IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-3, Issue-4 E-ISSN: 2347-2693

Fuzzy Logic and Genetic Algorithm for Data Mining based Intrusion Detection System: A Review Approach

Anshul Atre [1*] and Rajesh Singh [2]

[1,2] Department of CSE, NITM, Gwalior, M.P., India*

anshul011988@gmail.com[1],raj25682@gmail.com[2]

www.ijcseonline.org

Received: April /02/2015 Revised: April/11/2015 Accepted: April/23/2015 Published: April/30/ 2015

Abstract— Along with the modernization of technological era, the technological advancement has also raised concerns about the security of web activities. These activities are in a way or the other are attempted to be compromised by the adversary with the aim of gaining knowledge which may be somehow useful for him/her. In addition, terrorists are also utilizing web for fulfilling their inhuman goals which is currently an utmost concern for security agencies. Although there are many successful attempts have been made to restrict the existence of these illegitimate people, there still is a need for an effective affirmation solution. In respect to this, data mining comes out as a solution by bringing into existence a mining concept named Terrorist Network Mining. Terrorist network mining has proved as the most feasible solution where detection and analysis of terrorists is well performed. Still there were some improvements required to this concept which was efficiently done by combining fuzzy with genetic algorithms with the intrusion detection system (IDS) resulting into significant and efficient detection process. Hence the paper discusses about how well an intrusion detection system performs when combined with fuzzy data mining (reveal patterns whose behavior is intrusive) with genetic algorithm (leads to the success of efficient detection of intruders).

Keywords— Apriority algorithm, Data mining, Fuzzy logic, Genetic algorithm

I. INTRODUCTION

The comprehensive accessibility of the computer system and its data has been a keen area of interest for the intruders. Though it has come into notification that there exist many significant options where the illicit usage is restricted in effective manner there still is the requirement of improved technique. The intrusion detection system (IDS) is one such technique. IDS may be defined as a component of the computer and information security framework whose main goal is to differentiate between the normal activities of the system and behavior that can be classified as suspicious or intrusive [2]. The detection approaches by IDS group as, misuse detection and anomaly detection. In the misuse detection some intuitive data is pre-stored and a set of this intuitive pattern is used to detect the abnormal activities by performing a match of the patterns with the user current activity. While the anomaly detection does not perform any matching instead it traces the deviation of the systems behavior from its normal pattern.

IDS have been used as elite with data mining techniques for detecting the abnormal activities. The IDS combined with data mining adheres to apprehension process of the illegitimate activities. IDS in general, deal with Boolean values but including fuzzy logic in this switch to the discrete and continuous values. The IDS with fuzzy logic when combined with genetic algorithms

IDS have been used as elite with data mining techniques for detecting the abnormal activities. The IDS combined with data mining adheres to apprehension process of the illegitimate activities. IDS in general, deal with Boolean values but including fuzzy logic in this switch to the discrete and continuous values. The IDS with fuzzy logic when combined with genetic algorithms reveals a more effective and efficient apprehension process.

Fuzzy logic incorporated with data mining, association rule mining algorithm named apriori algorithm comes out as fuzzy association rule mining algorithm; where the support and confidence of conventional apriori algorithm are now not the constants instead are fuzzy values. Exceeding to the fuzzy logic in IDS is fuzzy logic along with genetic algorithm. Genetic algorithm lead to a stepping stone towards the success of efficient detection of intruders when combines with IDS along with fuzzy data mining.

II. FUZZY ASSOCIATION RULE MINING

Fuzzy logic is a multi-valued logic or probabilistic logic value, deals with reasoning that is approximate rather than fixed and exact. In contrast with traditional logic theory where binary sets have two-valued logic, true or false, fuzzy logic variables may have a truth value that ranges in degree between 0 and 1. Fuzzy logic has been extended to handle the concept of partial truth, where the truth value may range between completely true and completely false.

The fuzzy logic along with its approximate values provides greater results when fused with the association rule mining of data mining. The problem of mining association rules can be

described as follows: I={i1, i2,..., i } is a set of items, T={t1 , t2 , ... , t } is a set of

Transactions, $\subseteq \subset \subset \phi$

$^{m\,n}$ each of which contains items of the itemset I . Thus, each transaction ti is a set of items such that

ti I. An association rule is an implication of the form: X →Y, where X I, Y I and X ⋂Y=

. X (or Y) is a set of items, called itemset [AR.pdf]. The evaluation of the The support of the rule X →Y defines the percentage of transactions in T that contain

X ⋂Y,

defined| is| applicable to a transaction set T. It is formulated as [3]: supp (X →Y) = |X ⋂Y|/n where, X⋂Y is the number of transactions that contain all the items of the rule and n is the total number of transactions.

Implying how much frequent the rule The confidence of a rule describes the percentage of transactions containing X which also contain Y. It is defines as [3]:

$$\text{Supp}\ (X \rightarrow Y) = \frac{|X \cap Y|}{n}$$

$$\text{Conf}\ (X \rightarrow Y) = \frac{|X \cap Y|}{|X|}$$

The confidence of a rule describes the percentage of transactions containing X which also contain Y. It is defines as [3]:

$$\text{Conf}\ (X \rightarrow Y) = \frac{|X \cap Y|}{|X|}$$

The fuzzy association rule use different The constant s is then calculated as [1]:

$$s = \sum_{\substack{R1 \varepsilon S1 \\ R1 \varepsilon S2}} similarity(R1, R2)$$

The total similarity is given as [1]:

$$\text{Similarity (S1, S2)} = \frac{s}{|S1|} * \frac{s}{|S2|}$$

Interestingness of the rule generated is performed by calculating the two measures: support and confidence. Approach to examine the different and modified quality measures or algorithm derived from association rules. Fuzzy logic is appropriate for the intrusion detection problem because quantitative features such as the number of different connections or messages is often used for anomaly detection, and because security itself involves fuzziness. Hence the association rule combines fuzzy logic to calculate the association between the patterns using apriori algorithm. This extends and improve the conventional method by combining the fuzzy set theory so that it can deal with the both the discrete and continuous attributes in one database which is a normal situation in real world application [4] .
For association calculation, the itemset is considered consisting of fuzzy values. The

similarity between the two rules, R_1 and R_2 is calculated as [1]:

$$Similarity(R1, R2) = \begin{cases} 0 & ,if\ (X1 \neq Y1)\ or\ (X2 \neq Y2) \\ \max\left[0, 1 - \max\left[\frac{|c - c'|}{\max(c, c')}, \frac{|s - s'|}{\max(s, s')}\right]\right] & ,else \end{cases}$$

where, S1 and S2 are the item sets.

The novel approach of embedding fuzzy with apriori algorithm defines the new fuzzy valued support and confidence. Henceforth, the new minimum support (NMS) is calculated as [2]:

Similarly, the new minimum confidence (NMC) is calculated as [2]:

$$NMC = (I - C)\frac{(A - B)}{A}K + C$$

Here, A = Number Features, B = Current Feature, S = minimum Support and C = minimum Confidence. K is a user defined threshold (between 0 and 1) used to control the pruning step. If K increases, the number of rules generated decreases. Only the constraints of rule selection (not the support and confidence) in the candidate generation step of the Apriori algorithm change dynamically.

III. GENETIC ALGORITHM

Genetic algorithm is a heuristic search approach based on the Darwin's theory of natural selection. It consists of certain population, selected randomly, where this population is converted in binary form and is decided on the basis of a decision value named fitness value, selecting values or individuals fit to survive in the population. After deciding the population, the process of mutation (flipping of bits) and

crossover (exchange of bits) is performed to generate new population. The process of generating new population. The process of generating the reference rule set and the normal rule set

the reference rule set and the abnormal rule and Sref.abnorm is the similarity between

$$NMS = (I - S)\frac{(A - B)}{A}K + S$$

The above calculated fitness value set new population is continued till the convergence condition is achieved.

The use of genetic algorithm with data mining technology may improve detection process performance significantly[6], as the population generated every time may be used to distinguish the population into normal and abnormal rule set compared against a reference rule set. The individual is considered performing normal activities is decided by its own characteristic. The goal is to maximize the similarity of the reference rule set and normal rule sets mined from data that contains no intrusions while minimizing the similarity of the reference rule set and abnormal rule set mined from data containing intrusions. Genetic algorithms (GA) are used for feature selection (reducing the running time and improving the accuracy of the Apriori algorithm) and to optimize the membership functions [5]. The fitness value here is modified by which the initial population for GA implementation is generated [2].

$$Fitness = \frac{Sref.norm}{Sref.abnorm}$$

would be a high value causing no rules to be mined. This fitness value calculated is further modified in order to decrease the fitness value of an individual using a user defined threshold p [2].

$$if\,(Nr<Tr)\ Fitness = Fitness * \frac{Nr}{Tr}$$

$$if\,(Ar<Tr)\ Fitness = Fitness * \frac{Ar}{Tr}$$

where,

p is user threshold
Rr is the number of reference set rules
Nr isthe number of normal set rules
Ar is the number of abnormal set rules.

Generally, a prefix tree is constructed to store items with enough support and confidence. The Genetic algorithm combine with Genetic programming forms Genetic Network Programming (GNP) in which directed graph or prefix tree is constructed [4]. Using the NewMinSupp and NewMinConf, the GNP improves the performance by tuning the membership functions. After GA completion, the fuzzy membership function parameters and the selected features are saved.

IV. CONCLUSION AND FUTURE WORK

The ultimate goal of implementing the data mining along fuzzy and GA is to improve the efficiency of the detection process by extracting many rules that are statistically significant and they can be used for several purposes. In the future, the work is being done in the direction of including Fuzzy reasoning. Also the aim is towards on building distributions (probability density functions) of normal and intrusion accesses based on fuzzy GNP and hence then the data can be classified into normal class, known intrusion class and unknown intrusion class. The new Apriori algorithm suggested for improving the efficiency of data mining, by generating the candidate set of variable length itemset in place of fixed, can be implemented within this context where we modify the Apriori algorithm using lemmas with new candidate set generation.

REFERENCES

[1]. Y.Dhanalakshmi and Dr.I. Ramesh Babu, "Intrusion Detection Using Data Mining Along Fuzzy Logic and Genetic Algorithms", IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.2,February 2008.

[2]. German Florez,Susan M. Bridges, and Rayford B. Vaughn, "An Improved Algorithm for Fuzzy Data Mining for Intrusion Detection", Proceedings of Information Processing Society, 2002, Page(s): 457 – 462.

[3]. Jiawei Han & Micheline Kamber (2006) Data Mining;Concepts and Techniques, Second Edition, Morgan Kaufmann Publishers.

[4]. Ci Chen, Shingo Mabu, Chuan Yue, Kaoru Shimada, and Kotaro Hirasawa, "Network Intrusion Detection using Fuzzy Class Association Rule Mining Based on Genetic Network Programming", Proceedings of the 2009 IEEE International Conference on Systems, October 2009.

[5]. Shingo Mabu,, Ci Chen, Nannan Lu, Kaoru Shimada, and Kotaro Hirasawa, "An Intrusion-Detection Model Based on Fuzzy Class-Association-Rule Mining Using Genetic Network Programming", Proceedings of IEEE transactions on systems, man, and cybernetics—part c: applications and reviews, vol. 41, no. 1, January 2011.

[6]. Tan Jun-shan, He Wei1, Qing Yan ," Application of Genetic Algorithm in Data Mining". 2009 First Int. Workshop on Education Technology and Computer Science. 978-0-7695-3557-9/09 © 2009 IEEE. DOI10.1109/ETCS.2009.340. page 353.page 353.