Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2013, Article ID 524017, 6 pages
http://dx.doi.org/10.1155/2013/524017

Research Article

A Heuristic Feature Selection Approach for Text Categorization by Using Chaos Optimization and Genetic Algorithm

Hao Chen,[1] Wen Jiang,[2] Canbing Li,[3] and Rui Li[1]

[1] *School of Information Science and Engineering, Hunan University, Changsha 410082, China*
[2] *School of Software, Hunan Vocational College of Science and Technology, Changsha 410118, China*
[3] *College of Electrical and Information Engineering, Hunan University, Changsha 410082, China*

Correspondence should be addressed to Hao Chen; xschenhao@139.com

Received 10 October 2013; Revised 13 November 2013; Accepted 17 November 2013

Academic Editor: Gelan Yang

Due to the era of Big Data and the rapid growth in textual data, text classification becomes one of the key techniques for handling and organizing the text data. Feature selection is the most important step in automatic text categorization. In order to choose a subset of available features by eliminating unnecessary features to the classification task, a novel text categorization algorithm called chaos genetic feature selection optimization is proposed. The proposed algorithm selects the optimal subsets in both empirical and theoretical work in machine learning and presents a general framework for text categorization. Experimental results show that the proposed algorithm simplifies the feature selection process effectively and can obtain higher classification accuracy with a smaller feature set.

1. Introduction

Due to the era of Big Data and the rapid growth in textual data, feature selection (FS) is very important for organizing the data. Feature selection is also called attribute selection. Feature selection is a key step in automatic text categorization and machine learning systems, which automatically assigns the documents to a set of predefined classes based on their textual content. It is well known that feature selection is often used to deal with a high dimensional space of features whose main objective is to simplify a dataset by reducing its dimensionalities and identifying relevant underlying features. In the practical application of machine learning, the number of features which exist irrelevant and interdependent features usually very large. It easily leads to the following consequences. Firstly, there is more time consumption in features analysis and model training when the number of features is increasing. Secondly, it easily leads to "curse of dimensionality" when the number of features is increasing and results in the model becoming more complicated. Feature selection has been widely applied to various fields including text categorization [1], signal processing [2], data mining [3], machine learning [4], neural networks, and pattern recognition [5].

Given a feature set $X = \{X_1, \ldots, X_n\}$ with size n, there exist 2^n possible feature subsets and each feature subset is represented by a binary vector of dimension 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 [6]. As redundant features will affect the system classification accuracy and increase the computation time, we should eliminate the features with little information and ignore the redundant features that are strongly correlated. Feature selection can effectively deal with the problem because of its flexibility, computational efficiency, and capacity to handle high dimensional data [7].

In order to choose a subset of available features by eliminating unnecessary features to the categorization task, this paper makes use of FS method, together with machine learning knowledge, and proposes a novel heuristic algorithm for feature selection called chaos genetic feature selection optimization (CGFSO). Chaos is universal phenomenon in

many nonlinear systems that exhibits sensitive dependence on initial conditions and includes infinite unstable periodic motions [8]. Chaotic optimization algorithm (COA) first changes optimized variables into chaotic variables, examines each point in the entire solution space by change rule of chaotic variables, and accepts the better point as the present optimum solution [9]. Then, it takes the present optimum solution as the kernel and goes on searching the optimum solution by affixing a perturbation until the requirements are met. CGFSO 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 proposed algorithm selects the optimal subsets in both empirical and theoretical work in machine learning and presents a general framework for text categorization. Compared with other existing algorithms, the proposed algorithm simplifies the feature selection process effectively and can obtain higher classification accuracy with a smaller feature set.

The rest of this paper is organized as follows. Section 2 discusses on the prior research on feature selection. Section 3 proposes the CGFSO algorithm. Section 4 shows the experimental results and finally some conclusions are pointed out and future works are offered in Section 5.

2. Related Works

In this section we focus our discussion on the prior research on feature selection. Many scholars at home and abroad have made great contributions to the feature selection in both empirical and theoretical work, which are necessary and sufficient for solving the text categorization problem.

In order to achieve minimum classification error, Kanan and Faez [10] presented an improved ant colony optimization algorithm for feature selection in face recognition. Their algorithm can select the optimal feature subset in terms of shortest feature length and the best performance of classifier. Cao et al. [11] further developed this method by learning feature weights in kernel spaces. The proposed algorithm was often done as a data processing step, independent of classifier construction. To address the problem of jointly learning SVM (support vector machine) parameters and kernels, Zhen et al. [12] proposed a method for choosing SVM parameters including the parameters of kernels by minimizing the leave-one-out cross validation error.

Genetic algorithm (GA) is a parallel heuristic intelligent method, which is a popular technology for nonlinear optimization problem. Due to the advantages of GA, GA has been widely used an effective tool for FS in text categorization. Zhu et al. [13] proposed a combined feature subset selection method, called RICGA (ReliefF immune clonal genetic algorithm) based on the ReliefF algorithm, immune clonal algorithm, and GA. In the RICGA method, the paper first use ReliefF to get rid of irrelevant features then apply a modified genetic algorithm to acquire the finally feature subset. In order to extract feature set, Kim et al. [14] applied genetic algorithm to the feature selection problem and proposed a novel genetic algorithm feature selection (GAFS). Muni et al. [15] presented an online feature selection algorithm using genetic programming (GP). The proposed GP method simultaneously selected a good subset of features and constructed a classifier using the selected features. Waqas et al. [16] focused on multiobjective genetic algorithms for solving feature subset selection. The research showed that independent subsets of features are excellent in accuracy. AlSukker et al. [17] presented a novel modified genetic algorithm based on enhanced population diversity, parents' selection, and improved genetic operators. Practical results indicated the significance of the proposed GA variant in comparison to many other algorithms from the literature on different datasets. Mahrooghy et al. [18] employed filter-based feature selection genetic algorithm (FFSGA) to find an optimal set of features where redundant and irrelevant features are removed. The entropy index fitness function was used to evaluate the feature subsets. The results showed that using the feature selection technique not only improves the equitable threat score by almost 7% at some threshold values for the winter season, but also extremely decreases the dimensionality.

3. Application of CGFSO Algorithm

In this section, we focus our discussion on algorithms that explicitly attempt to select an optimal feature subset. It is usually difficult to obtain an optimal feature subset and has been proven to be NP-hard. Therefore, lots of heuristic algorithms have been used to perform feature selection of training including genetic algorithms, neural networks, and simulated annealing. In order to avoid the combinatorial search problem to find an optimal subset of m features, the most popular feature selection methods is the application of genetic algorithm, which always provide a suboptimal solution.

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 [19]. Chaos movement can nonrepeatedly cover all state in a certain range, according to its own rules [20]. COA shows a promising performance on nonlinear function optimization. However, the local search capability of COA is poor since its heuristic and stochastic properties often suffer from getting stuck in local optima. Thus, this paper takes advantage of the merit of GA and COA and a novel FS algorithm for text categorization; namely, CGFSO is proposed. The experimental results show that the proposed CGFSO finds subsets that result in the best accuracy, while finding compact feature subsets and performing faster than other traditional methods.

3.1. Chaotic Optimization Algorithm. COA is a novel approach of global optimization that has attracted widespread attention in recent years. In the COA, the well-known logistic map is normally described as follows:

$$x_{n+1} = \mu x_n \left(1 - x_n\right), \tag{1}$$

where μ is a control parameter, which cannot be bigger than 4, and x is a variable. It is easy to find that (1) is a deterministic dynamic system without any stochastic disturbance. When

$\mu = 4$ $(0 \le x_0 \le 1)$, the system above is completely in chaos state.

The basic process of chaos optimization algorithm generally includes two major steps. Firstly, define a chaotic sequences generator based on the logistic map. Generate a sequence of chaotic points and map it to a sequence of design points in the original design space. COA has a very sensitive dependence upon its initial condition and parameter. Chaotic sequences have been adopted instead of random sequences and somewhat good results have been shown in many applications. Then, calculate the objective function based on the generated design points, and choose the point with the minimum objective function as the current optimum. Secondly, the current optimum is assumed to be close to the global optimum after certain iterations, and it is viewed as the consult point with a little chaotic perturbation and explores the descent direction along axis directions in order. Repeat the above two steps until some specified convergence criterion is satisfied, then the global optimum is obtained. However, further numerical simulation showed that the method is effective only in small design space.

3.2. Chaos Genetic Feature Selection Optimization. Generally, a text categorization system consists of several essential parts including feature extraction and feature selection [21, 22]. In the feature selection stage can be used with the proposed algorithms to obtain a feature subset that allows the increase of the classification system accuracy and simplicity, and the reduction of the learning efforts. CGFSO 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. Firstly, generating a feature subset from the given feature set, then using the evaluation function to evaluate the feature subset. Evaluation results are compared with the stopping criterion, if the result of the evaluation is better than stopping criterion, then CGFSO algorithm automatically stops. Otherwise, CGFSO algorithm continues to produce the next feature subset. Feature subsets elected general also verify its validity.

In CGFSO algorithm, each individual in the population represents a candidate solution to the feature selection problem [23]. The first thing to consider is the algorithm coding problem, and we set the number of features as the length of chromosomes. If the individual (chromosome) a is represented as a string $a_1 a_2 \dots a_i \dots a_n$, each gene a_i corresponds to the ith feature. If $a_i = 1$, it means that the corresponding feature is selected. If $a_i = 0$, it indicates that the ith feature is ignored.

The solution quality in terms of classification accuracy is evaluated by classifying the training data sets using the selected features. Classification accuracy and feature cost are the two key factors used to design a fitness function. The test accuracy measures the number of examples that are correctly classified. Thus, the individual who has high classification accuracy and low total feature cost produces a high fitness value. The individual with high fitness value has high probability to be selected to the next generation. A solution obtaining higher accuracy and with fewer features will get a greater quality function value. Therefore, the fitness function can be defined as follows:

$$f(x) = \sqrt{\text{Precision}(x)^2 + \text{Recall}(x)^2} - \lambda \times \frac{\delta(x) \times \cos t(x)}{\text{Precision}(x) + \text{Recall}(x) + 1} + \cos t_{\max}, \quad (2)$$

where $\text{Precision}(x)$ is the test precision ratio, $\text{Recall}(x)$ is the test recall ratio, $\cos t(x)$ is the sum of measurement costs of the feature subset represented by x, and λ $(0 \le \lambda \le 1)$ is the adjustment coefficient. $\cos t_{\max}$ is an upper bound on the costs of candidate solutions. In this case, $\cos t_{\max}$ is simply the sum of the costs associated with all of the features. $\delta(x) = 1$ indicates that feature x is selected; otherwise, $\delta(x) = 0$ indicates that feature x is ignored.

The main steps of the CGFSO algorithm can be summarized as follows.

Step 1. Give the population size *PopSize*, the crossover probabilities p_c, the mutation probabilities p_m, and the termination generation G_m. Then randomly initialize the initial population $P(k)$, and set evolution generation $k = 0$.

Step 2. Evaluate the fitness of initial population $P(k)$ according to objective function.

Step 3. Select *PopSize*/5 individuals with larger fitness to the next generation population $P(k + 1)$.

Step 4. Perform the crossover operation for $P(k)$ to generate $Q(k)$.

Step 5. Perform logistic chaotic mutation for the population $Q(k)$ to generate the population $L(k)$.

Step 6. Compute individual fitness after logistic chaotic mutation. If the fitness value after mutation is larger than the old one, then substitute the old one with it, and obtain the next generation population $P(k + 1) = Q(k) \cup L(k)$.

Step 7. $k = k + 1$; if stopping conditions are satisfied, the algorithm ends, and then output the best feature subset; otherwise, go back to Step 2 until the maximum evolution iterations are completed.

4. Experimental Results

In this section, a series of simulation experiments were conducted to show the effectiveness and superiority of the CGFSO algorithm for text categorization problems. In order to provide an overview on the base accuracy of the classifiers, the Reuters collection was taken in our experiments. We uses Reuters-21567 that are 5213 documents in training set and 2016 documents in test set and adopt the top ten classes. Experimental platform use Dell computer with CPU Xeon 3.06 GHz (24P8122) and 2 GB of RAM. We implement the proposed CGFSO algorithm and other two FS methods such as GA and SVM; that is, the parameters of CGFSO and GA are set as follows: the size of the population is 100, the maximum

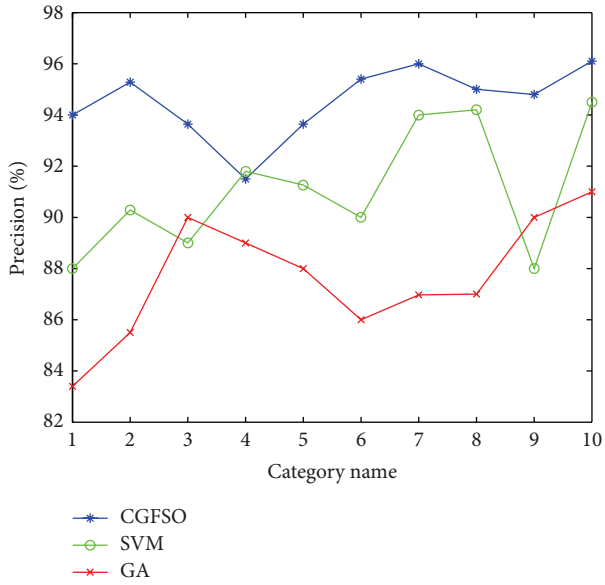

Figure 1: The precision of the three feature selection algorithms.

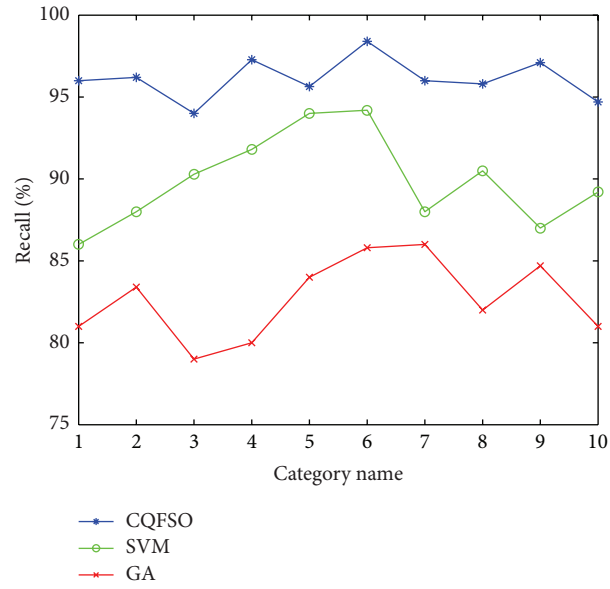

Figure 2: The recall of the three feature selection algorithms.

number of generations is 500, crossover probability is 0.7, and mutation probability is 0.2. Since the experimental result depend on the population randomly generated by the CGFSO and GA algorithms, so we have performed 20 simulations on each data set.

4.1. Precision and Recall. In most text categorization, the performance of feature selection techniques is particularly important. Several norms such as precision and recall are often 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. Precision and recall are defined as follows.

Definition 1. Assume that TP_i represents the number of test documents correctly classified under ith category (C_i) and FP_i denotes the number of test documents incorrectly classified C_i; then classification precision can be formulated as

$$\text{Precision}\,(i) = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i}. \tag{3}$$

Definition 2. Assume that TP_i represents the number of test documents correctly classified under ith category (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; then classification recall can be formulated as

$$\text{Recall}\,(i) = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i}. \tag{4}$$

4.2. Simulation Experiment. To analyse the performance of the feature selection algorithms, we will show the results obtained using the proposed approach. Figures 1–5 show the

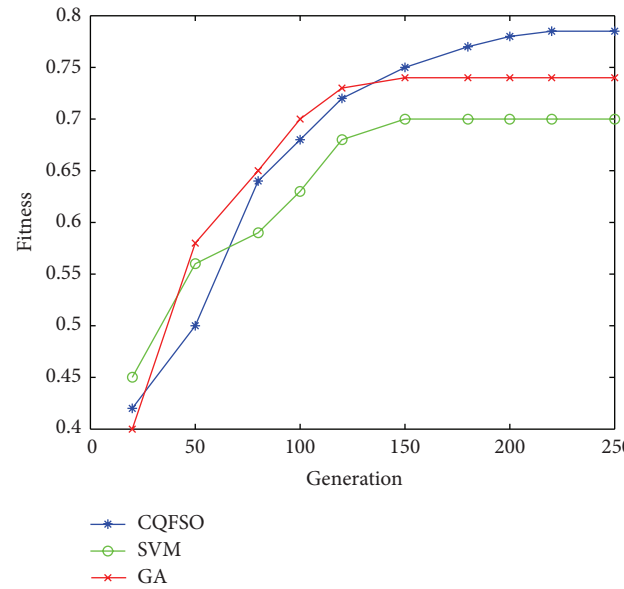

Figure 3: The fitness value of the three feature selection algorithms.

performance of our proposed method against the GA and SVM for the ten most frequent categories with respect to classification accuracy. Figure 1 is the precision of GA, SVM, and CGFSO with different categories. Figure 2 shows the recall of the GA, SVM, and CGFSO. Figure 3 shows the average fitness in the solutions obtained by the algorithms GA, SVM, and CGFSO. Figure 4 is the precision of GA, SVM, and CGFSO with different number of features. Figure 5 is the recall of GA, SVM, and CGFSO with different number of features.

From the experimental result in Figure 1, it can be seen that the precision of CGFSO is the highest in most cases and its maximum value is close to 96%. However, the precision

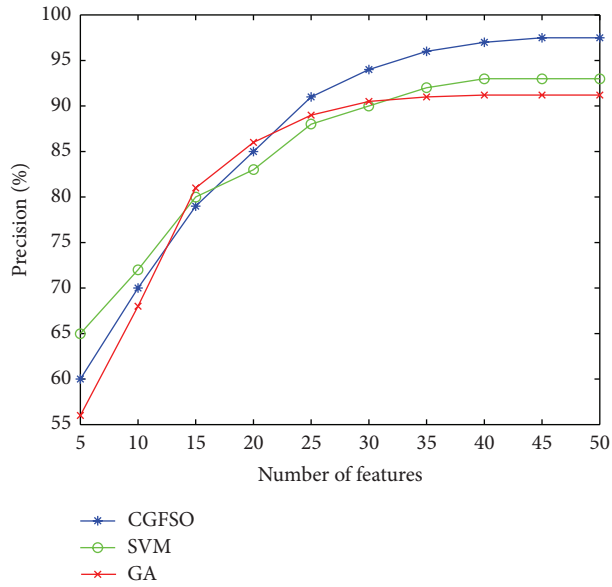

Figure 4: The precision of algorithms with different number of features.

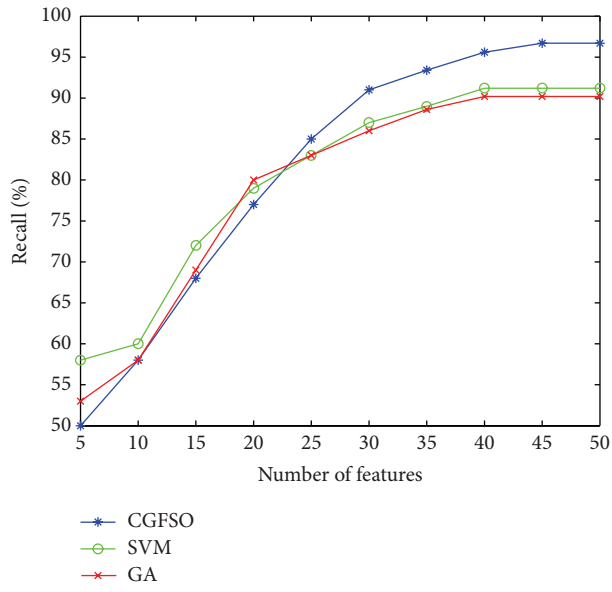

Figure 5: The recall of algorithms with different number of features.

of GA is relatively minimal and its minimum value is close to 83.1%. From the experimental results in Figure 2, it can be seen that CGFSO is significantly better than the other two algorithms in the aspect of recall. The maximum recall of CGFSO is close to 98%. However, the recall of GA is relatively minimal and its minimum value is close to 78.5%. From the experimental results in Figures 1 and 2, we can easily see that CGFSO algorithm can obtain better performance with a smaller feature set than other two algorithms, especially in the aspect of recall.

From the experimental results in Figure 3, the average fitness of CGFSO is the highest in most cases and its maximum value is close to 0.78. The performance of SVM and GA is relatively close. Because CGFSO effectively combines the advantages of chaos optimization algorithm and genetic algorithm, and effectively expands the range of feasible solution. When a gradual increase is in the number of features, the precision and recall of the three feature selection algorithms are gradually increased. As can be seen from Figures 4 and 5, the overall performance of CGFSO is significantly superior to GA and SVM. It is worth noting that our approach has the least number of support vectors compared with other feature selection approaches.

It can be seen from the experimental results that CGFSO learning process effectively and efficiently reduces the complexity of the system in the feature selection stage.

5. Conclusions

Due to the era of Big Data and the rapid growth in textual data, text classification has become a way to process and organize the text data. In order to achieve the goal of this paper, we designed a new text classification algorithm based on genetic algorithm and chaotic optimization algorithm. The experimental results show that the CQFSO yields the best result of these three methods. The experiment also demonstrated that the CQFSO yields better accuracy even with a large data set since it achieved better performance with the lower number of features. In the future, we will design a new heuristic feature selection algorithm, apply it to text classification field, and will involve experiments with other kinds of datasets.

Acknowledgment

This work is partially supported by the National Science Foundation of China under Grant nos. 61370226 and 61272546.

References

[1] M. H. Aghdam, N. Ghasem-Aghaee, and M. E. Basiri, "Application of ant colony optimization for feature selection in text categorization," in *Proceedings of the IEEE Congress on Evolutionary Computation (CEC '08)*, pp. 2867–2873, IEEE Press, Hong Kong, June 2008.

[2] T. W. Liao, "Feature extraction and selection from acoustic emission signals with an application in grinding wheel condition monitoring," *Engineering Applications of Artificial Intelligence*, vol. 23, no. 1, pp. 74–84, 2010.

[3] P. E. N. Lutu and A. P. Engelbrecht, "A decision rule-based method for feature selection in predictive data mining," *Expert Systems with Applications*, vol. 37, no. 1, pp. 602–609, 2010.

[4] S. Nemati, M. E. Basiri, N. Ghasem-Aghaee, and M. H. Aghdam, "A novel ACO-GA hybrid algorithm for feature selection in protein function prediction," *Expert Systems with Applications*, vol. 36, no. 10, pp. 12086–12094, 2009.

[5] S. M. Awaidah and S. A. Mahmoud, "A multiple feature/resolution scheme to Arabic (Indian) numerals recognition using hidden Markov models," *Signal Processing*, vol. 89, no. 6, pp. 1176–1184, 2009.

[6] S.-W. Lin, K.-C. Ying, S.-C. Chen, and Z.-J. Lee, "Particle swarm optimization for parameter determination and feature selection of support vector machines," *Expert Systems with Applications*, vol. 35, no. 4, pp. 1817–1824, 2008.

[7] 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.

[8] H.-J. Lu, H.-M. Zhang, and L.-H. Ma, "New optimization algorithm based on chaos," *Journal of Zhejiang University SCIENCE A*, vol. 7, no. 4, pp. 539–542, 2006.

[9] C.-G. Fei and Z.-Z. Han, "A novel chaotic optimization algorithm and its applications," *Journal of Harbin Institute of Technology*, vol. 17, no. 2, pp. 254–258, 2010.

[10] H. R. Kanan and K. Faez, "An improved feature selection method based on ant colony optimization (ACO) evaluated on face recognition system," *Applied Mathematics and Computation*, vol. 205, no. 2, pp. 716–725, 2008.

[11] B. Cao, D. Shen, J.-T. Sun, Q. Yang, and Z. Chen, "Feature selection in a kernel space," in *Proceedings of the 24th International Conference on Machine Learning (ICML '07)*, pp. 121–128, June 2007.

[12] Z. Zhen, X. Zeng, H. Wang, and L. Han, "A global evaluation criterion for feature selection in text categorization using Kullback-Leibler divergence," in *Proceedings of the International Conference of Soft Computing and Pattern Recognition (SoCPaR '11)*, pp. 440–445, October 2011.

[13] Y. Zhu, X. Shan, and J. Guo, "Modified genetic algorithm based feature subset selection in intrusion detection system," in *Proceedings of the IEEE International Symposium on Communications and Information Technology (ISCIT '05)*, pp. 10–13, IEEE Computer Society, Wuhan China, October 2005.

[14] H.-D. Kim, C.-H. Park, H.-C. Yang, and K.-B. Sim, "Genetic algorithm based feature selection method development for pattern recognition," in *SICE-ICASE International Joint Conference*, pp. 1020–1025, October 2006.

[15] D. P. Muni, N. R. Pal, and J. Das, "Genetic programming for simultaneous feature selection and classifier design," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 36, no. 1, pp. 106–117, 2006.

[16] K. Waqas, R. Baig, and S. Ali, "Feature subset selection using multi-objective genetic algorithms," in *Proceedings of the 13th IEEE International Multitopic Conference (INMIC '09)*, pp. 1–6, December 2009.

[17] A. AlSukker, R. N. Khushaba, and A. Al-Ani, "Enhancing the diversity of genetic algorithm for improved feature selection," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC '10)*, pp. 1325–1331, October 2010.

[18] M. Mahrooghy, H. Y. Nicolas, G. A. Valentine, A. James, and Y. Shantia, "On the use of the genetic algorithm filter-based feature selection technique for satellite precipitation estimation," *IEEE Geoscience and Remote Sensing Letters*, vol. 9, no. 5, pp. 963–967, 2012.

[19] M. E. ElAlami, "A filter model for feature subset selection based on genetic algorithm," *Knowledge-Based Systems*, vol. 22, no. 5, pp. 356–362, 2009.

[20] S. Han, W. Pedrycz, and C. Han, "Nonlinear channel blind equalization using hybrid genetic algorithm with simulated annealing," *Mathematical and Computer Modelling*, vol. 41, no. 6-7, pp. 697–709, 2005.

[21] H. Kim, P. Rowland, and H. Park, "Dimension reduction in text classification with support vector machines," *Journal of Machine Learning Research*, vol. 6, pp. 37–53, 2005.

[22] E. S. Correa, M. T. A. Steiner, A. A. Freitas, and C. Carnieri, "A genetic algorithm for solving a capacitated p-median problem," *Numerical Algorithms*, vol. 35, no. 2–4, pp. 373–388, 2004.

[23] I.-S. Oh, J.-S. Lee, and B.-R. Moon, "Hybrid genetic algorithms for feature selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 11, pp. 1424–1437, 2004.