A Feature Selection Algorithm for Big Data Based on Genetic Algorithm

Bo Tian(✉) and Weizhi Xiong

Big Data Institute, Tongren University, Tongren, Guizhou, China
tianbomail@163.com

Abstract. Features selection is an important task since it has significant impact on the data mining performance. This paper present an algorithm to perform feature selection based on the adaptive genetic algorithm. First, the method to compute the crossover probability and mutation probability were proposed. Therefore, the subset feature selection operation can be seen as a process of evolution, and realized adaptive feature subsets selection and optimization. Experimental results demonstrate that the proposed algorithm achieves notable classification accuracy improvements and reduced the total computing time compare to the conventional algorithm.

Keywords: Big data · Feature selection · Genetic algorithm

1 Introduction

With the development of the information and computer technologies, big data mining has been widely employed in various industries, such as finance and stock market, traffic, tourism, health records, social security, science data, and so forth. As everyone knows, the big data containing three problems that is velocity, variety and volume problems, and big data comprises various types of data [1]. These problems are the main challenges for big data mining. Therefore, the traditional data mine method is lead to extreme time consuming and complexity [2]. It cannot meet the demand of big data process and time requirement obviously.

In data mining algorithms, the feature selection model is a key issue. It select the most important feature to improve the performance of classifier, which is employed to predict classes for new samples. There are some studies have been made according to this problem. As an efficient algorithm, Heuristics has been widely applied in feature selection [3], which most is top down supervised learning. Additionally, the heuristics require full set of data in training processing. It is not suitable to dynamic stream processing environment [4], and many feature selection algorithm is only developed for some special application areas. As a general rule, the convergence speed of traditional algorithm is still slow and may not converge to a global minimum [5]. Therefore, it necessary to design a high performance algorithm for feature selection.

F. Qiao et al. (eds.), *Recent Developments in Mechatronics and Intelligent Robotics*,
Advances in Intelligent Systems and Computing 690, DOI 10.1007/978-3-319-65978-7_24

2 Feature Selection Algorithm Based on Genetic Algorithm

Genetic algorithm is an efficient heuristic method which is widely employed to get global solution within solution space. Therefore, to solve the feature selection problem, an adaptive genetic algorithm based was proposed. And the genetic operators such as selection, crossover and mutation were designed in following section.

2.1 Initial Population

Initial population contains n features as $F = (f_1 , f_2, \cdots\cdots, f_n)$. The initial population which referred as $G = Ln$. L is represent the number of bits to the corresponding feature. In order to improve calculative efficiency and accelerate the convergence, we set its value to 5.

The fitness value of individual was calculated according to the following formula.

$$P = \frac{kr_{cf}}{\sqrt{k + k(k+1)r_{ff}}} \tag{1}$$

Where k is constant. The r_{cf} and r_{ff} represent the forward and backward time delay respectively. In calculating process, if $P > \varepsilon$, then the algorithm is terminated. Where the ε is threshold.

2.2 Selection

In order to avoiding premature convergence sometimes occurs, the adaptive selection according to the following probability.

$$Q = \alpha P + \frac{(e - e^{k/k}\max)}{e + e^{k/k}\max} \tag{2}$$

Where Q is the fitness value of next generation. The selection operation are produced according to the probability Q.

2.3 Crossover

The adaptive crossover probability can be computed as follows:

$$P_c = \begin{cases} P_{c2}, f_{avg} > f', P_{c2} \leq 1 \\ P_{c1}(f\max - f')/(f\max - f_{avg}), f_{avg} \leq f', P_{c1} \leq 1; \end{cases} \tag{3}$$

Where f' is the individual which fitness is bigger than another in crossover process. f represented the fitness of individual will be crossover. P_{c1} , P_{c2} , P_{m1} , P_{m2} were parameters and $f_{\max}, f_{avg}$ is the max fitness and average fitness of last generation, respectively.

2.4 Mutation

The purpose of mutation is to introduce a slight perturbation to increase the diversity of trial individuals after crossover, preventing trial individuals from clustering and causing premature convergence of solution. The probability of mutation is calculated as follows:

$$P_m = \begin{cases} P_{m2}, f_{avg} > f,\ P_{m2} \leq 1 \\ P_{m1}(f\max - f)/(f\max - f_{avg}), f_{avg} \leq f,\ P_{m1} \leq 1; \end{cases} \tag{4}$$

The steps involved the proposed algorithm are listed below:

(Step 1) Initialize the population. Crossover probability and the probability of mutation were computed by using (3) and (4) respectively.

(Step 2) The fitness of individual were computed by using (1). And the algorithm is whether return determined by termination condition.

(Step 3) The adaptive selection was completed according to the (2), and the next generation was obtained.

(Step 4) The crossover probability was computed by using (3), A random λ was generation for individual as pair. If $P_c > \lambda$, then start crossover operation.

(Step 5) The mutation probability was computed by using (4), and a random η in the range [0, 1] was generation. When $P_m > \eta$, the mutation is started.

(Step 6) The next generation is obtained. If the termination condition is satisfied, then the optimal feature subset is return, else goto (step 2).

3 Experimental Results

In this section, we evaluate the performance of our proposed feature selection algorithm through computer simulation. The simulation is performed by the matlab. To validate the performance of proposed algorithm, the BIF and C-F which are two kinds of represented algorithms in feature selection are compared [6]. The common dataset is obtained from UCI machine learning repository [7]. The dataset is given as Table 1.

Table 1. The datasets using in experiments.

Dataset	The number of features	Sample size	The number of classes
Anneal	898	38	6
Mfeat-factors	2000	216	10
Mushroom	8124	22	2
Spectrometer	531	100	4
Winc	13	178	3

In order to ensuring the fairness of experiment, Each algorithm would select the same number of features. As the classical leaning algorithm, the decision tree is used in test. And the experimental environment is the Weka. The leaning algorithm run three times and took the average. The result is show as Table 2.

Table 2. Size of feature subset for BIF algorithm

Datasets	Algorithm	The number of selected features	Percentage
Anneal	BIF	14	36.8
	C-F	9	23.6
	Proposed algorithm	7	18.4
Mfeat-factors	BIF	114	52.7
	C-F	173	80.1
	Proposed algorithm	126	58.3
Mushroom	BIF	15	68.1
	C-F	9	40.9
	Proposed algorithm	7	31.8
Spectrometer	BIF	51	51.0
	C-F	45	45.0
	Proposed algorithm	42	42.0
Winc	BIF	8	61.5
	C-F	7	53.8
	Proposed algorithm	4	30.7

The results shown in Tables 1 and 2 revealed that our proposed algorithm could select the smallest feature subset, because it can take out of the sample which has been identified. Furthermore, the classification accuracy of each feature subset was shown as following.

Experiment results are shown in Table 3 for each algorithm. From the simulation results, we can observe that the proposed algorithm can has better classification accuracy over the competing algorithm by using fewer features.

Table 3. Classification accuracy of each feature subset

Dataset	Accuracy		
	BIF	C-F	Proposed algorithm
Anneal	89.24	85.67	89.56
Mfeat-factors	90.43	88.72	92.33
Mushroom	84.39	76.84	86.94
Spectrometer	80.17	74.21	84.26
Winc	74.96	83.31	86.99

The accuracy of classification are shown in Figs. 1 and 2 for each algorithm. From the result, we can conclude that the proposed algorithm have better accuracy when comparing with other algorithm.

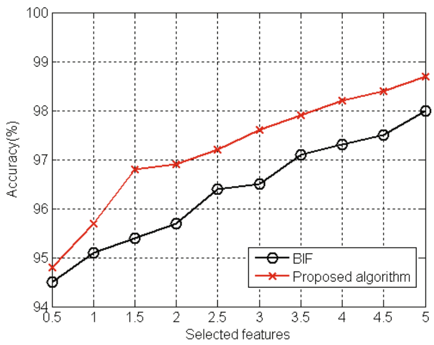

Fig. 1. The accuracy of classification for Anneal

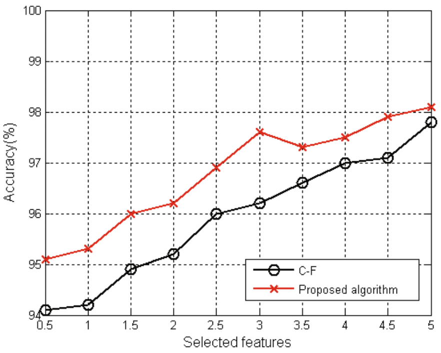

Fig. 2. The accuracy of classification for Winc

4 Conclusions

A feature selection algorithm based on the adaptive genetic was proposed in this paper. The method to compute the crossover probability and mutation probability were designed according to adaptive genetic. The algorithm realized adaptive feature subsets selection and optimization. Experimental results show that the proposed algorithm achieves notable classification accuracy improvements, and reduced the total computing time, comparisons with the conventional scheme.

Acknowledgements. This study was supported by the Science and Technology Cooperation Project of Guizhou province under Grant No. Qiankehe LH [2015] 7251. Science Foundation of Tongren University under Grant No. trxyDH1503.

References

1. Qiu, M., Ming, Z., Li, J.: Phase-change memory optimization for green cloud with genetic algorithm. IEEE Trans. Comput. **64**, 3528–3540 (2015)
2. Chiang, C.L.: Improved genetic algorithm for power economic dispatch of units with valve-point effects and fuels. IEEE Trans. Power Syst. **20**, 1690–1698 (2005)
3. Ronowicz, J., Thommes, M., Kleinebudde, P.: A data mining approach to optimize pellets manufacturing process based on a decision tree algorithm. Eur. J. Pharm. Sci. **73**, 44–51 (2015)
4. Yang, H., Fong, S.: Countering the concept-drift problems in big data by an incrementally optimized stream mining model. J. Syst. Softw. **102**, 158–165 (2015)
5. Ying, X.J., Xin, X.W.: Several feature selection algorithms based on the discernibility of a feature subset and support vector machines. Chin. J. Comput. **37**, 1705–1710 (2014)
6. Pinheiro, R.H.W., Cavalcanti, G.D.C.: A global-ranking local feature selection method for text categorization. Expert Syst. Appl. **39**, 2851–2857 (2012)
7. Information on http://www.ics.uci.edu/~mlearn/MLRepository.html