Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 40 (2007) 2373–2391
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

Genetic-fuzzy rule mining approach and evaluation of feature selection techniques for anomaly intrusion detection

Chi-Ho Tsang, Sam Kwong*, Hanli Wang

Department of Computer Science, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong, PR China

Received 7 June 2005; received in revised form 10 July 2006; accepted 13 December 2006

Abstract

Classification of intrusion attacks and normal network traffic is a challenging and critical problem in pattern recognition and network security. In this paper, we present a novel intrusion detection approach to extract both accurate and interpretable fuzzy IF–THEN rules from network traffic data for classification. The proposed fuzzy rule-based system is evolved from an agent-based evolutionary framework and multi-objective optimization. In addition, the proposed system can also act as a genetic feature selection wrapper to search for an optimal feature subset for dimensionality reduction. To evaluate the classification and feature selection performance of our approach, it is compared with some well-known classifiers as well as feature selection filters and wrappers. The extensive experimental results on the KDD-Cup99 intrusion detection benchmark data set demonstrate that the proposed approach produces interpretable fuzzy systems, and outperforms other classifiers and wrappers by providing the highest detection accuracy for intrusion attacks and low false alarm rate for normal network traffic with minimized number of features.

Keywords: Fuzzy classifier; Genetic algorithms; Multi-objective optimization; Feature selection; Intrusion detection

1. Introduction

Intrusion detection based on statistical pattern recognition approaches has attracted a wide range of interest over the last 10 years in response to the growing demand of reliable and intelligent intrusion detection systems (IDS), which are required to detect sophisticated and polymorphous intrusion attacks. In general, intrusion detection approaches are usually categorized into misuse and anomaly detection approaches in the literature. Misuse detection approach can reliably identify intrusion attacks in relation to the known signatures of discovered vulnerabilities. However, emergent intervention of security experts is required to define accurate rules or signatures, which limits the application of misuse detection approach to build intelligent IDS. On the other hand, the anomaly detection approach usually deals with statistical analysis and pattern recognition problems. It is able to detect novel attacks without a priori knowledge about them if the classification model has the generalization capability to extract intrusion pattern and knowledge during training. Unfortunately, it commonly suffers from high false positive rate (FPR) on classifying normal network traffic nowadays. To overcome the anomaly intrusion detection problem, the data mining [1], machine learning [2] and immune system [3] approaches have been proposed in recent years.

Learning classification rules from network data is one of the most effective methods to automate and simplify the manual development of intrusion signatures, and predict novel attacks if the generalized knowledge can be extracted from data. One of the key challenges in building an anomaly rule-based IDS is to ensure that it can automatically extract optimal classification rules from training data, and the extracted rules should be (i) accurate and sufficient to detect both known and unseen intrusion attacks and recognize normal network traffic, and (ii) linguistically interpretable for human comprehension. To extract rule-based knowledge from network data, Lee et al. [4] propose to apply association rules to capture the behaviours and relations in programs execution and user activities, and use frequent episodes to model the sequential patterns in system

* Corresponding author. Tel.: +852 2788 7704; fax: +852 2788 8614.
E-mail address: cssamk@cityu.edu.hk (S. Kwong).

0031-3203/$30.00
doi:10.1016/j.patcog.2006.12.009

audits or network data. However, since the quantitative features in the intrusion data are partitioned into the interval with crisp boundary, there might exist a sharp boundary problem for classification. In order to solve this problem, the fuzzy logic [5], which provides the partial membership in set theory, is applied in Ref. [6] to integrate with the association rules and frequent episodes. The application of fuzzy logic in intrusion detection can also be found in Ref. [7], which effectively detects port scanning and denial-of-service attacks.

Genetic algorithm (GA) [8] has been successfully applied to solve many combinatorial optimization problems. The application of GA to the evolution of fuzzy rules can be found in Refs. [7,9] for intrusion detection. In Ref. [9], a simple GA is applied to generate and evolve the fuzzy classifiers that use complete expression tree and triangular membership function for the formulation of chromosome. To evaluate the fitness of individual solutions, the weighted sum of fitness values of multiple objective functions is proposed in Ref. [9] where the proposed weights are user-defined and cannot be optimized dynamically for different cases. In Ref. [10], a large number of fuzzy rules are first generated for each class with the use of fuzzy association rules. After that, a boosting GA based on the iterative rule learning approach is applied for each class to search its fuzzy rules required for classification, in which the rules can be extracted and included in the system for evaluation. However, it only optimizes classification accuracy and omits the necessity of interpretability optimization. In Ref. [11], a simple GA is employed as the searching strategy in a feature selection wrapper that applies RIPPER [12] as the induction algorithm for rule learning and classification. The above-mentioned works have somehow successfully demonstrated the effectiveness of applying GA to select feature subset and generate fuzzy rule-based IDS, however, only on the optimization of classification accuracy.

In general, there is always a trade-off between the accuracy and interpretability such that the acquisition of fuzzy IF–THEN rules, which achieves good accuracy, does not imply the fuzzy system is interpretable for human comprehension. As discussed in Refs. [13,14], besides the importance of classification performance, it is also desirable to obtain highly interpretable knowledge in IDS to assist security experts for intrusion analysis. Therefore, the optimizations of both accuracy and interpretability should be necessarily taken into account for building anomaly rule-based IDS. To achieve this goal, a multi-objective genetic fuzzy intrusion detection system (MOGFIDS) is proposed in this work, which applies an agent-based evolutionary computation framework to generate and evolve an accurate and interpretable fuzzy knowledge base for classification. To the best of our knowledge, this is the first work in applying multi-objective genetic fuzzy system concerning with both accuracy and interpretability for anomaly rule-based intrusion detection.

In addition, the proposed MOGFIDS can be considered as a genetic wrapper that searches for a near-optimal feature subset from network traffic data. This helps to reduce the computational overhead for classification and improve the generalization capability of MOGFIDS. Feature selection (FS), which is known to be an NP-hard problem [15], has been extensively studied in the last two decades. Given a set of N features, the goal is to select a desired subset of size M from 2^N possible subsets in order to minimize the classification error and alleviate the curse of dimensionality for computational cost. In general, the optimality of feature subset can substantially improve the interpretability of rule-based classifiers since the optimal minimal number of features minimizes the number of classification rules generated from data. The FS techniques can be broadly classified into filter-based and wrapper-based approaches in the supervised learning paradigm. The filter-based approaches select features using estimation criterion based on the statistics of learning data, and are independent of the induction classifier. The wrapper-based approaches employ induction classifier as a black box using cross-validation or bootstrap techniques to evaluate the feature subset candidates suggested by different search algorithms, such that the accuracy of the classifier can often be maximized. Wrapper-based approaches generally produce better subsets than filter-based approaches, but they are more computationally expensive than filter-based approaches due to the repeated runs of classifier, in particular for very high-dimensional feature domains. As there is no single FS technique that has proven superior for all problem domains, the first sub-goal of this work is to search for a near-optimal feature subset using some well-known filter-based approaches as a baseline reference, and the second sub-goal is to evaluate the effectiveness of MOGFIDS in comparison with some wrapper-based approaches, in searching near-optimal feature subset for intrusion detection.

The rest of this paper is organized as follows. Section 2 highlights the interpretability issues of genetic fuzzy rule-based system (GFRBS). Our proposed multi-objective genetic fuzzy rule-mining approach is described in Section 3 in detail. Section 4 discusses the experimental results including the performance comparisons of MOGFIDS with other feature selection approaches for intrusion detection. Finally, we draw the conclusions in Section 5.

2. Genetic fuzzy rule-based systems and interpretability

Fuzzy rule-based systems, inspired by the fuzzy set theory [5], have been successfully applied to solve many complex and non-linear problems by constructing fuzzy IF–THEN rules for classification and modeling control. GFRBSs employ evolutionary approach to learn and extract knowledge from training data. The optimization criteria in GFRBS include linguistic variables, parameters of fuzzy membership functions, fuzzy rules and the number of rules. In traditional GFRBS, the classification performance and interpretability (also known as transparency), which are often contradictive to each other, are not addressed simultaneously. Redundant fuzzy rules and fuzzy sets, as well as inappropriate fuzzy set topology would be undesirably constructed if the interpretability criterion is not optimized. The poor interpretability of such fuzzy systems can potentially degrade the performance as well as the usefulness of fuzzy rule-based IDS. In this section, we briefly discern the interpretability with the following factors, which are discussed in our previous work [16] in detail.