Optimal Feature Selection Algorithm Based on Quantum-Inspired Clone Genetic Strategy in Text Categorization

Hao Chen
School of Information Science and Engineering
Central South University
Hunan, China
+86 731 8617575
xschenhao@gmail.com

Beiji Zou
School of Information Science and Engineering
Central South University
Hunan, China
+86 731 8877701
bjzou@vip.163.com

ABSTRACT
Information overload is a serious issue in the modern society. As a powerful method to help people out of being "lost" in too much useless information, automatic text categorization is getting more and more important. Feature selection is the most important step in text categorization. To improve the performance of text categorization, we present a new text categorization method called quantum-inspired clone genetic algorithm (QCGA). The experimental results show that the QCGA algorithm is superior to other common methods.

Categories and Subject Descriptors
I.7 DOCUMENT AND TEXT PROCESSING; E.2 DATA STORAGE REPRESENTATIONS, Contiguous representations**

General Terms
Algorithms, Documentation

Keywords
Genetic algorithm , quantum-inspired clone genetic algorithm, text categorization, feature selection

1. INTRODUCTION
Due to the rapid growth in textual data, automatic methods for organizing the data are needed. Automatic text categorization is one of these methods, it automatically assigns the documents to a set of predefined classes based on their textual content, and whose main goal is to reduce the considerable manual process required for the task. Text categorization is a crucial and well-proven instrument for organizing large volumes of textual information.

Feature selection (FS) is a commonly used step in machine learning, especially when dealing with a high dimensional space of features [1]. The objective of feature selection is to simplify a dataset by reducing its dimensionality and identifying relevant underlying features without sacrificing predictive accuracy. FS is extensive and it spreads throughout many fields, including text categorization, data mining and pattern recognition.

In this paper, we propose a quantum-inspired clone genetic algorithm (QCGA) to the problem of FS in text categorization. Proposed algorithm is applied to text features of bag of words model in which a document is considered as a set of words or terms and each position in the input feature vector corresponds to a given term in original document. The organization of the remaining content is as follows. Section 2 presents the related works. Section 3 describes the QCGA algorithm. Section 4 reports computational experiments and finally the conclusion and future works are offered in the last section.

Copyright is held by the author/owner(s).
GEC'09, June 12–14, 2009, Shanghai, China.
ACM 978-1-60558-326-6/09/06.

2. RELATED WORKS
Recently, text categorization has become a key technology to deal with and organize a large number of documents. Several approaches are applied to the problem of FS in text categorization. Genetic algorithm (GA) is optimization techniques based on the mechanism of natural selection. They used operations found in natural genetics to guide itself through the paths in the search space. Because of their advantages, recently, GA has been widely used a tool for FS in data mining. GA attempts to achieve better solutions by application of knowledge from previous iterations [2].

Applying GA to the FS problem is straightforward: the chromosomes of the individuals contain one bit for each feature, and the value of the bit determines whether the feature will be used in the classification. Using the wrapper approach, the individuals are evaluated by training the classifiers using the feature subset indicated by the chromosome and using the resulting accuracy to calculate the fitness. In [3], GA is due to find an optimal binary vector in which each bit corresponds to a feature. A '1' or '0' suggests that the feature is selected or dropped, respectively. The aim is to find the binary vector with the smallest number of 1's such that the classifier performance is maximized. This criterion is often modified to reduce the dimensionality of the feature vector at the same time. Punch, W.F. et al. use an m-ary vector to assign weights to features instead of abruptly dropping or including them as in the binary case [4]. Adriana, P. et al. presents a GA, called Olex-GA for the induction of rule-based text categorization [5]. GA have been used to search for feature subsets in conjunction with several categorization methods such as neural networks, , and k-nearest neighbors [6]. Besides selecting feature subsets, GA can extract new features by searching for a vector of numeric coefficients that is used to transform linearly the original features. In this case, a value of zero in the transformation vector is equivalent to avoiding the feature. Raymer et al. combined the linear transformation with

explicit FS flags in the chromosomes [7], and reported an advantage over the pure transformation method. However, these papers focused on FS and did not deal with attribute reduction and parameter optimization for the algorithm. Also, most of the proposed FS techniques are not designed to handle multiple selection criteria such as classification accuracy, feature measurement cost.

Although GA has a powerful quality of global search, it is liable to raise the problem of prematurely convergence in the practical application, and has low search efficiency in the late evolving period [8]. Clonal selection genetic (CSA) establishes the idea that the cells are selected when they recognize the antigens and proliferate. When exposed to antigens, the immune cells which may recognize and eliminate the antigens can be selected in the body and mount an effective response against them. Quantum computing (QC) is based on the concepts of qubits and superposition of states of quantum mechanics. QC can represent a linear superposition of solutions due to its probabilistic representation. Thus, we propose to abstract the merit of GA, CSA and QC, and a FS method in text categorization, namely quantum-inspired clone genetic algorithm (QCGA) is proposed. The experiments demonstrate that, in most cases, the proposed QCGA finds subsets that result in the best accuracy, while finding compact feature subsets, and performing faster than other common methods.

3. THE APPLICATION OF QGCA

FS is one of the applications of subset problems. Given a feature set of size n, the FS problem is to find a minimal feature subset of size $k(k<n)$ while retaining a suitably high accuracy in representing the original features.

Generally, a text categorization system consists of several essential parts including feature extraction and feature selection. After preprocessing of text documents, feature extraction is used to transform the input text document into a feature set(feature vector).FS is applied to the feature set to reduce the dimensionality of it. This process is shown in Fig.1.QGCA is used to explore the space of all subsets of given feature set. The performance of selected feature subsets is measured by invoking an evaluation function with the corresponding reduced feature space and measuring the specified classification result. The best feature subset found is then output as the recommended set of feature to be used in the actual design of the classification system.

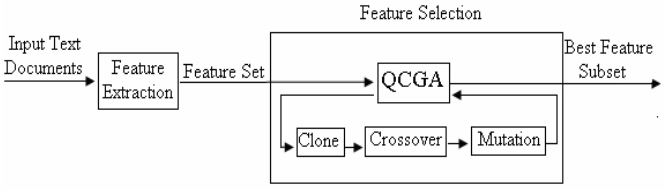

Fig. 1. The process of feature selection in text categorization

The main steps of proposed feature selection algorithm are as follows:

Step 1: Generate an initial population $A(t)$, and set parameters of algorithm, $t=0$;

Step 2: Evaluate the affinity of initial population;

Step 3: Perform the clone, crossover and mutation for $A(t)$ to generate $A(t+1)$;

Step 4: $t=t+1$;if stopping condition is satisfied, then output the best feature subset; otherwise,go back to step 2.

In QCGA, each antibody in the population represents a candidate solution to the feature subset selection problem. Let m be the total number of features available to choose from to represent the patterns to be classified. Given m such features, there exist 2^m possible feature subsets. Thus, for large values of m ,an exhaustive search is not feasible. Each feature subset is represented by a binary vector of dimension m .If a bit is a 1, it means that the corresponding feature is selected. A value of 0 indicates that the corresponding feature is not selected.

Classification accuracy and the feature cost are the two criteria used to design an affinity function. Thus, for the individual (antibody) with high classification accuracy and low total feature cost produce a high affinity value. We solve the multiple criteria problem by creating a single objective affinity function that combines the two goals into one. The antibody with high affinity value has high probability to be preserved to the next generation. We designed an affinity function as follows:

$$f(x)=accuracy(x)-\frac{F(x)\times cost(x)}{accuracy(x)+1}+cost_{\max} \quad (1)$$

Here, $f(x)$ is the affinity function of the feature subset represented by x ; $accuracy(x)$ is the test accuracy; $cost(x)$ is the sum of measurement costs of the feature subset represented by x ; and $cost_{\max}$ is an upper bound on the costs of candidate solutions. In this case, $cost_{\max}$ is simply the sum of the costs associated with all of the features. $F(x)=1$ represents that feature x is selected; otherwise, $F(x)=0$ represents that feature x is not selected.

Clone is the process of antibody proliferation [9]. A new population is generated by selecting the best antibodies from the population, these antibodies will be cloned according to its affinities, after clone process, a temporary population is generated. Antibodies are sorted by descending order in our algorithms. The number of clones reproduced for each antibody is proportional to its affinity with the antigen. The number of clones is given as follows:

$$N_c=round(\frac{\beta\times N}{i}) \quad (2)$$

where N_c is the number of clones generated for each antibody, β is a multiplying factor, N is the total number of antibodies, i is the index of current antibody in the population, and $round(*)$ is the function that it rounds its variable toward the closest integer. After the clone step, the population becomes:

$$B(k)=\{A(k),A_1'(k),A_2'(k),...,A_i'(k),...,A_n'(k)\} \quad (3)$$

where $A_i^{'}(k)=\{a_{i1},a_{i2},...,a_{ij},....,a_{iN_C-1}\}, a_{ij}=a_i, j=1,2,...,N_C$

In QCGA, one point crossover operator is used for qubit, which is illustrated as follows. In particular, one crossover position is randomly determined (e.g. position i), and then the qubit of the parents before position i are reserved while the qubits after position i are exchanged.

$$\begin{bmatrix}\alpha_1 ... & \alpha_i & \alpha_{i+1} & ... & \alpha_m \\ \beta_1 ... & \beta_i & \beta_{i+1} & ... & \beta_m\end{bmatrix}\Rightarrow\begin{bmatrix}\alpha_1 ... & \alpha_i & \alpha_{i+1}^{'} & ... & \alpha_m^{'} \\ \beta_1 ... & \beta_i & \beta_{i+1}^{'} & ... & \beta_m^{'}\end{bmatrix}$$

$$\begin{bmatrix}\alpha_1^{'} ... & \alpha_i^{'} & \alpha_{i+1}^{'} & ... & \alpha_m^{'} \\ \beta_1^{'} ... & \beta_i^{'} & \beta_{i+1}^{'} & ... & \beta_m^{'}\end{bmatrix}\Rightarrow\begin{bmatrix}\alpha_1^{'} ... & \alpha_i^{'} & \alpha_{i+1} & ... & \alpha_m \\ \beta_1^{'} ... & \beta_i^{'} & \beta_{i+1} & ... & \beta_m\end{bmatrix}$$

The mutation operator helps prevent early convergence of the QCGA by changing characteristics of antibody in the population. Such changes in the antibody also results in the QCGA ability to 'jump' to far away solutions, hopefully to unexplored areas of the solution space. In QCGA, mutation for qubit is illustrated as follows. In particular, one position is randomly determined (e.g. position i), and than the corresponding α_i and β_i are exchanged

$$\begin{bmatrix}\alpha_1 \alpha_2 & ... & \alpha_i & ... & \alpha_m \\ \beta_1 \beta_2 & ... & \beta_i & ... & \beta_m\end{bmatrix}\Rightarrow\begin{bmatrix}\alpha_1 \alpha_2 & ... & \beta_i & ... & \alpha_m \\ \beta_1 \beta_2 & ... & \alpha_i & ... & \beta_m\end{bmatrix}$$

4. EXPERIMENTAL RESULTS

A series of experiments was conducted to show the utility of proposed FS algorithm. All experiments have been run on a machine with 2.33GHz Xeon CPU and 2GB of RAM. We implement proposed QCGA algorithm and other three FS algorithms in Matlab7.0. We adjusted the parameters of the QCGA by experiments, and finally selected the following combination of the parameters: the maximum number of generations $MaxGen=150$, $k=5$, $n=120$, and the size of the population is 100.

4.1 DATASET

To provide an overview on the base line accuracy of the classifiers and to compare them with various studies, the Reuters collection was taken in our experiments. The Reuters collection consists of stories from Reuter's news agency which classified under categories related to economics, and accounts for most of the experimental works in text categorization so far. We used Reuters-21567.In Reuters-21567 dataset, we adopt the top ten classes; 5213 documents in training set and 2016 documents in test set. The maximum class has 2096 documents, occupying 40.2% of training set. The minimum class has 68 documents, occupying 1.3% of training set. Table 1 shows the ten most frequent categories along with the number of training and test example in each.

Table 1 Number of train/test documents

NO.	Category name	Number of train	Number of test
1	Agent	1286	512
2	Field	160	62
3	Meta	318	106
4	Method	68	37
5	Money	2096	798
6	Set	153	72
7	Term	324	96
8	Text	126	45
9	Task	248	98
10	World	434	190

4.2 SIMULATION

In most text categorization, we are interested in the performance of the proposed feature selection techniques. Several measures such as precision and recall are used to evaluate the performance of feature selection algorithm. Precision is defined as the ratio of correct topic cases to the total predicted topic cases. Recall is defined as the proportion of the correct topic cases to the total cases. They showed in the following equations:

$$Precision(i)=\frac{TP_i}{TP_i+FP_i} \quad (4)$$

$$Recall(i)=\frac{TP_i}{TP_i+FN_i} \quad (5)$$

where TP_i is the number of test documents correctly classified under ith category(C_i). FP_i is the number of test documents incorrectly classified C_i ,and FN_i is the number of test documents incorrectly classified under other categories these probabilities may be estimated in terms of the contingency table for C_i .

Fig.2-4 shows the performance of our proposed method against the GA in [5] and Olex-GA in [7] for the ten most frequent categories. The precision of GA, Olex-GA and QCGA is shown in Fig.2. The recall of the GA, Olex-GA and QCGA is shown in Fig.3. The antibody of the GA and QCGA is shown in Fig.4.As the generation continues, further improvement is found in average population affinity as demonstrated in Fig. 4.

The average precision for GA, Olex-GA and QCGA are 70.9%, 82.8% and 91.4% respectively. With the exception of categories Method, the precision of each category for QCGA is higher than GA and Olex-GA. This indicates that the QCGA algorithm perform at generally high precision. The average recall results for the GA, Olex-GA and QCGA are 81.3%, 90.2%and 96.1%, and 96.1% respectively. The QCGA can classify documents into the correct category mapping to precision, with a high recall ratio. Analyzing the precision and recall show in Fig.2-3, we see that on average, the QCGA algorithm obtain a higher accuracy value than the GA and Olex-GA.

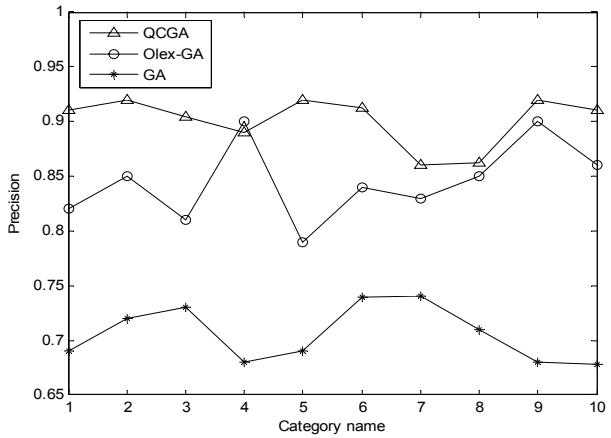

Fig. 2 The precision of the GA, Olex-GA and QCGA

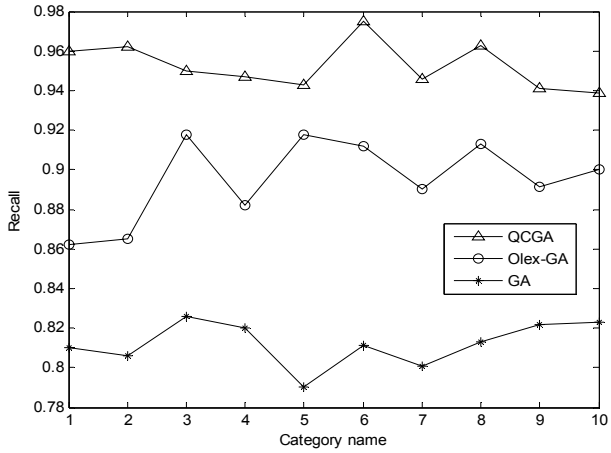

Fig. 3 The recall of the GA, Olex-GA and QCGA

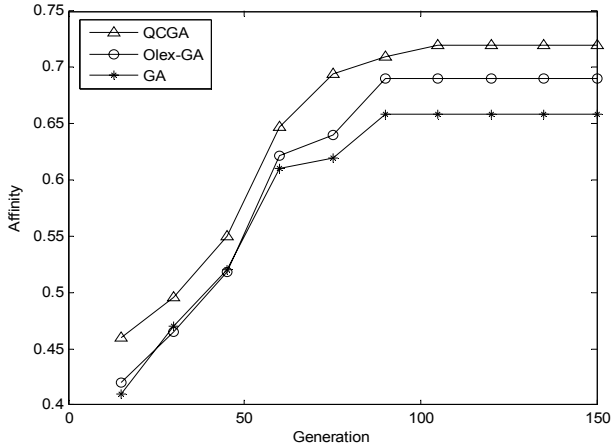

Fig. 4 The affinity of GA, Olex-GA and QCGA

In Fig.4, the simulated result shows that the average affinity of GA, Olex-GA and QCGA are 0.66, 0.69, and 0.723 respectively. This indicates that the QCGA yields a better classification result than the other two methods. Because QGCA can guide search to the optimal minimal subset every time.

5. CONCLUSIONS

With the rapid development of the online information, text classification becomes one of the key techniques for handling and organizing the text data. In this paper, we have proposed a text classification method using a QCGA algorithm. The experimental results show that the QGCA yields the best result of these three methods. The experiment also demonstrated that the QGCA yields better accuracy even with a large data set since it achieved better performance with the lower number of features. In future research, we intend to investigate the performance of proposed feature selection algorithm by taking advantage of using more complex classifiers in that. Another research direction will involve experiments with other kinds of datasets.

6. ACKNOWLEDGMENTS

This paper is partly supported by Major Program of National Natural Science Foundation of China(Grant No. 90715043).

7. REFERENCES

[1] Kim, H., Howland, P. Dimension reduction in text classification with support vector machines, Journal of Machine Learning Research, 2005, 12(6):37-53.

[2] Jorng,T.H, Ching,C.Y. Applying genetic algorithms to query optimization in document retrieval[J].Information Processing and Management,2000(36): 737-759.

[3] Yang.Y, Honavar, V. Feature subset selection using a genetic algorithm. IEEE Intelligent Systems, 1998, 13(2):44-49.

[4] Punch, W.F., Goodman, E.D Further research on feature selection and classification using genetic algorithm. Proceedings of the Fifth International Conference on Genetic Algorithm, San Mateo, CA, Morgan Kaufmann,1993,pp:557-564.

[5] Adriana, P., Veronica L.P.A Genetic Algorithm for Text Classification Rule Induction [J]. Berlin, Heidelberg: Springer-Verlag, LNAI 5212, 2008:188-203.

[6] Kudo, M.Sklansky, K., Comparison of algorithms that select feature for pattern classifier, Pattern Recognition, 2000, 33(2):25-41.

[7] Raymer, M., Punch, W., Goodman, E. Dimensionality reduction using genetic algorithm. IEEE Transactions on Evolutionary Computing, 2000, 12(4):164-171.

[8] YU Hong-mei, YAO Ping-jing. Combined genetic algorithm/ simulated annealing algorithm for large-scale system energy integration [J] .Computers and Chemical Engineering , Elsevier Science Ltd , 2000 , 8(24) :2023 - 2035.

[9] LI Ji, ZHONG Jiang, WU Zhong-Fu. An Artificial Immune Algorithm for Job Scheduling in Grid Environment with Fuzzy Processing Time[J].Computer Science, 2006, 33(2):35-37