condition is satisfied, if satisfied then turn the eighth step.

4. Select the function to perform operations in accordance with the fitness of each individual.

5. Press crossover probability Pc perform crossover operation.

6. By mutation probability Pm perform mutation.

7. If you have got a new generation of groups of individuals constituted by N, then turn the second step, otherwise turn fourth.

8. Output Results terminated.

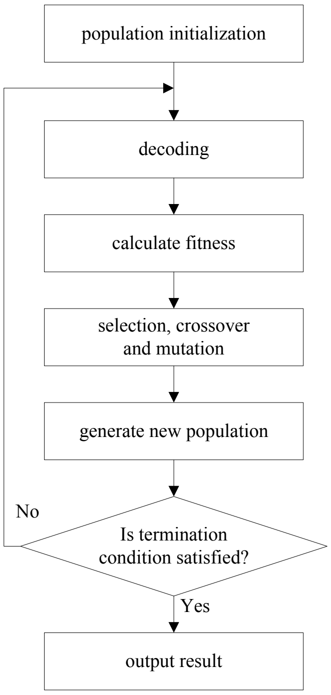
population initialization
decoding
calculate fitness
selection, crossover and mutation
generate new population
No
Is termination condition satisfied?
Yes
output result

Figure 2. Genetic Algorithm Flowchart

IV. Feature Selection Method Based on Genetic Algorithms

A. Feature Extraction Process

1. Pretreatment: Removal of network text in HTML format, the location information is retained text words, and text segmentation and POS tagging.

2. Heuristic calculations: Computing text TFIDF, correlation, and other heuristics for each location and POS words.

3. Heuristic integration: According to multiple heuristics fusion model, four heuristic words are fused, and calculate the composite score.

4. Output: Finally, sorted according to the size of each feature score, and select the optimum output characteristic.

B. Genetic Algorithm to Optimize Weight Parameters

Because of the genetic algorithm is simple, easy to understand, easy to implement, and in solving combinatorial optimization problems have a strong advantage, therefore, the paper uses genetic algorithms to optimize the parameters of weight, resulting in the best combination of a range of parameter weights. Four parameters defining the weights in the range of (0,1), and their value to 1. Then select the appropriate initial value based on experience, and through iterative calculation to obtain the right parameters for each heuristic weight. Use of genetic algorithm to obtain the weight of each characteristic parameters specific process is described as follows:

1. Initializing each characteristic parameter weighting, a = 0.3, b = 0.2, c = 0.2, d = 0.4.

2. Coding: The algorithm uses the decimal coding chromosome encoded.

3. Corresponding recall rate was calculated using the weights of each parameter to recall the fitness function as a chromosome, the recall rate is calculated as rec=n/N, n represents the number of features consistent with the marked characteristics, N represents the total number of features centralized document marked.

4. Crossover and mutation: genetic algorithm convergence speed and quality of the solution depends largely on the crossover and mutation probabilities. In order to prevent trapped in local optimum algorithm and accelerate algorithm search efficiency, optimum population only allow individuals to participate in crossover and mutation, and the current population is not the best individual participation.

5. Termination condition: Poor populations Best Contemporary chromosome fitness value and the previous generation of the population of the best fitness value of chromosomes absolute value is not more than 10^{-5}. Select parameters using genetic algorithm to optimize the heavy weight of each heuristic can effectively avoid subjectivity is determined by the subjective experience of the parameters in order to achieve adaptively tuning parameters based on the training data. The results show that the experiments below, the use of genetic algorithms to obtain the parameters of the weighting feature extraction method herein allows to obtain a good extraction.

C. Analysis of Test Results

Experimental data using the Reuters-21578 classification corpus as experimental data, it's downloaded at: http://www.daviddlewis.com/resources/testcollections/reuters21578 。 A test collection for text categorization contains, at minimum, a set of texts and, for each text, a specification of what categories that text belongs to. For the Reuters-21578 collection the documents are Reuters newswire stories, and the categories are five different sets of content related categories. For each document, a human indexer decided which categories from which sets that document belonged to. The category sets are as follows:

TABLE I. CATEGORY SETS

Category Set	Number of Categories
EXCHANGES	39
ORGS	56
PEOPLE	267
PLACES	175
TOPICS	135

Experiments based on the frequency of feature extraction methods, extraction methods and performance of the proposed method were compared based on the characteristics associated degree.

TABLE II. COMPARATIVE RESULTS

methods	rec
TF-IDF	77.59
CF	82.96
Proposed algorithm	85.17

V. SUMMARY

Link between the proposed method can effectively use words and words between the intrinsic properties, characterization Chinese texts through a variety of heuristic features, the feature word more comprehensive test. The experimental results show that this method can effectively integrate the advantages of different factors, compared with the traditional method, this method has some advantages, making this method has some practical value in terms of text mining.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (61163034, 61373067), the Grassland Excellent Talents Project of Inner Mongolia Autonomous Region(2013), the supported By Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region(NJYT-14-A09), the Inner Mongolia Natural Science Foundation(2013MS0911), the 321 Talents Project the two level of Inner Mongolia Autonomous Region(2010), the Inner Mongolia talent development fund(2011), the First Trade Leading Talent of Tongliao, and the Science Research Project of Inner Mongolia University(NMD1229, NMD1322). The Corresponding author is Zhili Pei, and the Corresponding email is zhilipei@sina.com.

REFERENCES

[1] LI Gang, DAI Qiangbin. Keywords auto matic indexing based on lexical chains [J]. Document, InformationandKnowl-edge, 2011,12 (3): 67-71

[2] ZHU Haodong, LI Hongchan. Feature selection based on mutual information and rough set theory [J]. ComputerEn-gineering, 2011,37 (15): 181-183.

[3] JAVED K, BABRI H A, SAEED M. Feature selection based on class-dependent densities for high-dimensional binary data [J]. IEEE Transactions on Knowledge and Data Engineering, 2012,24 (3): 465-477.

[4] GHEYAS I A, SMITH L S. Feature subset selection in large dimensionality domains [J]. Pattern Recognition, 2010,43 (1): 5-13.

[5] Iio Jun. Basic techniques in text mining using open-source tools. Proceedings of the 9th International Symposium on Open Collaboration, WikiSym + OpenSym. [C] 2013: 175-182

[6] Da Costa Pinho Isis, Epstein Daniel, Reategui Eliseo Berni, Corrêa Ygor. The use of text mining to build a pedagogical agent capable of mediating synchronous online discussions in the context of foreign language learning. 43rd IEEE Annual Frontiers in Education Conference, FIE 2013[C]. 2013: 393-399

[7] Samhaa R. El-Beltagy,Ahmed Rafea. KP-Miner: A keyphrase extraction system for English and Arabic documents[J]. Information Systems . 2008 (1)

[8] BONG C H, NARAYANAN K. An empirical study of feature selection for text categorization based on term weightage[C]: Proceedings of the 2004 IEEE/W IC/ACM International Conference on Web Intelligence. Washington, DC: [s. n.]: IEEE Computer Society , 2004: 599-602.

[9] Soucy, P.&Mineau, G.W. Feature Selection strategies for text categorization[C]: Proceedings of the 16th Conference of Canadian Conference on AI, . Halifax, Canada: the Canadian Society for Computational Studies of Intelligence', Vol. 2671 of Lecture Notesin Computer Science, Springer-Verlag New York, Inc, 2003: 505–509.

[10] C.D. Manning,P. Raghavan,P.,H. Schutze.Introduction to Information Retrieval. Journal of Women s Health . 2008

[11] Computer and telecommunications Xiaoyan Zhang, Ying Hua. Text mining methods and applied research [J]. 2011(12):68-69

[12] Salton G,Wong A,Yang CS.A vector space model for automatic indexing. Communications of the ACM . 1975

[13] Fang Li.Text mining of certain key technologies [D]. Beijing University of Chemical Technology .2010.

[14] Carter P H. The Rocchio classifier and second generation wavelets. In: Proceedings of the International Society for Optical Engineering. Orlando, USA: SPIE, 2007: 1-11.