ISSN: 1813-4890

# Discovering Comprehensible Anomaly Detection Rules based on an Scalable Genetic Algorithm in Computer Cluster

Lei Zhao

School of Electronic Information, Shanghai Dianji University, Shanghai 201306, China

## Abstract

**The development of big data and cloud computing has made it a critical task for detecting anomalous behaviors and events in computer network technology. This work presents a scalable genetic algorithm (SGA) based classification algorithm to discover comprehensible IF-THEN rules for network anomaly detection using the big data of Management Information Base. The SGA can be used to find anomaly detection classification rules from a cluster of computers with a message passing interface (MPI) standard. A new chromosome encoding and a new fitness function scheme are also given. Experiments are done on a real data set of a large-size company using the algorithm. Both the theoretical and experimental results show that the method is effective and efficient for network anomaly detection in computer clusters.**

## Keywords

**Scalable Genetic Algorithm; Classfication Rule; Network Anomaly Detection.**

## 1. Introduction

Recently, plenty of techniques and algorithms have been proposed for network anomaly detection including machine learning methods [1], [2], statistical analysis [3], [4], the rule based method [5], and pattern matching methods [6], [7]. Most of these methods are designed to create models for normal or abnormal data and then attempt to detect deviations from the models in observed data. For example, the machine learning method is regarded as a powerful tool for training labeled network data i.e., with instances pre-classified as being an abnormal or not [2].The machine learning tools can automatically train and build detection rules and models. As a result, machine learning methods have become unique for the last decade.

In this paper, a new method is proposed for network anomaly detection using the big data in MIB based on a scalable genetic algorithm (SGA). Similar work has been done using the genetic algorithm in medical domain on a non-big data basis, for example, diagnosis of breast cancer and prediction of recurrence of breast cancer [8]. We also propose a new chromosome encoding scheme and a new fitness function design method for finding comprehensible rules. Experiments are done on a real data set of a large-size company using the algorithm. Both the theoretical and experimental results show that the method is effective and efficient for network anomaly detection in computer clusters.

The paper is organized as follows. Section 2 briefly reviews the basic characteristics of genetic algorithms for the task of classification task. Section 3 gives the proposed method for discovering comprehensible classification rules using GA in cluster. In Section 4, experiments and evaluation are carried out. The paper will conclude in Section 5.

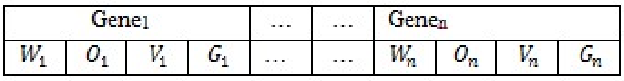
# 2. Overview of Classfication Rule Construction based on Genetic Algorithm

## 2.1. Classfication Rule Representation

Classification is one of the most important tasks in machine learning and data mining, as well as has been used for different areas of intelligent systems. The classification on a database consists of assigning records to one of a set of pre-defined classes which is designated by the researchers. The discovered knowledge is usually represented in the form of IF-THEN prediction rules. In this context, a typical IF-THEN rule can be represented as IF (A1 op V1 ∧ A2 op V2 ∧ …∧ An op Vn ) THEN C, where Ai (i=1,2,…,n) are attributes of the data set used for prediction, op is relational operators such as =, <, and > etc. In the simplest case, there is only one attribute for prediction. The discovered rules can be evaluated by several criteria, such as the classification accuracy rate, the confidence of the prediction and comprehensibility, etc.

## 2.2. Chromosome Encoding

We use the Michigan encoding method in which one record in the database corresponds to one single chromosome (individual) [8]. In the natural environment, a chromosome is composed of n genes, where each gene corresponds to a condition in the IF part of a rule, and the entire chromosome is in line with the whole IF part of the rule. Figure 1 shows a typical encoding scheme for chromosome designed in this work. It can be seen that one gene is composed of four parts including Wi,Oi,Viand Gi. Wiis the weight field which taking real values between [0..1]. The field operator Oi is a variable taking relational operators such as "= ", "< " and "> ". Vi is the value field contains one of the values belonging to the domain of attribute Ai. Gi is the information gain field consists of the value of the information gain for attribute Ai, see Fig. 1.



| Gene1 | | | | … | … | Gene$_n$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| $W_1$ | $O_1$ | $V_1$ | $G_1$ | … | … | $W_n$ | $O_n$ | $V_n$ | $G_n$ |

**Fig. 1** Encoding scheme for a chromsome (individual)

## 2.3. Fitness Function

The fitness function is designed for evaluating the quality of each classification rule. In this paper, we consider the four kinds of rules shown in Table 1.

**Table 1.** Four Kinds of Classification Rules

| ClassficationRules | If A<br>Then C | If A<br>Then not C | If not A<br>Then C | If not A<br>Then not C |
|---|---|---|---|---|
| Number of the rule | pp | pn | np | nn |

As shown on the above table, there are four kinds of classification rules including "If A Then C","If A Then not C", "If not A Then C" and "If not A Then not C". We compute the numbers of these rules respectively, for example, the number of the rule "If A Then C" is pp.

In [8], the fitness function is combined by two indicators, namely the sensitivity (Se) and specificity (Sp), defined as:

$$Se=pp/(pp+nn) \tag{1}$$

$$Sp=pn(pn+np) \tag{2}$$

The fitness function is computed as the product of Sp and Se, i.e.:

$$fitness=Se \times Sp \quad (3)$$

# 3. Discovering Comprehensible Anomaly Rules based on Scalable Genetic Algorithm in Computer Cluster

## 3.1. Criteria for Evaluating Comprehensibility of Classification Rules

In [8], the comprehensibility of a classification rule is considered that has relation with the rule's length, i.e., the simpler the rule, the more comprehensible the rule. We believe that the way of relating the length of the rule with its comprehensibility can lead to a new indicator called simplicity (Si). In this paper, we present a new method for computing the comprehensibility of a classification rule. We believe that a comprehensible classification rule can provide more information than a common one. Furthermore, we introduce the concept of attribute's information gain for computing the comprehensibility of a classification rule. The information gain of an attribute is defined by:

$$InfoGain(A)=I(S1,S2,\ldots,Sm)-E(A) \quad (4)$$

where I(S1,S2,…,Sm) is the information needed for classification by a given attribute A. E(A) is the entropy of attribute A[1]. The comprehensibility of a classification rule is defined as:

$$compre= \sum InfoGain(Ak)/ \sum InforGain(An) \quad (5)$$

where k is the length of the rule and n is the number of all attributes used for classification. The comprehensibility of the classification rule is the ratio between the information gain of the attributes in the rule and the information gain of all attributes. The more information gain of the attributes used for prediction, the more the comprehensibility of the rule.

## 3.2. Fitness Function

To obtain a comprehensible classification rule, we modify (3) as:

$$fitness=w1 \times Se \times Sp+w2 \times compre \quad (6)$$

where w1 and w2 are weights and satisfy w1+w2=1.

## 3.3. Mining Genetic Algorithm on a Computer Cluster

In order to scale to very large data sets, we first connect the machines of different departments to form a computer cluster environment. Then we distribute the data to multiple machines in the cluster. The data is distributed equally between the machines in parallel. For a cluster of N machines, a distribute method can be used and each of the machine has 1/N of the whole data . Finally, the GA algorithm is run on the machine respectively and the final result of the GA classification is obtained by merging the partial results from all the machines in the cluster.

Algorithm 1 is designed for building a distributed GA index. It includes a broadcast process and a merge process. Each process in the cluster runs in parallel and reads a fraction of the network data set. All processes build the GA index in parallel using their respective data set.

In order to search the distributed index the query is sent from a client to one of the computers in the MPI cluster, which is called the master server, see Fig. 2.

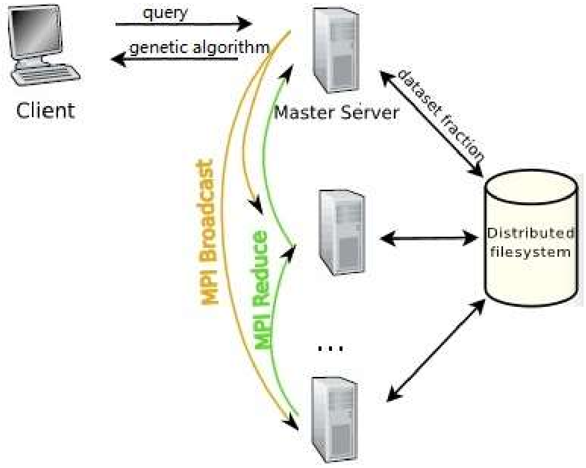


**Fig. 2** The scaling genetic algorithm on a cluster using MPI standard.

```
Algorithm 1 Scaling a distributed index on a computer cluster
Input: query Q, GA parameters P.
1: MPI broadcast(Q,P).
2: run genetic search with query Q and parameters P.
3: merge results using a MPI reduce operation.
4: return the results.
```

The master is the main server and is located at the center of the cluster. It can broadcast the query to all of the processes in the cluster running theirs own GA on fraction of the data. When the classification is complete, a message passing interface reduce operation is used to merge the results back to the master process and the final result is returned to the client. The server can also be used to assign computing tasks. For example, it can automatically detect the running abilities of all client computers and create an assign table. Client computers with higher computing abilities can be assigned with more percentage of the whole task while a poor client computer with a lighter task. All client computer runs equally and using a merge algorithms to form the final results. We designed the task assigning algorithms shown in algorithm 2.

```
Algorithm 2 Task assigning algorithm
Input: computer parameters of each client.
1:  server send the task assign query s.
2:  each client receive s.
3: client computer own computing abilities ai.
4: client send ai to the server.
5: server buid the assign table.
6: server sends the assign task to each client.
```

## 3.4. Experiments

In our experiment, we consider four kinds of anomaly activities in computer network including 100% CPU occupancy failure, application server down failure, database shutdown failure and worm attack failure which are injected to the network artificially. We select 720 records (24 hours) including 11 attributes. Table Ⅵ shows four rules obtained using GA, one rule for each class. For each class, the algorithm was run three times. In all experiments, the initial population size is set to 720 and the maximum generation size is 200. The probability of selection, crossover and mutation are 80%, 100% and 5% respectively.

For comparison, we obtained the corresponding false positive and detection rate of different classification techniques. Figure 5 shows ROC (Receiver Operator Characteristic) curves of three classification methods, the Gas, the decision tree and the SVM method. To more accurately compare these methods, we calculated the performance value using the area under the ROC and presented the figures in Table Ⅶ. It is meaningful to notice that the GAs method has a larger area than the other two methods.

**Table 2.** Discovered Rules of Four Kinds Network Failure

| Class | Rule | Fitness |
|---|---|---|
| CPU occupancy | If ifInOctets<5000∧ifInUcastPkts =0∧ifOutOctcts =0 | 0.912 |
| Application server shutdown | If ifInOctets<2000∧ifOutDiscards =0∧ifOutOctcts =0 | 0.869 |
| Database shutdown failure | If ifInOctets<15000∧ifOutDiscards =0∧ifOutOctcts =0 | 0.823 |

## 4. Conclusion

We have proposed a new approach for network anomaly detection by mining comprehensible classification rules from big data in MIB. Experimental results show that the method is effective and efficient. A key contribution of our work is the design of a scalable genetic algorithm using MPI standard for mining big data. Another contribution is that we have designed a new chromosome encoding scheme and a new fitness function scheme. Future work also includes design more effective and efficient MPI procedures and task assigning algorithm.

As future work, we plan to further investigate the big network anomaly detection by discover classification rules with other indicators of the network system by different classification methods. The correlation of these indicators will also be taken into account to extract useful and non-redundant rules.

## References

[1] M. Thottan and C. Ji, “Anomaly detection in IP networks,” IEEE Trans. Signal Processing, pp. 2191–2204, August 2003.

[2] A. Patcha and J.M. Park, “An overview of anomaly detection techniques:Existing solutions and latest technological trends,” Computer Networks, vol.51, no.12, pp. 3448–3470, 2007.

[3] W. Lee and D. Xiang, “Information-theoretic measures for anomaly detection,” IEEE Symp. on Security and Privacy, pp. 130–143.

[4] A. Lakhina, M. Crovella, and C. Diot, “Mining anomalies using traffic feature distributions, ” in Proc. ACM SIGCOMM 2005, Aug. 2005.

[5] P. Gogoi, B. Borah and D. K. Bhattacharyya, “Network Anomaly identification using supervised classifier”, Informatica,vol.37, pp.93–105, July 2013.

[6] M. Thottan and C. Ji. “Adaptive thresholding for proactive network problem detection,” IEEE Network: Special Issue on Network Management, Oct. 1998.

[7] P. Gogoi, B. Borah and D. K. Bhattacharyya, “Network Anomaly Identification using Supervised Classifier,” Informatica,vol.37, pp.93-105, July 2013.

[8] A. A. Freitas, “A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery,” In: A. Ghosh and S. Tsutsui (eds.): Advances in Evolutionary Computation,Springer-Verlag, pp. 819–845, 2001.

[9] J. Manyika, M. Chui, B. Brown, “Big data: The next frontier for innovation, competition, and productivity,” McKinsey Global Institute, pp.15–17,2011.