A Novel Methodology for Database Knowledge Discovery

RICHARD A. WASNIOWSKI
Computer Science Department
California State University
Carson, CA 90747, USA

Abstract: - This paper presents the rough set and genetic algorithms application to knowledge discovery in databases (RSGAKD). The purpose of the methodology is to use specified data for knowledge extraction from computer security logs. The methodology is outlined in terms of its objectives, scope, constraints, assumptions, and tools. The framework introduces rough set based knowledge approach. Where appropriate the procedures associated with the RSGAKD methodology are described.

Key-Words:- Database, Knowledge Discovery, Intrusion Detection, Rough Sets.

1 Introduction

Intrusions detections for computer systems are rapidly becoming one of the most important threats. A major concern is the high rate of false alarms produced by Intrusion Detection Systems which undermine the applicability of such systems. When the attacks are analyzed it is observed that most of them are similar and it is generally agreed that classification can help designers and programmers to better understand attacks and build more secure detection and response systems. In this paper we suggest the extraction of knowledge from computer security logs to learn various attack patterns. The design of practical knowledge extraction systems faces however problems because computer security logs tend to be very large and incomplete. To deal with this problem we investigate the capabilities of Rough Sets approach to data reduction and genetic algorithm. The methodology developed has been applied to a real application and the results are discussed.

2 Mining patterns in event logs

The critical component of knowledge based systems is the knowledge database which contains facts and rules that represent human expert domain knowledge. In order to make a knowledge base complete knowledge experts employ a variety of techniques, such as querying and interviewing, for eliciting information. The term Knowledge Discovery in Databases (KDD) was introduced at the KDD workshop [5]. Knowledge Discovery in Databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. The overall KDD process includes data selection, data preprocessing, data transformation, data mining, interpretation and evaluation. Data mining is a step in the KDD process consisting of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce particular patterns. It is defined as an analytic process designed to explore large amounts of data in the search for consistent patterns and systematic relationships between variables, and to validate the findings by applying the detected patterns to new subsets of data. In this paper we employ rough sets and genetic algorithm for mining patterns in computer security event logs. Rough set theory (RST) introduced by Pawlak [16] provides proper tools to analyze the set of attributes, and to handle vagueness and uncertainty. Using rough set theory, the minimal attribute set can be computed without losing any essential information. Genetic algorithms were introduced by Holland [9].In the following section we introduce rough sets and genetic algorithms.

3. Framework Overview

We developed an intrusion detection architecture called Fuzzy Agent-Based Intrusion Detection System (FABIDS) [15]. The architecture of this system is presented below:

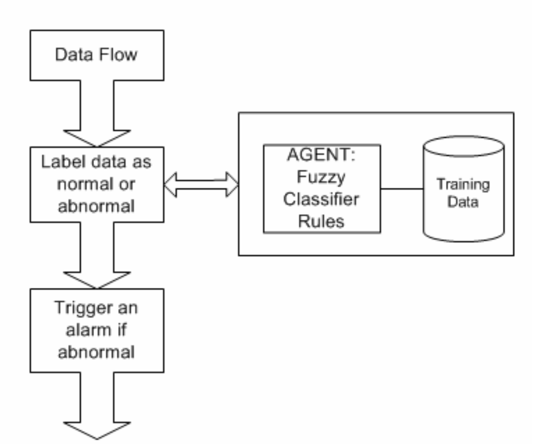
Data Flow
Label data as normal or abnormal
AGENT: Fuzzy Classifier Rules
Training Data
Trigger an alarm if abnormal

Fig.1: FABIDS Architecture

The system consists of multiple intelligent decision support modules, such as fuzzy inference module, classifier, database, etc. This framework integrates several modules such as data collecting sensors, database, fuzzy pattern classification etc [15]. The heart of the controller inference engine is a set of if-then rules whose antecedents and consequences are made up of linguistic variables and associated fuzzy membership functions. Consequences from fired rules are numerically aggregated by fuzzy set union and then defuzzified to yield a single crisp output as the control. Since the differences between the normal and abnormal activities are not distinct, but rather fuzzy, the Fuzzy Inference module can reduce false alarms in determining intrusive activities. For detailed description see [15]. We are processing log files using event correlation engine similar to Risto Vaarandi's powerful Perl event correlation engine described in his article 'A Data Clustering Algorithm for Mining Patterns From Event Logs' published in Proceedings of the 2003 IEEE Workshop on IP Operations and Management [18]. Our algorithm, with similarities to the Apriori and Max-Miner algorithms is implemented in Java.

The framework for knowledge discovery from computer security logs simulation results includes the following main parts. Data quality improvement tools, computer security logs can include anomalous data and it is indispensable to use anomaly detection tools and query language tools to detect inconsistencies that might exist in the database. Knowledge discovery tools, the data mining technique to be applied depends very much on the application domain and the nature of the data available. Knowledge verification tools, the set of discovered rules has to be verified for accuracy, consistency and usefulness for the knowledge base being developed (see Fig. 2)

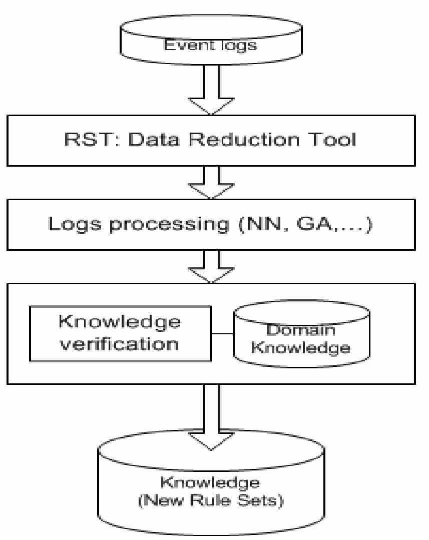

Fig. 2: A framework for rule discovery from log files.

This paper focuses on how to integrate two main concepts of this framework: rough sets and genetic algorithms

3.1 Rough Set Based Data Reduction

The rough set theory, introduced by Z. Pawlak provides sound mathematical tools to deal with inconsistencies in data sets. The rough set approach does not require any preliminary or additional information about data such as probability in probability theory or grade of membership in fuzzy set theory. In the rough set theory inconsistencies are not removed from consideration. Instead, lower and upper approximations of the concept are computed. An information system (IS) consists of a set of objects. IS is usually represented as a pair, $S = (U, A)$, where U and A are finite, nonempty sets called the universe and a set of attributes. The domain of a, denoted Va, is associated with every attribute, $a \in A$. Then, an indiscernibility relation is defined as a binary relation I(B) on U for any subset B of A: (x, y) ∈ I(B) if and only if a(x) = a(y) for every a ∈ A, where a(x) denotes the value of attribute a for element x. I(B) is an equivalence relation on U, the partition determined by B is denoted by U/I(B) or simple U/B. An equivalence class of I(B) containing x is denoted by B(x), and objects x and y are B-indiscernible if (x, y) belongs to A reduct is "a minimal set of condition attributes that preserves the degree of dependency. Some redundant attributes can be removed from the IS without loss of information based on the concept of a reduct. However, some attributes cannot be removed to keep the information, which means an attribute has different degree of significance from another one Applications of Rough Set Theory for Data Mining The rough set approach, a non-numerical analysis, has been used in many applications for rule extraction, classification, and recognition of dependencies among attributes. The rough set method can be a successful data-mining tool by itself. However, it can also be utilized as a part of data mining to find out the relationships between data. In both cases, rough set methods have been successfully applied to analyze a set of data. In this section, rough set applications are introduced with an emphasis on rough set aspects. ROSETTA is a software system for data mining based on discernibility analysis. ROSETTA is applicable to general data mining and knowledge discovery process; the system can be used for preprocessing of data, computing reducts, generating rules, and validating the extracted rules. Outcomes of ROSETTA can be represented in various forms such as tables, reducts, rules, confusion matrixes, partitions, and set approximations. In addition, ROSETTA has the following features: preprocessing of data tables with missing values, discretization of numerical attributes, filtering of reducts and rules according to specified evaluation criteria, and classification of new objects with synthesized rules using voting schemes. The rules generated by ROSETTA are if-then rules, which has a conjunctive antecedent and a disjunctive consequent. Discovered patterns are validated based on quantities such as support counts, probabilities, and user-supplied information about attribute costs.

3.1 Genetic Algorithms

The Genetic Algorithms (GA), pioneered by Holland [9] is an algorithm that transforms a population of individual objects each with an associated fitness value, into a new generation of population. It uses the principle of reproduction and survival of the fittest and naturally occurring genetic operations such as crossover, recombination which is in fact the principal genetic operation, and mutation. Each individual in the population represents a possible solution to a given problem. The GA attempts to find a best solution to the problem by genetically breeding the population of individuals. Crossover is the principle genetic operator employed, with mutation included as an operator of secondary importance. The three dominant different approaches of crossover are one-point crossover, two-point crossover and uniform crossover. With one-point cross-over one value representing the position in the string is chosen at random, and the values of the bit string are interchanged to one side of particular point. With two-point crossover two positions representing the positions in the string are chosen at random, and the values of the bit strings in-between the two values are interchanged. In uniform crossover individual vector positions between parents are swapped with a 50% probability. It must be pointed out that individuals are selected for recombination almost always based on their fitness. In preparing to use the conventional GA operating the user must determine: the representation scheme, the fitness measure, the parameters and variables for controlling the algorithm, and finally a way of designating the result and a criterion for terminating a run. The primary parameters for controlling the GA are the population size, M, and the maximum number of generations to be run. Each run of the GA requires specification of a termination criterion for deciding when to terminate a run and a method of result designation.

4 Experiments

In our experiments we used various sets of data such as the KDD Cup 1999 Intrusion detection contest data, SNORT logs and real-time data. The KDD Cup 1999 data was prepared by the 1998 DARPA Intrusion Detection Evaluation program by MIT Lincoln Labs. Lincoln labs set up an environment consisting of a local-area network simulating a typical U.S. Air Force LAN. They acquired nine weeks of raw TCPdump data that was processed into connection records. The original data contains 744MB of data with 4.94 million records. The dataset has 41 attributes for each connection record plus one class label specifying one of 24 attacks or normal condition.

All these attacks fall into four major categories: Denial of Service (DoS), Remote to User (R2U), User to Root (U2R) and Probing (Probe). While conducting the research for this paper we were provided full access to the SNORT logs from one of the departmental servers [15] The FABIDS sensors were also monitoring network devices: a NAT Router/Firewall and a web server. The NAT Router/Firewall was configured to allow Internet access to the web server, by mapping selected ports to the web server behind the NAT Router/Firewall on the internal network. This configuration was selected to allow a single attack to simultaneously attack the NAT Router/Firewall and the web server so we could generate events logs with identical timestamps to ensure that we could successfully merge data from multiple sensors.

5 Conclusions

In this paper, we have discussed the framework and algorithms for discovering knowledge from computer security logs. The data mining solution is based on the discovery of the principal attributes and the implicit relationships in the given data set. In addition, we have presented techniques based on RS and GA for rule and knowledge discovery. The approach discussed in this paper has been applied to a real application in computer security logs. The methodology discussed in this paper has potential for exploitation in intrusion detection systems design, and will be further developed.

References

[1] Cohen, P.R., and Feigenbaum, E.A., editors (1982) The Handbook of Artificial Intelligence - Vol. 3 AddisonWesley Publishing Co: Reading, MA, U.S.A.

[2] Duntsch, I., and Gediga, G. (1998) "Uncertainty measures of rough set prediction", Artificial Intelligence, no. 106, pp. 109-137.

[3] Efron, B. (1982) The Jackknife, the Bootstrap and Other Resampling Plans. CBMS-NSF Regional Conference Series in Applied Mathematics, no. 38. Society for Industrial and Applied Mathematics: Philadelphia, PA, U.S.A.

[4] Fayyad, U.M., Piatetsky-Shapiro, G., and Smyth, P. (1996) "The KDD process for extracting useful knowledge from volumes of data", Communications of the ACM, v. 39, no. 11, pp. 27-34.

[5] Frawley, W.J., Piatetsky-Shapiro, G., and Matheus, C. (1992) "Knowledge discovery in databases: An overview." AI Magazine, pp. 57-70.

[6] Gawrys, M., and Sienkiewicz, J. (1993) Rough Set Library User's Manual (Version. 2.0, September 1993). Institute of Computer Science, Warsaw University of Technology, Warsaw, Poland.

[7] Gunter, B. (1997) "Tree-based classification and regression - Part 2: Assessing classification performance." Quality Progress, Dec, pp. 83-84.

[8] Nguyen H S,Skowron A. Quantization of real value attributes. Proceedings of Second Joint Annual Conf. on Information Science, Wrightsville Beach,North Carolina,pp34-37,1995

[9] Holland John H. Adaptation in Natural and Artificial System.Ann Arbor: The University of Michigan Press,1975

[10] Zong-Ben Xu, Zan-Kan Nie, Wen-Xiu Zhang, Almost Sure Strong Convergence of A Class of Genetic Algorithms With Parent-Offsprings Competition, ACTA MATHEMATICAE APPLICATAE, SINICA (2002),25(1),(pp167-175)

[11] Krzysztof J.Cios, Witold Pedrycz, Roman W. Swiniarski, Data Mining Methods for Knowledge Discovery, Kluwer Academic Publishers, 1998

[12] Kennedy, G.J. (1993) A Systematic Approach to the Specification of an Information Systems Development System. PhD Thesis, Department of Information Science, University of Otago, Dunedin, New Zealand.

[13] Koperski, K., and Han, J. (1995) "Discovery of spatial association rules in geographic information systems."

[14] M.J., and Herring, J.R., editors, Advances in Spatial Databases: 4th International Symposium, SSD '95, Portland, ME, U.S.A., August 1995, Proceedings. Lecture Notes in Computer Science: Goos, G. and Hartmanis, J. (Eds.) no. 951, pp. 47-66. Springer: Berlin.

[15] R. Wasniowski, RSDKD, January 2005

[16] Pawlak, Z. (1982) "Rough sets." International Journal of Computer and Information Sciences, v. 11, no. 5, pp. 341-356.

[17] Pawlak, Z. (1991) Rough Sets: Theoretical Aspects of Reasoning About Data. Theory and Decision Library. Series D: System Theory, Knowledge Engineering and Problem Solving, no. 9. Kluwer Academic Publishers: Dordrecht, The Netherlands.

[18]Risto Vaarandi, “SEC - a Lightweight Event Correlation Tool”, Proceedings of the 2nd IEEE Workshop on IP Operations and Management, 2002.

[19]D. Bulatovic and D. Velasevic, “A Distributed Intrusion Detection System Based on Bayesian Alarm Networks,” In Proc. of CQRE’99, LNCS 1740, pp. 219–228, 1999.

[20] C.A. Carver, J.M. Hill, J.R. Surdu, and U.W. Pooch. “A Methodology for using Intelligent Agents to Provide Automated Intrusion Response.” In Proc. of the IEEE Systems, Man, and Cybernetics Information Assurance and Security Workshop, West Point, NY, 2000