

Research Article

Application of Global Optimization Methods for Feature Selection and Machine Learning

Shaohua Wu,¹ Yong Hu,¹ Wei Wang,¹ Xinyong Feng,¹ and Wanneng Shu²

¹ College of Electronics and Information Engineering, Sichuan University, Chengdu 610064, China

² College of Computer Science, South-Central University for Nationalities, Wuhan 430074, China

Correspondence should be addressed to Xinyong Feng; xinyong_feng@sohu.com

Received 2 September 2013; Revised 12 October 2013; Accepted 14 October 2013

Academic Editor: Gelan Yang

Copyright © 2013 Shaohua Wu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The feature selection process constitutes a commonly encountered problem of global combinatorial optimization. The process reduces the number of features by removing irrelevant and redundant data. This paper proposed a novel immune clonal genetic algorithm based on immune clonal algorithm designed to solve the feature selection problem. The proposed algorithm has more exploration and exploitation abilities due to the clonal selection theory, and each antibody in the search space specifies a subset of the possible features. Experimental results show that the proposed algorithm simplifies the feature selection process effectively and obtains higher classification accuracy than other feature selection algorithms.

1. Introduction

With the explosive development of massive data, it is difficult to analyze and extract high level knowledge from data. The increasing trend of high-dimensional data collection and problem representation calls for the use of feature selection in many machine learning tasks [1]. Machine learning is the most commonly used technique to address larger and more complex tasks by analyzing the most relevant information already present in databases [2]. Machine learning is programming computers to optimize a performance criterion using example data or past experience. The selection of relevant features and elimination of irrelevant ones are the key problems in machine learning that have become an open issue in the field of machine learning [3]. Feature selection (FS) is frequently used as a preprocessing step to machine learning that chooses a subset of features from the original set of features forming patterns in a training dataset. In recent years, feature selection has been successfully applied in classification problem, such as data mining applications, information retrieval processing, and pattern classification. FS has recently become an area of intense interests and research.

Feature selection is a preprocessing technique for effective data analysis in the emerging field of data mining which is aimed at choosing a subset of original features so that

the feature space is optimally reduced according to the predetermined targets [4]. Feature selection is one of the most important means which can influence the classification accuracy rate and improve the predictive accuracy of algorithms by reducing the dimensionality, removing irrelevant features, and reducing the amount of data needed for the learning process [5, 6]. FS has been an important field of research and development since 1970's and proven to be effective in removing irrelevant features, reducing the cost of feature measurement and dimensionality, increasing classifier efficiency and classification accuracy rate, and enhancing comprehensibility of learned results.

Both theoretical analysis and empirical evidence show that irrelevant and redundant features affecting the speed and accuracy of learning algorithms and thus should be eliminated as well. An efficient and robust feature selection approach including genetic algorithms (GA) and immune clone algorithm (ICA) can eliminate noisy, irrelevant, and redundant data that have been tried out for feature selection.

In order to find a subset of features that are most relevant to the classification task, this paper makes use of FS technique, together with machine learning knowledge, and proposes a novel optimization algorithm for feature selection called immune clonal genetic feature selection algorithm (ICGFSA). We describe the feature selection for selection

of optimal subsets in both empirical and theoretical work in machine learning, and we present a general framework that we use to compare different algorithms. Experimental results show that the proposed algorithm simplifies the feature selection process effectively and either obtains higher classification accuracy or uses fewer features than other feature selection algorithms.

The structure of the rest of the paper is organized as follows. A brief survey is given in Section 2. We study the classification accuracy and formalize it as a mathematical optimization model in Section 3. Section 4 explains the details of the ICGFSA. Several experiments conducted to evaluate the effectiveness of the proposed approach are presented in Section 5. Finally, Section 6 concludes the paper and discusses some future research directions.

2. Related Works

In this section, we focus our discussion on the prior research on feature selection and machine learning. There has been substantial work on feature selection for selection of optimal subsets from the original dataset, which are necessary and sufficient for solving the classification problem.

Extreme learning machine (ELM) is a new learning algorithm for Single Layer Feed-forward Neural network (SLFN) whose learning speed is faster than traditional feed-forward network learning algorithm like back propagation algorithm while obtaining better generalization performance [7]. Support vector machines (SVM) is a very popular machine learning method used in many applications, such as classification. It finds the maximum margin hyperplane between two classes using the training data and applying an optimization technique [8]. SVM has shown good generalization performance on many classification problems.

Genetic algorithm has been proven to be very effective solution in a great variety of approximately optimum search problems. Recently, Huang and Wang proposed a genetic algorithm to simultaneously optimize the parameters and input feature subset of support vector machine (SVM) without loss of accuracy in classification problems [9]. In [10], a hybrid genetic algorithm is adopted to find a subset of features that are most relevant to the classification task. Two stages of optimization are involved. The inner and outer optimizations cooperate with each other and achieve the high global predictive accuracy as well as the high local search efficiency. Reference [11] proposed and investigated the use of a genetic algorithm method for simultaneously aiming at a higher accuracy level for the software effort estimates.

To further settle the feature selection problems, Mr. Liu et al. proposed an improved feature selection (IFS) method by integrating MSPSO, SVM with F -score method [12]. Reference [13] proposed a new evolutionary algorithm called Intelligent Dynamic Swarm (IDS), that is, a modified Particle Swarm Optimization. To evaluate the classification accuracy of IT-IN and remaining four feature selection algorithms, Naive Bayes, SVM, and ELM classifiers are used for ten UCI repository datasets. Deisy et al. proposed IT-IN performs better than the existing above algorithms in terms of number of features [14].

The feature selection process constitutes a commonly encountered problem of global combinatorial optimization. Chuang et al. presented a novel optimization algorithm called catfish binary particle swarm optimization, in which the so-called catfish effect is applied to improve the performance of binary particle swarm optimization [15]. Reference [16] proposed a new information gain and divergence-based feature selection method for statistical machine learning-based text categorization without relying on more complex dependence models. Han et al. study employs feature selection (FS) techniques, such as mutual-information-based filter and genetic algorithm-based wrapper, to help search for the important sensors in data driven chiller FDD applications, so as to improve FDD performance while saving initial sensor cost.

3. Classification Accuracy and F -Score

In this section, the proposed feature selection model will be discussed. In general, feature selection problem can be described as follows.

Definition 1. Assume that $TR = \{D, F, C\}$ represents a training dataset with m features or attributes and n instances, $D = \{o_1, \dots, o_j, \dots, o_n\}$ denotes the instances, $F = \{f_1, \dots, f_i, \dots, f_m\}$ denotes feature space of D constructed from m features, which gives an optimal performance for the classifier, and $C = \{c_1, \dots, c_i, \dots, c_k\}$ represents the set of classes where instances are tagged.

Definition 2. Assume that $o_j = (v_{j1}, \dots, v_{jm})$ represents a value vector of features, where v_{ji} is the value of o_j corresponding to the feature f_i , $o_j \in D$.

The feature selection approaches are used to generate a feature subset F based on the relevance and feature interaction of data samples. The main goal of classification learning is to characterize the relationship between F and C . Assume that F_1 is the subset of already-selected features, F_2 is the subset of unselected features, and $F = F_1 \cup F_2$, $F_1 \cap F_2 = \phi$. Therefore, any optimal feature subset obtained by selection algorithms should preserve the existing relationship between F and C hidden in the dataset.

The best subset of features is selected by evaluating a number of predefined criteria, such as classification accuracy and F -score. In order to evaluate the classification accuracy rate, the specific equation on classification accuracy is defined as follows.

Definition 3. Assume that S is the set of data items to be classified and sc is the class of the item s . If $\text{classify}(s)$ returns the classification accuracy rates of s , then classification accuracy can be formulated as

$$\text{acc}(S) = \frac{\sum_{i=1}^{|S|} \text{ass}(s_i)}{|S|}, \quad s_i \in S, \quad (1)$$

$$\text{ass}(s) = \begin{cases} 1, & \text{classify}(s) = sc, \\ 0, & \text{otherwise,} \end{cases}$$

where $|S|$ represents the number of elements in the collection S , $s \in S$.

F -score is an effective approach which measures the discrimination of two sets of real numbers. The larger the F -score is, the more this feature is discriminative.

Definition 4. Given training vectors X_k . If the number of the j th dataset is n_j , then the F -score of the i th feature is defined as

$$F(s_i) = \frac{\sum_{j=1}^m (\bar{x}_{i,j} - \bar{x}_i)^2}{\sum_{j=1}^m (1/(n_j + 1)) \sum_{k=1}^{n_j} (x_{i,j}^k - x_{i,j})^2}, \quad (2)$$

where \bar{x}_i , $\bar{x}_{i,j}$ are the average of the i th feature of the whole dataset and the j th dataset, respectively; $x_{i,j}^k$ is the i th feature of the k th instance in the j th dataset; m is the number of datasets. $k = 1, 2, \dots, m$ and $j = 1, 2, \dots, l$.

4. Heuristic Feature Selection Algorithm

In this section, we focus our discussion on algorithms that explicitly attempt to select an optimal feature subset. Finding an optimal feature subset is usually difficult, and feature selection for selection of optimal subsets has been shown to be NP-hard. Therefore, a number of heuristic algorithms have been used to perform feature selection of training and testing data, such as genetic algorithms, particle swarm optimization, neural networks, and simulated annealing.

Genetic algorithms have been proven as an intelligent optimization algorithm that can find the optimal solution to a problem in the sense of probability in a random manner [17]. However, standard genetic algorithms have some weaknesses, such as premature convergence and poor local search ability. On the other hand, some other heuristic algorithms, such as particle swarm optimization, simulated annealing, and clonal selection algorithm usually have powerful local search ability.

4.1. Basic Idea. In order to obtain the higher classification accuracy rate and higher efficiency of standard genetic algorithms, some hybrid GA for feature selection have been developed by combining the powerful global search ability of GA with some efficient local search heuristic algorithms. In this paper, a novel immune clonal genetic algorithm based on immune clonal algorithm, called ICGFSA, is designed to solve the feature selection problem. Immune clone algorithm is a simulation of the immune system which has the ability to identify the bacteria and designed diversity, and its search target has certain dispersion and independence. ICA can effectively maintain the diversity between populations of antibodies but also accelerate the global convergence speed [18]. The ICGFSA algorithm has more exploration and exploitation abilities due to the clonal selection theory that an antibody has the possibility to clone some similar antibodies in the solution space with each antibody in the search space specifying a subset of the possible features. The experimental results show the superiority of the ICGFSA in terms of the prediction accuracy with smaller subset of features. The overall scheme of the proposed algorithm framework is outlined in Figure 1.

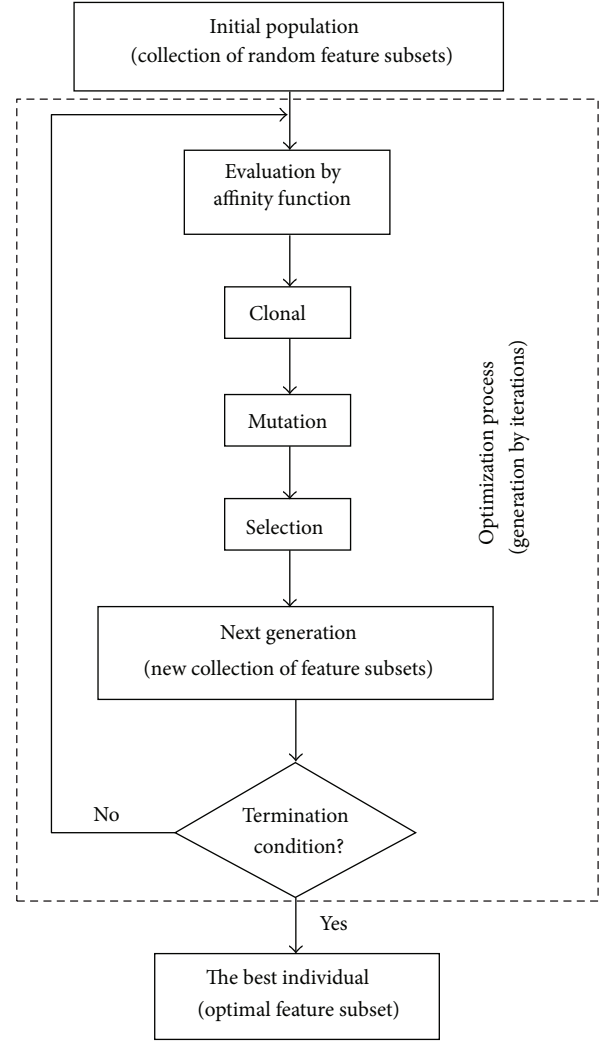


FIGURE 1: Feature selection by ICGFSA algorithm.

4.2. Encoding. In the ICGFSA algorithm, each antibody in the population represents a candidate solution to the feature selection problem. The algorithm uses the binary coding method that “1” means “selected” and “0” means “unselected” [19]. Therefore, the chromosomes represents by a string of binary digits of zeros and ones and each gene in chromosome corresponds to a feature.

4.3. Affinity Function. We design an affinity function that combines classification accuracy rate with F -score, which is the evaluation criterion for the feature selection. The affinity function is defined as follows:

$$\text{affinity}(i) = \lambda_1 \times \text{acc}(s_i) + \lambda_2 \times \frac{1}{|S|} \times \frac{\sum_{j=1}^{|S|} F(\text{FS}(s_j))}{\sum_{j=1}^{|S|} F(s_j)}. \quad (3)$$

In which, $\text{FS}(s_j)$ is equal to the instance of feature i when feature i is selected, otherwise $\text{FS}(s_j)$ is equal to 0, $\lambda_1 + \lambda_2 = 1$.

TABLE 1: Description of dataset.

No.	Datasets	Instances	Features	Classes
1	Liver	345	6	2
2	WDBC	569	30	2
3	Soybean	685	35	19
4	Glass	214	9	6
5	Wine	178	13	3
6	PDF	800	213	2

4.4. Basic Operation. In this section focuses on the three main operations of ICGFSA, including clonal, mutation, and selection. Mutation operation will take the binary mutation operation in standard genetic algorithm [20].

Clonal is essentially the larger antibody affinity for a certain scale replication. Clone size is calculated as follows:

$$\text{size}(i) = \left\lceil \frac{|D|}{|F|} \times \frac{\text{affinity}(i)}{\sum_{j=1}^N \text{affinity}(i)} \right\rceil. \quad (4)$$

In which, $|D|$ and $|F|$ are the number of elements in the set D and F , respectively. N represents the number of antibodies in the population.

The basic idea of selection operation is as follows. Firstly, select the n highest affinity antibodies and generate a number of clones for them. Secondly, antibodies that have been selected directly are retained to the next generation [21].

5. Experimental Results and Discussion

5.1. Parameter Setting. In this section, in order to investigate the effectiveness and superiority of the ICGFSA algorithm for classification problems, the same conditions were used to compare with other feature selection methods such as GA and SVM; that is, the parameters of ICGFSA and GA are set as follows: population size is 50, maximum generations is 500, crossover probability is 0.7, and mutation probability is 0.2. For each dataset we have performed 50 simulations, since the test results depend on the population randomly generated by the ICGFSA algorithm.

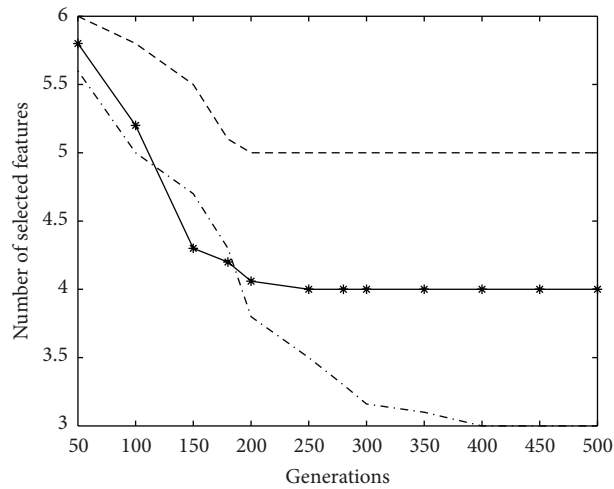
5.2. Benchmark Datasets. To evaluate the performance of ICGFSA algorithms, the following benchmark datasets are selected for simulation experiments: Liver, WDBC, Soybean, Glass, and Wine. These datasets were obtained from the UCI machine learning repository [22] and most of them are frequently used in a comprehensive testing. They suit for feature selection methods under different conditions. Furthermore, to evaluate the algorithms for real Internet data, we also use malicious PDF file datasets from Virus Total [23]. Table 1 is given some general information about these datasets, such as instances, features, and classes.

5.3. Experimental Results. Figure 2 is the number of selected features with different generations in benchmark datasets using ICGFSA, GA, and SVM, respectively. As seen from Figure 2, it can be observed that the number of selected features is decreased with the number of generations increasing,

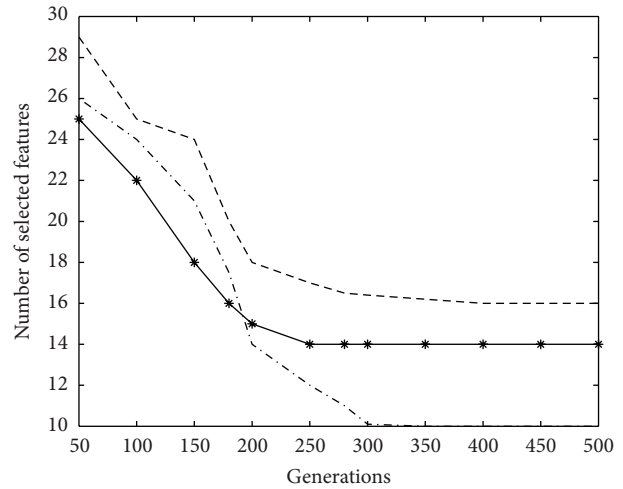
and ICGFSA can converge to the optimal subsets of required number features since it is the stochastic search algorithms. In the Liver dataset, the number of features selected keeps decreasing, while the number of iterations keeps increasing, until ICGFSA obtained nearly 90% classification accuracy, which indicates that a good feature selection algorithm not only decreases the number of features, but also selects features relevant for improving classification accuracy. It can be observed from Figure 3(b) that when the number of iterations increases beyond certain value (say 300), the performance will no longer be improved. In the Wine dataset, there are several critical points (153, 198, 297, etc.) where the trend has been shifted or changed sharply. In the Soybean and Glass datasets, three algorithms have the best performances and significant improvements in the number of selection features.

We carried out extensive experiments to verify the ICGFSA algorithm. The running times that find the best subset of required numbers of features and number of selected features in benchmark datasets using ICGFSA, GA, and SVM are recorded in Table 2. It can be observed from Table 2 that ICGFSA algorithm can achieve significant feature reduction that selects only a small portion from the original features which better than the other two algorithms. ICGFSA is more effective than GA and SVM and, moreover, produces improvements of conventional feature selection algorithms over SVM which is known to give the best classification accuracy. From the experimental results we can obviously see that ICGFSA has the least feature number and clonal selection operations can greatly enforce the local searching ability and make the algorithm fast enough to reach its optimum, which indicates ICGFSA has the ability to break through the local optimal solution when applied to large-scale feature selection problems. It can be concluded that the ICGFSA is relatively simple and can effectively reduce the computational complexity of implementation process.

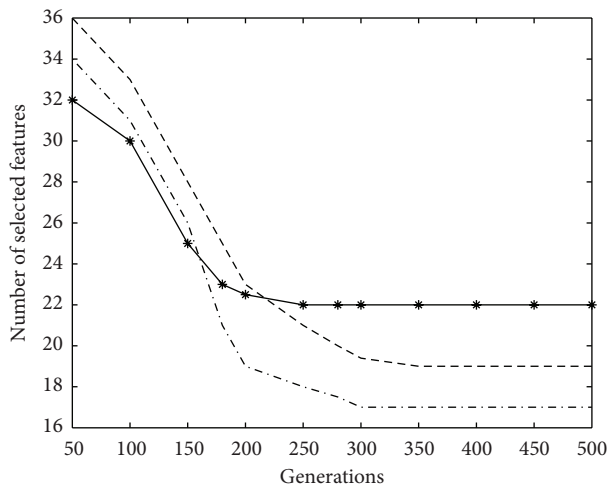
Finally, we inspect the classification accuracy for the proposed algorithm. Figure 3 shows the global best classification accuracies with different generations in benchmark datasets using ICGFSA, GA, and SVM, respectively. In the Liver dataset, the global best classification accuracy of ICGFSA is 88.69%. However, the global best classification accuracy of GA and SVM are only 85.12% and 87.54%, respectively. In the WDBC dataset, the global best classification accuracy of ICGFSA is 84.89%. However, the global best classification accuracy of GA and SVM is only 79.36% and 84.72%, respectively. In the Soybean dataset, the global best classification accuracy of ICGFSA and SVM is 84.96% and 84.94%, respectively. However, the global best classification accuracy of GA is only 77.68%. In the Glass dataset, the global best classification accuracy of ICGFSA is 87.96%. However, the global best classification accuracy of GA and SVM is only 84.17% and 86.35%, respectively. In the Wine dataset, the ICGFSA obtained 94.8% classification accuracy before reaching the maximum number of iterations. In the PDF dataset, the global best classification accuracy of ICGFSA and SVM is 94.16% and 93.97%, respectively. However, the global best classification accuracy of GA is only 92.14%. ICGFSA method is consistently more effective than GA and SVN methods on six datasets.



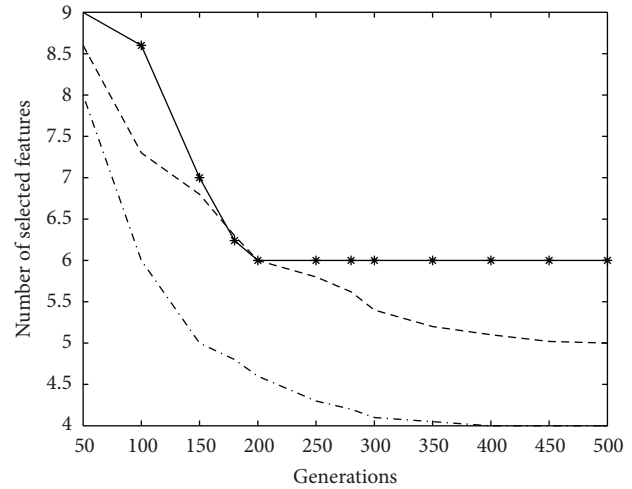
(a) Liver dataset



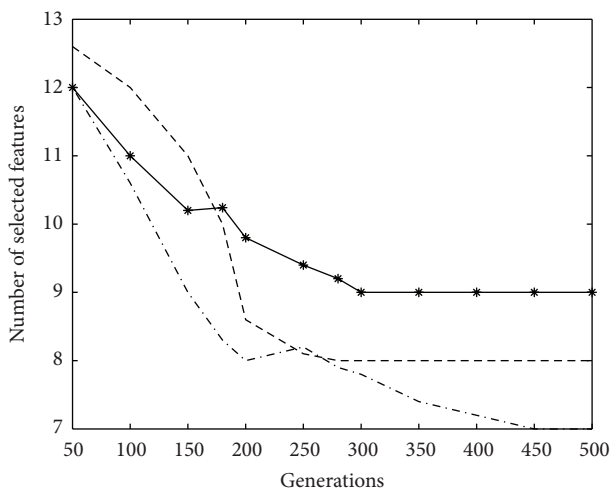
(b) WDBC dataset



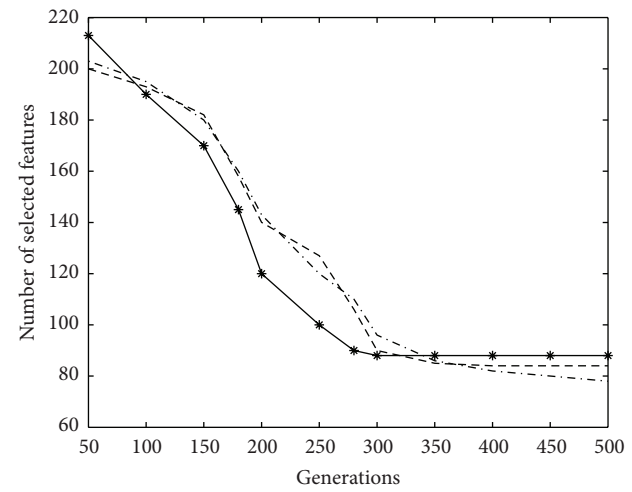
(c) Soybean dataset



(d) Glass dataset



(e) Wine dataset

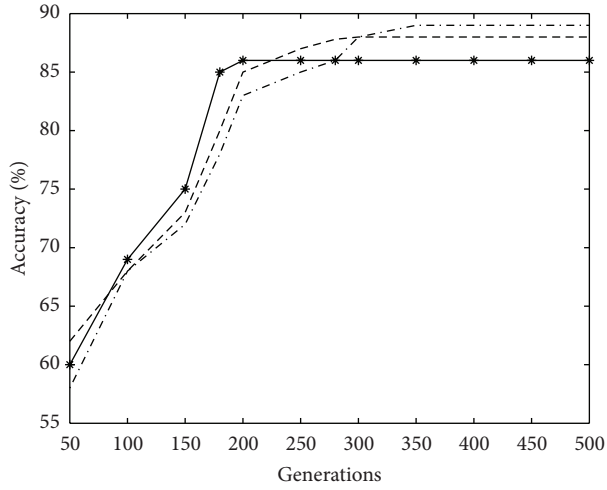


(f) PDF dataset

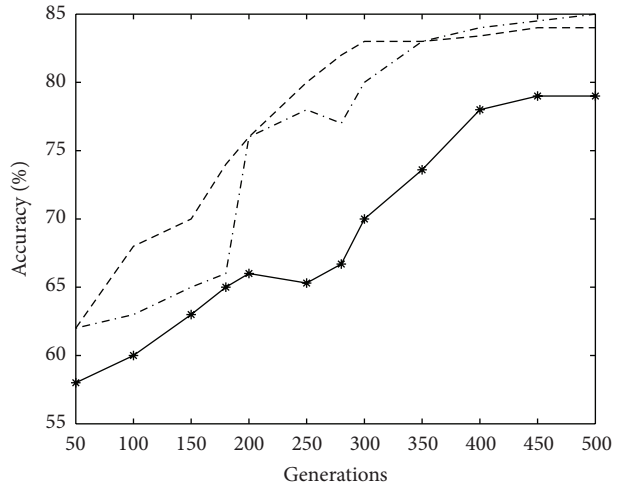
--- ICGFSA
 —●— GA
 --- SVM

--- ICGFSA
 —●— GA
 --- SVM

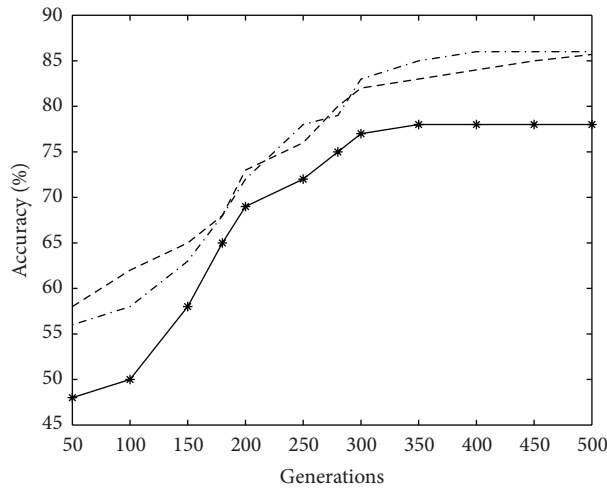
FIGURE 2: Number of selected features with different generations in benchmark datasets.



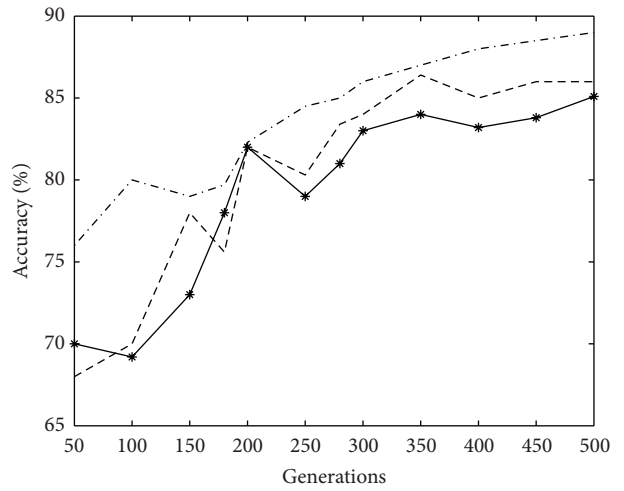
(a) Liver dataset



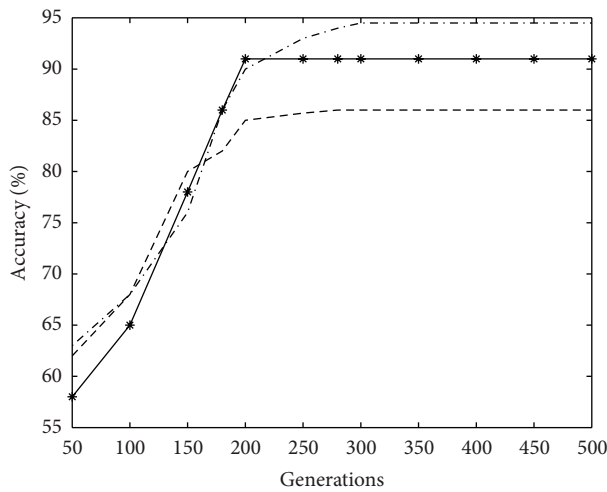
(b) WDBC dataset



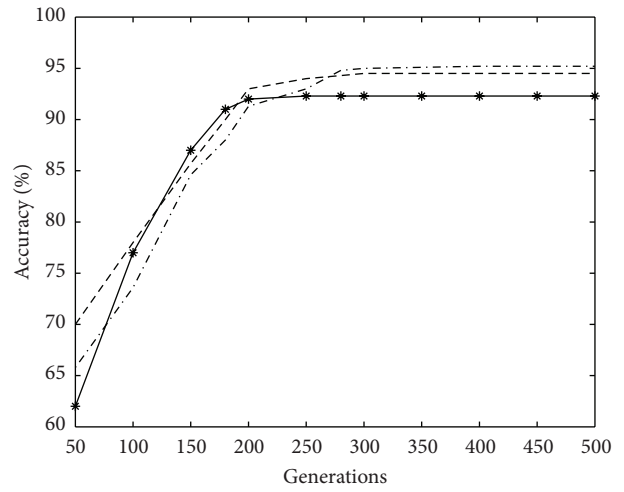
(c) Soybean dataset



(d) Glass dataset



(e) Wine dataset



(f) PDF dataset

--- ICGFSA
 —●— GA
 - - - SVM

--- ICGFSA
 —●— GA
 - - - SVM

FIGURE 3: Global classification accuracies with different generations in benchmark datasets.

TABLE 2: Running time and number of selected features for three feature selection algorithms.

Datasets	Running time (seconds)			Number of selected features		
	ICGFSA	GA	SVM	ICGFSA	GA	SVM
Liver	12.3	11.1	11.2	3	4	5
WDBC	12.6	12.9	13.1	10	14	16
Soybean	13.2	14.7	14.9	17	22	19
Glass	11.8	12.3	11.7	4	6	5
Wine	9.6	10.8	9.5	7	9	8
PDF	830.1	832.5	822	78	89	83

The numerical results and statistical analysis show that the proposed ICGFSA algorithm performs significantly better than the other two algorithms in terms of running time and higher classification accuracy. ICGFSA can reduce the feature vocabulary with best performance in accuracy. It can be concluded that an effective feature selection algorithm is helpful in reducing the computational complexity of analyzing dataset. As long as the chosen features contain enough feature classification information, higher classification accuracy can be achieved.

6. Conclusions

Machine learning is a science of the artificial intelligence. The field's main objectives of study are computer algorithms that improve their performance through experience. In this paper, the main work in machine learning field is on methods for handling datasets containing large amounts of irrelevant attributes. For the high dimensionality of feature space and the large amounts of irrelevant feature, we propose a new feature selection method base on genetic algorithm and immune clonal algorithm. In the future, ICGFSA algorithm will be applied to more datasets for testing performance.

Acknowledgments

This research work was supported by the Hubei Key Laboratory of Intelligent Wireless Communications (Grant no. IWC2012007) and the Special Fund for Basic Scientific Research of Central Colleges, South-Central University for Nationalities (Grant no. CZY11005).

References

- [1] T. Peters, D. W. Bulger, T.-H. Loi, J. Y. H. Yang, and D. Ma, "Two-step cross-entropy feature selection for microarrays-power through complementarity," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 4, pp. 1148–1151, 2011.
- [2] W.-C. Yeh, "A two-stage discrete particle swarm optimization for the problem of multiple multi-level redundancy allocation in series systems," *Expert Systems with Applications*, vol. 36, no. 5, pp. 9192–9200, 2009.
- [3] L.-Y. Chuang, H.-W. Chang, C.-J. Tu, and C.-H. Yang, "Improved binary PSO for feature selection using gene expression data," *Computational Biology and Chemistry*, vol. 32, no. 1, pp. 29–37, 2008.
- [4] B. Hammer and K. Gersmann, "A note on the universal approximation capability of support vector machines," *Neural Processing Letters*, vol. 17, no. 1, pp. 43–53, 2003.
- [5] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of Machine Learning Research*, vol. 5, pp. 1205–1224, 2004.
- [6] G. Qu, S. Hariri, and M. Yousif, "A new dependency and correlation analysis for features," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 9, pp. 1199–1206, 2005.
- [7] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1–3, pp. 489–501, 2006.
- [8] J. G. Dy, C. E. Brodley, A. Kak, L. S. Broderick, and A. M. Aisen, "Unsupervised feature selection applied to content-based retrieval of lung images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 3, pp. 373–378, 2003.
- [9] C.-L. Huang and C.-J. Wang, "A GA-based feature selection and parameters optimization for support vector machines," *Expert Systems with Applications*, vol. 31, no. 2, pp. 231–240, 2006.
- [10] J. Huang, Y. Cai, and X. Xu, "A hybrid genetic algorithm for feature selection wrapper based on mutual information," *Pattern Recognition Letters*, vol. 28, no. 13, pp. 1825–1844, 2007.
- [11] A. L. I. Oliveira, P. L. Braga, R. M. F. Lima, and M. L. Cornélio, "GA-based method for feature selection and parameters optimization for machine learning regression applied to software effort estimation," *Information and Software Technology*, vol. 52, no. 11, pp. 1155–1166, 2010.
- [12] Y. Liu, G. Wang, H. Chen, H. Dong, X. Zhu, and S. Wang, "An improved particle swarm optimization for feature selection," *Journal of Bionic Engineering*, vol. 8, no. 2, pp. 191–200, 2011.
- [13] C. Bae, W.-C. Yeh, Y. Y. Chung, and S.-L. Liu, "Feature selection with Intelligent Dynamic Swarm and rough set," *Expert Systems with Applications*, vol. 37, no. 10, pp. 7026–7032, 2010.
- [14] C. Deisy, S. Baskar, N. Ramraj, J. S. Koori, and P. Jeevanandam, "A novel information theoretic-interact algorithm (IT-IN) for feature selection using three machine learning algorithms," *Expert Systems with Applications*, vol. 37, no. 12, pp. 7589–7597, 2010.
- [15] L.-Y. Chuang, S.-W. Tsai, and C.-H. Yang, "Improved binary particle swarm optimization using catfish effect for feature selection," *Expert Systems with Applications*, vol. 38, no. 10, pp. 12699–12707, 2011.
- [16] C. Lee and G. G. Lee, "Information gain and divergence-based feature selection for machine learning-based text categorization," *Information Processing and Management*, vol. 42, no. 1, pp. 155–165, 2006.

- [17] J. Huang, Y. Cai, and X. Xu, "A hybrid genetic algorithm for feature selection wrapper based on mutual information," *Pattern Recognition Letters*, vol. 28, no. 13, pp. 1825–1844, 2007.
- [18] L. N. De Castro and F. J. Von Zuben, "Learning and optimization using the clonal selection principle," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 3, pp. 239–251, 2002.
- [19] H. Han, B. Gu, T. Wang, and Z. R. Li, "Important sensors for chiller fault detection and diagnosis (FDD) from the perspective of feature selection and machine learning," *International Journal of Refrigeration*, vol. 34, no. 2, pp. 586–599, 2011.
- [20] P. Kumsawat, K. Attakitmongcol, and A. Srikaew, "A new approach for optimization in image watermarking by using genetic algorithms," *IEEE Transactions on Signal Processing*, vol. 53, no. 12, pp. 4707–4719, 2005.
- [21] R. Meiri and J. Zahavi, "Using simulated annealing to optimize the feature selection problem in marketing applications," *European Journal of Operational Research*, vol. 171, no. 3, pp. 842–858, 2006.
- [22] C. L. Blake and C. J. Merz, "UCI repository of machine learning databases," Department of Information and Computer Science, University of California, Irvine, Calif, USA, 1998, <http://www.ics.uci.edu/mllearn/MLRepository.html>.
- [23] VirusTotal: <http://www.virustotal.com>.



Hindawi

Submit your manuscripts at
<http://www.hindawi.com>

