A Feature Selection Method for Anomaly Detection Based on Improved Genetic Algorithm

Shi Chen, Zhiping Huang, Zhen Zuo & Xiaojun Guo
College of Mechatronics Engineering and Automation, National University of Defense Technology, Changsha, 410073, China

ABSTRACT: Since anomaly detection systems often need to handle large amounts of data, feature selection, which is an effective method for reducing data complexity, is usually applied for anomaly detection. In this paper, an improved genetic algorithm based feature selection method is proposed to obtain optimal features subset with not only considering the performance of classifier but the features generation costs. An optimal weighted nearest neighbor classifier is also adopted to improve the detection performance with the selected features. The experiment results on NSL-KDD dataset show that the proposed method achieves a better or similar performance with 99.66% detection rate and 0.70% false negative rate, when compared with that based on all features.
KEYWORD: Anomaly detection; feature selection; genetic algorithm

1 INTRODUCTION

Due to the proper use of features subset not only can reduce the complexity of model but also improve the classification performance, feature selection become an important part of anomaly detection systems (H.J. Liao et al, 2013) (P.G.-T. P. Garcıa-Teodoro et al, 2009). Moreover, applying a lower dimensional space to represent the data can avoids over-fitting and the *curse of dimensionality*. As a result, it is infeasible to construct an anomaly detection system based on all features, and feature selection is more and more indispensable.

Various feature selection or dimensionality reduction methods for anomaly detection have been proposed throughout the past decades. Principal Component Analysis (PCA) is one of the most common techniques for solving the problem of high dimensionality of features space (S. Lakhina et al, 2010). In order to detect network attacks, Tang, Jiang, and Zhao applied information gain method to select more discriminative features by using k-means clustering algorithm combined with SVM classifier (P. Tang et al, 2010). The cuttlefish algorithm (CFA) was firstly applied as a feature selection method for IDSs by Eesa, Orman and Brifcani (A.S. Eesa et al, 2015). After the optimal features subset was obtained, a decision tree (DT) classifier was used as a classifier to train and test the dataset with the selected features. Stein, Chen, Wu, and Hua proposed a hybrid genetic-DT model that the genetic algorithm (GA) was adopted as a search strategy to obtain an optimal subset of features, and a DT based on C4.5 algorithm was constructed as the evaluator on the selected features (G. Stein et al, 2005). De la Hoz, Ortiz and Ortega built a wrapper model based on NSGA-II algorithm to select proper feature subsets in DARPA/NSL-KDD datasets (E. de la Hoz et al, 2014). All above feature selection methods did not take into account the effort to generate a feature. Félix Iglesias, Tanja and Zseby analyzed features generation costs based on the standard IP Flow Information Export records in detail (F. Iglesias et al, 2014). They drew a conclusion that the costs for feature generation were relevant in a practical system. Considering the effort of features generation, the costs of features in the NSL-KDD dataset are not equal. The feature generation costs could be categorized into S-small, M-medium and H-high (F. Iglesias et al, 2014). If we attempt to obtain less costly features subset and still achieve desired classification performance, we can construct a higher efficient and effective model than that with costly features.

In this paper, we propose an improved GA based feature selection method for anomaly detection. The work of this paper is aiming to select features with not only considering the performance of classifier but the features generation costs. The NSL-KDD (2009) dataset (E. Bagheri et al, 2009), which become to be a benchmark dataset for anomaly detection, is used to evaluate the proposed model. In order to further improve the detection performances,

an optimal weighted nearest neighbor (OWNN) classifier (R.J. Samworth, 2012) is utilized.

2 METHODOLOGIES

Feature selection can be treated as an optimize process, GA is known as a perfect algorithm for solving optimization problems and has been applied in numbers of works to select the significative subset of features (G. Stein et al, 2005) (E. de la Hoz et al, 2014) (T. Shon et al, 2006). Since the traditional GA is easy to converge to the local optimal solution, a crowding based niche technology is employed to improve the global optimization ability of GA. In this section, we introduce the processes that using the improved GA for selecting features with the OWNN classifier.

A feature selection wrapper model based on the improved GA is constructed. In this model, the optimal features subset search part is the improved GA and the evaluation part is OWNN classifier. Figure 1 shows the description of this model in detail.

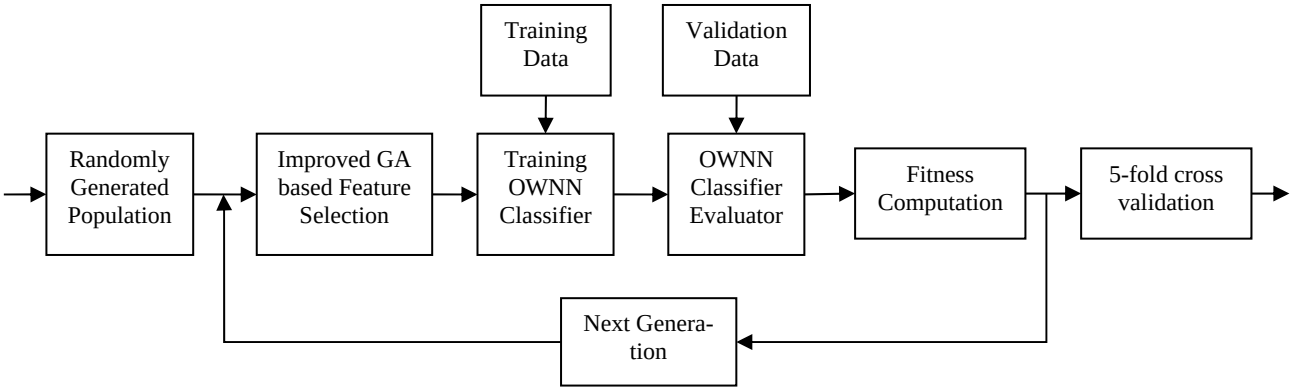

Figure 1 Improved GA/OWNN Classifier Hybrid

In order to build the model of feature selection based on the improved GA and OWNN classifier, each individual's chromosome is coded in binary value to represent a subset of features and the number of neighbors (*k*) considered in OWNN classifier. For the first step, we transform features and the *k* value into binary gene strings that each feature and the *k* value are converted into '0' or '1'. Specifically, each individual has 60 binary bits that the first 41 bits represent the features to select or not and the last 19 bits represent the *k* value. For the first 41 bits, '1' means the corresponding feature is selected and '0' means not. On the other hand, *k* value is represented by the number of "1" contained in the last 19 bits, and plus one. The reason for adding one is that k=0 is not a valid input for OWNN. As a result, the subset of features and *k* value are used as the input for the OWNN classifier. The fitness of each individual depends on the classification performance with the following objective function:

$$f(X) = DR - FNR + \frac{1}{1+C^2(X)} \tag{1}$$

Where, DR and FNR are the OWNN classifier based anomaly detection rate and false negative rate, respectively. $C(X)$ is selected features generation costs function (refer to (F. Iglesias et al, 2014) for detailed information), the less costly the better. We can calculate DR and FNR by using the formulas as follows:

$$DR = \frac{TP}{TP + FP} \tag{2}$$

$$FNR = \frac{FN}{TP + FN} \tag{3}$$

Where, TP, true positive is the number of anomaly examples correctly predicted as anomaly. FP, false positive is the number of normal examples falsely predicted as anomaly. FN, false negative is the number of anomaly examples falsely predicted as normal.

$C(X)$ in formula (1) can be derived as follows:

$$\begin{aligned} C(X) &= C(X(x_i)), \alpha = 1, \beta = 2, \eta = 3, i = 1, 2, \cdots, 41. \\ &= \alpha(x_\alpha) + \beta(x_\beta) + \eta(x_\eta) \\ &= \alpha(x_1 + x_2 + x_3 + x_5 + x_6) + \beta(x_7 + x_9 + x_{23} + x_{24} + x_{29} + x_{30} + \cdots + x_{37}) \\ &\quad + \eta(x_4 + x_8 + x_{10} + x_{11} + \cdots + x_{22} + x_{25} + \cdots + x_{28} + x_{38} + \cdots + x_{41}) \end{aligned} \tag{4}$$

Where, $X(x_i)$ is an individual with 41 features and x_i is the ith feature in the individual with a binary value .The coefficients α, β, η mean S-small, M-medium and H-high of the cost for feature generation, respectively. $x_\alpha, x_\beta, x_\eta$ are sets of features with the coefficient α, β, η, respectively.

After the values of objective function for all individuals in the current generation have been calculated, the GA starts to generate new individuals for next generation. We build a roulette wheel based on the fitness of individuals for parental selection and use single-point crossover and a bit level mutation. Moreover, we adopt the niche technology with the crowding mechanism in the evolution for maintaining the diversity of population. Two parent individuals compete with two offspring based on Hamming distance. In order to prevent losing the high-quality individuals generated during the evolutionary process, an elite strategy, that the optimal individual is directly selected to the next generation, is adopted to improve the efficiency of GA. The steps above are iteratively executed until the number of iterations reaches 300. In order to reduce computational time, a simple validation is applied during the process of optimization and the train data set of NSL-KDD is split into new train data set (70% of examples) and new test data set (30% of examples). After significant features subset is obtained, 5-fold cross validation is adopted to guarantee the validity of results obtained by simpler validations (Table 2).

3 DATASET AND DATA PREPROCESSING

In this paper, the NSL-KDD (2009) data set is applied to demo the superiority of the proposed algorithm. NSL-KDD comes from the KDD'99 data set which has some inherent problems (J. McHugh 2000). NSL-KDD is proposed to solve some of these problems by Tavallaee (E. Bagheri et al, 2009). The most recent reference works in IDSs have adopted the NSL-KDD data set to evaluate their procedures (S. Lakhina et al, 2010) (E. de la Hoz et al, 2014) (F. Iglesias et al, 2014). We use the 20% subset of the NSL-KDD training data (25,192 examples) for our experiments.

Data preprocessing includes encoding nominal variables and normalization. The NSL-KDD data set consists of 41 features, which could be classified into three types: continuous, nominal and binary features. We just simply replace each different values of a nominal feature with an integer number. In order to avoid features with greater values dominating those features with smaller values, data normalization is needed to perform before conducting experiments. In this paper, continuous variables are normalized to zero mean and unity variance using the Eq. (5).

$$\hat{x} = \frac{x - \overline{x}}{\sigma} \tag{5}$$

Where $\overline{x}$ and σ are the mean and the standard deviation of variable x, respectively.

4 EXPERIMENTS RESULT AND DISCUSSIONS

As show in Eq. (2) and (3), we choose DR and FNR to evaluate the proposed scheme. The experiments environment is an Intel Core 2.90 GHz computer, in which RAM is 4 GB and operating system is Windows 7. The experimental codes are written in the R programming language. In our experiments, the size of initial population is 100, crossover rate is 0.9, mutation rate is 0.05, and the maximum number of iteration is 300.

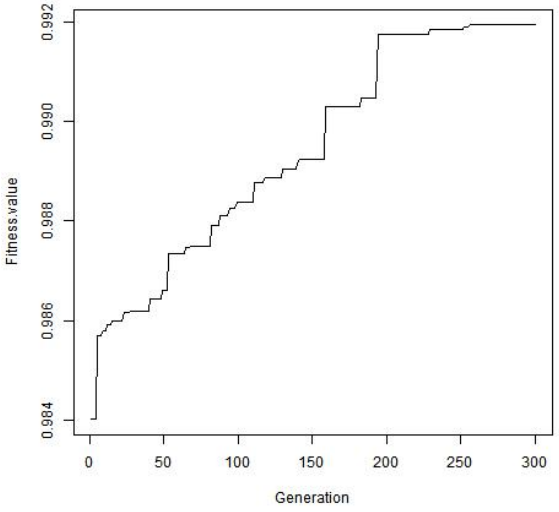

Figure 2 Fitness Value over Generations

Table 1 Improved GA Feature Selection Results

Generation	Number of Neighbors Considered	Number of selected features	DR (%)	FPR (%)	Costs
01-50	k = 9	1,3,4,5,6,8,9,10,14,15,17,19,22,23,26,38,40,41 (18)	99.69	1.08	44
51-100	k = 11	1,2,3,4,5,6,8,9,10,15,17,20,25,26,27,41 (16)	99.72	0.95	37
101-150	k = 11	1,2,3,4,5,6,7,8,9,10,14,15,17,20,25,27,40,41 (18)	99.74	0.87	42
151-200	k = 11	2,3,4,5,7,8,9,10,12,14,15,17,20,25,27,40,41 (17)	99.87	0.75	43
201-250	k = 9	2,3,4,5,7,8,9,10,12,14,15,17,25,27,40,41 (16)	99.87	0.74	40
251-300	k = 9	2,3,4,5,7,9,10,12,14,15,17,25,27,40,41 (15)	99.87	0.74	37

Detail experimental results are shown in Figure 2 and Table 1. As shown in Figure 2, the increase of fitness value of the optimal individual is not obvious after 200 generations, so we can conclude that the final generation is well-optimized. Looking at Table 1, with the evolution of genes, DR is increasing, and FPR is decreasing as same as the costs for features generation. By comparing the generation 251-300 with the generation 151-200, more features do not always ensure better performance of classification because some features are redundant and irrelevant.

Table 2 shows the performances of k-NN, GA+OWNN, and IGA+OWNN methods after 5-fold cross validation. Compared with k-NN and GA+OWNN, our proposed method not only achieves desired results in terms of DR and FNR, it also considerably decreases the costs for features generation, and the dimension of feature space is reduced from 41 to 15.

Table 2 Evaluation of Significant Subsets after 5-fold Cross Validation

Method	Number of Neighbors Considered	Number of selected features	DR (%)	FPR (%)	Costs
k-NN	$k = 1$	all (41)	99.30	0.77	100
GA+OWNN	$k = 7$	2,3,8,10,21,23,25,27,28,32,33,34,35,36,38,39,41 (17)	99.39	0.72	41
IGA+OWNN	$k = 9$	2,3,4,5,7,9,10,12,14,15,17,25,27,40,41 (15)	99.66	0.70	37

5 SUMMARY

In this paper, we present an improved GA based feature selection method for anomaly detection. We apply the crowding mechanism based niche technology to improve the global optimization ability and convergence speed of GA, and adopt OWNN classifier as an evaluator to improve the classification performance with the selected features subset. Experiments results show that this method can effectively reduce the dimension of feature space with desirable detection rate and false negative rate, and a features subset with lower generation costs is selected. In the future, we hope to use multi-object GA and other classifiers to increase the performance of anomaly detection and verify it under more realistic environments.

ACKNOWLEDGMENTS

This work is supported by the National High Technology Research and Development Program of China (Grant No. 2015AA7115089) and the National Natural Science Foundation of China (Grant No. 61374008).

REFERENCES

A.S. Eesa, Z. Orman, A.M.A. Brifcani, A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems, EXPERT SYSTEMS WITH APPLICATIONS. 42 (2015) 2670-2679.

E. Bagheri, W. Lu, A.A. Ghorbani, A detailed analysis of the KDD CUP 99 data set, IEEE Symposium on Computational Intelligence for Security and Defense Applications, CISDA 2009.

E. de la Hoz, A. Ortiz, J. Ortega, A. Martinez-Alvarez, Feature selection by multi-objective optimisation: Application to network anomaly detection by hierarchical self-organising maps, KNOWLEDGE-BASED SYSTEMS. 71 (2014) 322-338.

F. Iglesias, T. Zseby, Analysis of network traffic features for anomaly detection, Machine Learning. 101 (2014) 59-84.

G. Stein, B. Chen, A. Wu, K. Hua, Decision tree classifier for network intrusion detection with GA-based feature selection, ACM'05. 2 (2005) 2136-2141.

H.J. Liao, C.H.R. Lin, Y.C. Lin, K.Y. Tung, Intrusion detection system: A comprehensive review, JOURNAL OF NETWORK AND COMPUTER APPLICATIONS. 36 (2013) 16-24.

Information on http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

J. McHugh, Testing intrusion detection systems: A critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by Lincoln laboratory, ACM Transactions on Information and System Security. 3 (2000) 262-294.

P.G.-T. P. Garcıa-Teodoro, G. Macia-Fernandez, E.Vazquez, Anomaly-based network intrusion detection: Techniques, systems and challenges, COMPUTERS & SECURITY.28 (2009) 18-28.

P. Tang, R.A. Jiang, M. Zhao, editors. Feature selection and design of intrusion detection system based on k-means and triangle area support vector machine, 2010 Second International Conference on Future Networks. (2010) 144 -148.

R.J. Samworth, Optimal weighted nearest neighbour classifiers, ANNALS OF STATISTICS. 40 (2012) 273-2763.

S. Lakhina, S. Joseph, B. Verma, Feature Reduction using Principal Component Analysis for Effective Anomaly-Based Intrusion Detection on NSL-KDD, International Journal of Engineering Science and Technology. 2 (2010) 1790-1799.

The NSL-KDD dataset. Information on http://nsl.cs.unb.ca/nsl-kdd.

T. Shon, X. Kovah, J. Moon, Applying genetic algorithm for classifying anomalous TCP/IP packets, Neurocomputing. 69 (2006) 2429-2433.