Expert Systems with Applications 37 (2010) 7162–7166

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

# Random effects logistic regression model for anomaly detection

Min Seok Mok, So Young Sohn *, Yong Han Ju

*Department of Information and Industrial Engineering, Yonsei University, 134 Shinchon-dong, Seoul 120-749, Republic of Korea*

ARTICLE INFO

*Keywords:*
Anomaly detection
Intrusion
Random effects
KDD-99

ABSTRACT

As the influence of the internet continues to expand as a medium for communications and commerce, the threat from spammers, system attackers, and criminal enterprises has grown accordingly. This paper proposes a random effects logistic regression model to predict anomaly detection. Unlike the previous studies on anomaly detection, a random effects model was applied, which accommodates not only the risk factors of the exposures but also the uncertainty not explained by such factors. The specific factors of the risk category such as retained 'protocol type' and 'logged in' are included in the proposed model. The research is based on a sample of 49,427 random observations for 42 variables of the KDD-cup 1999 (Data Mining and Knowledge Discovery competition) data set that contains 'normal' and 'anomaly' connections. The proposed model has a classification accuracy of 98.94% for the training data set, while that for the validation data set is 98.68%.



## 1. Introduction

As advances in networking technology help to connect people around the globe, the internet continues to expand its influence as a medium for communications and commerce. At a similar speed, the threat from spammers, system attackers, and criminal enterprises has continually escalated.

Intrusion Detection Systems (IDS) analyze audit trail data to detect any unusual user behavior. In addition, IDS detects hostile activities or exploits in a network (Depren, Topallar, Anarim, & Ciliz, 2005). Although the idea behind intrusion detection is that simple patterns of legitimate user behavior can be captured and the behavior of an anomalous user can be distinguished and identified from normal users (Anderson, 1980), abnormal behavior detection is still a difficult task to implement because of unpredictable attacks (Wang, 2005).

Statistical analysis is the most widely used technique, which defines normal behavior by collecting data relating to the behavior of legitimate users over a period of time (Anderson, Lunt, Javits, Tamaru, & Valdes, 1995). Statistical techniques have been adapted to anomaly detection, which includes principal component analysis (Shyul, Chen, Sarinnapakorn1, & Chang, 2003), cluster and multivariate analysis (Taylor & Alves-Foss, 2002), Bayesian analysis (Barbard, Wu, & Jajodia, 2001), frequency and simple significance tests (Masum, Ye, Chen, & Noh, 2000; Qin & Hwang, 2004; Zhou & Lang, 2003), and multinomial logistic regression (Wang, 2005). Gowadia, Farkas, and Valtorta (2005) adapted the occurrence probability of specific attacks in the existing Bayesian Networks-based anomaly detection system. By observing the input parameters, they were able to anticipate the occurrence probability of specific attacks corresponding to the sequence of input parameters. Lee, Kim, and Kwon (2008) proposed a method for proactive detection of DDoS attacks by exploiting its architecture; which consists of a selection of handlers and agents, communication and compromise, and attack by cluster analysis. Wu and Zhang (2006) presented novel anomaly detection and a clustering algorithm for network anomaly detection based on factor analysis and the Mahalanobis distance. Depren et al. (2007) proposed a novel IDS architecture utilizing both anomaly and misuse detection approaches. The proposed anomaly detection module used a Self-Organizing Map (SOM) structure to model normal behavior. SOM is a neural network model for analyzing and visualizing high dimensional data. Arranz, Cruz, Sanz-Bobi, Ruiz, and Coutino (2008) used neural network for detection of anomalies.

Statistical approaches to anomaly detection have several advantages and disadvantages. First, the disadvantage is that skilled attackers can be accustomed to statistical anomaly detection, also known as the inability to decipher the difference between abnormal and normal behavior. It can also be difficult to determine thresholds that balance the likelihood of false positives with the likelihood of false negatives. In addition, statistical methods need accurate statistical distributions, but not all behaviors can be modeled using purely statistical methods (Patcha & Park, 2007). However, the advantage is not only the ability to detect novel attacks or unknown attacks, but also the systems do not require prior knowledge of security flaws or attacks. Statistical approaches can provide accurate notification of malicious activities that typically

* Corresponding author. Tel.: +82 2 2123 4014; fax: +82 2 364 7807.
*E-mail address:* sohns@yonsei.ac.kr (S.Y. Sohn).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.04.017

occur over extended periods of time and are good indicators of impending attacks.

One of the popular statistical approaches is a fixed effect logistic regression model, which accommodates predictors for anomaly behavior. However, this model does not accommodate variation that cannot be explained by such predictors.

Accordingly, in this paper a random effects logistic regression model is proposed. The advantage of using such a random effects model for anomaly detection is to accommodate not only the network environment characteristics but also the uncertainty that cannot be explained by such network environment characteristics.

The random effects model has been frequently used to accommodate both 'between cluster variation' as well as 'within cluster variation' (Sohn, 1996, 1997, 1999, 2002; Sohn & Choi, 2006; Sohn & Park, 1998).

The outline of this study is as follows: Section 2 introduces the anomaly detection, and Section 3 deals with the random effects logistic regression model for anomaly detection. Section 4 contains an empirical case study and its results. Finally, in Section 5, the results of the study are summarized.

## 2. Anomaly detection

An anomaly is defined as a violation of the security policy of the system; anomaly detection thus refers to the mechanisms that are developed to detect violations of system security policy (Chebrolu, Abraham, & Thomas, 2005; Shiu, Liu, & Yeung, 1997). Anomaly detection is based on the assumption that intrusive activities are noticeably different from normal system activities and are thus detectable. Generally, an anomaly will cause loss of integrity, confidentiality, denial of resources, or unauthorized use of resources.

Anomaly detection seeks to identify activities that vary from established patterns for users, or groups of users. It typically involves the creation of knowledge bases compiled from profiles of previously monitored activities. In addition, anomaly detection component should be integrated in the modeling phase of the knowledge based system (Vanthienen, Muus, & Wets, 1998). Anomaly detection is usually achieved through one of the following:

- Threshold detection, detecting abnormal activity on the server or network, for example abnormal consumption of the CPU for one server, or abnormal saturation of the network.
- Statistical measures learned from historical values.
- Rule-based measures with expert systems.
- Non-linear algorithms such as Neural Networks or Genetic Algorithms (Planquart, 2001).

Sometimes anomalous activities that are not intrusive are flagged as intrusive, although they are false positives. Actual intrusive activities that go undetected are called false negatives. This is a serious issue, and is far more serious than the problem of false positives. One of the main issues of anomaly detection systems is the selection of threshold levels so that neither of the above problems is unreasonably magnified. Anomaly detection is usually computationally expensive because of the overhead of keeping track of and possibly updating several system profiles (Mukkamala, Janoski, & Sung, 2005; Mukkamala, Sung, & Abraham, 2005).

An anomaly detection system is capable of detecting all types of malicious network traffic and computer usage. This includes network attacks against vulnerable services, data driven attacks on applications, host-based attacks such as privilege escalation, unauthorized logins and access to sensitive files and malware. An anomaly detection system is a dynamic monitoring entity that complements the static monitoring abilities of a firewall. The network packets that are collected are analyzed for rule violations by a pattern recognition algorithm. When rule violations are detected, the anomaly detection system alerts the administrator.

## 3. Random effects logistic regression model

In this section, a system attack event is assumed to be an anomaly and a random effects logistical regression model for anomaly detection is proposed. It was assumed that all enterprises can be split into two exclusive subsets: a normal or an attack group. Additionally, each event is associated with information about a network condition, which can be used for the prediction of the event level such as normal or attack. Under such circumstances, the proposed random effects model is as follows.

Let $y_i$ be a binary random variable where it is 1 if an attack is observed for event i, or 0 otherwise ($i = 1,\ldots,n$) and $p_i$ be the probability of attack of event $i$. Then it can be assumed that $y_i$ for a given probability $p_i$ follows a binomial distribution:

$$y_i|p_i \sim \mathrm{Bin}(1,p_i), \tag{1}$$

with the following probability mass function:

$$p(y_i|p_i) = p_i^{y_i}(1-p_i)^{1-y_i}. \tag{2}$$

The conditional mean is $E(y_i|p_i) = p_i$ and the corresponding variance is $V(y_i|p_i) = p_i(1-p_i)$. This conditional mean can vary over different events. Part of the variation can be explained by the network conditional characteristics of events and the remaining part of the variation is due to random error. It is assumed that $p_i$ is random following a beta distribution with the expected value of $p_i$ being a function of related characteristics, $z_{1i},\ldots,z_{ki}$. This means that the expected probability of an attack event would vary due to the network conditional characteristics represented by the $1 \times \boldsymbol{k}$ covariate vector, $z_{1i}\cdots z_{ki}$. The choice of a beta distribution is due to the fact that it aptly describes the distribution of the probability and its conjugate relationship to a binomial distribution. One of the models that describe such a situation is as follows:

$$p_i \sim \mathrm{beta}(c_i, c_i\exp(-\gamma_0-\gamma_1 z_{1i}-\cdots-\gamma_k z_{ki})), \tag{3}$$

with the following probability density function:

$$f(p_i) = \frac{\Gamma(c_i + c_i\exp(-\gamma_0-\gamma_1 z_{1i}-\cdots-\gamma_k z_{ki}))}{\Gamma(c_i)\Gamma(c_i\exp(-\gamma_0-\gamma_1 z_{1i}-\cdots-\gamma_k z_{ki}))}p_i^{c_i-1}(1-p_i)^{c_i\exp(-\gamma_0-\gamma_1 z_{1i}-\cdots-\gamma_k z_{ki})-1}, \tag{4}$$

where $c_i = 1+\exp(\gamma_0+\gamma_1 z_{1i}+\cdots+\gamma_k z_{ki})$ and $\gamma_k$ denote the regression coefficient of $z_{ki}$. Subsequently, the expected value and variance of $p_i$ can be obtained as follows:

$$E(p_i) = 1/(1+\exp(-\gamma_0-\gamma_1 z_{1i}-\cdots-\gamma_k z_{ki})), \tag{5}$$

$$V(p_i) = \frac{\exp(-\gamma_0-\gamma_1 z_{1i}-\cdots-\gamma_k z_{ki})}{(1+\exp(-\gamma_0-\gamma_1 z_{1i}-\cdots-\gamma_k z_{ki}))^2(1+c_i+c_i\exp(-\gamma_0-\gamma_1 z_{1i}-\cdots-\gamma_k z_{ki}))}. \tag{6}$$

Note that the expected probability in (5) forms a logistic regression. Using (2) and (4), the random effects distribution $f(p_i)$ can be updated as follows: