DATAMINING BASED STRATIFIED INTRUSION DETECTION

Prof. Ramamoorthy S
Research Scholar
Sathyabama University, Chennai
Email:mailrmoorthy@yahoo.com

Dr .V. Shanthi
Professor
St.Joseph's College of Engineering, Chennai
Email:drvshanthi@yahoo.co.in

ABSTRACT

The intrusion detection systems focus on low-level attacks, and only generate isolated alerts. They can't find logical relations among alerts. In addition, IDS's accuracy is low; a lot of alerts are false alerts. So it is difficult for human users or intrusion response systems to understand the alerts and take appropriate actions. To solve this problem different intrusion scenario detection methods are proposed. In this paper a data mining based clustering method is used to find the attack scenarios. Usually an attack consists of many steps in which corresponding alerts are generated, so we call each step is an attack scenario. In each step an attacker will perform a task to get certain target. The alerts generated in each step can be used as the feature of corresponding clustering approach.

Keywords: intrusion detection system.

1.INTRODUCTION TO THE INTRUSION DETECTION

Intrusion detection is detection of intrusion behavior, it collects information of the key part of computer network and system, then analyzes them to detect whether occur the action of disobey security strategy [1]. Intrusion Detection System (IDS) is the software or combination of software and hardware to detect intrusion behavior. IDS can examine intrusion attack before system is damaged, and make use of alerting and defense system to deport the intrusion attack. In the process of intrusion attack, It can reduce the loss resulted in [2]. After system attacked, the related attack information is collected, and as security system knowledge, it is added to the strategy set, thus can strengthen system security defence ability, and avoid system being intruded by the same intrusion again.

2.THE USE OF CLUSTERING IN INFORMATION RETRIEVAL

In choosing a cluster method for use in experimental IR (Information Retrieval), two, often conflicting, criteria have frequently been used. To list some of the more important of these: (1) The method produces a clustering which is unlikely to be altered drastically when Further objects are incorporated, i.e. it is stable under growth. (2) The method is stable in the sense that small errors in the description of the objects lead to small changes in the clustering ;(3) the method is independent of the initial ordering of the objects[3].

3.SIMPLE RANDOM SAMPLING:

Simple random sampling is where the researcher has a list which approximates listing all members of the population, then draws from that list using a random number generator or possibly takes every nth subject (called interval sampling). Conventional significance testing is appropriate for simple random samples [4,6,7]. *Simple random sampling with replacement (SWSWR)* is where selections are replaced back into the sampling frame such that repeat selections are possible. *Simple random sampling without replacement (SRSWOR)* does not allow the same random selection to be made more than once [5]. Confidence intervals are slightly (usually trivially) smaller (more precise) for SRSWOR samples compared to simple random samples. Most computer programs use formulas which assume SRSWOR sampling.

4.SIMPLE RANDOM SAMPLING IN STRATIFIED

It is simple random sampling of each stratum of the population. For instance, in a study of college students, a simple random sample may be drawn from each class (freshman, sophomore, junior, senior) in proportion to class size [8,9]. This guarantees the resulting sample will be proportionate to known sizes in the population. One may simultaneously stratify for additional variables, such as gender, with separate simple random samples of freshman women, freshmen men, sophomore women, etc. The finer the stratification, the more precision compared to unstratified simple random sampling [10]. That is, confidence intervals will be narrower for stratified sampling than for simple random sampling of the same population. The more heterogenous the means of the strata are on a variable of interest, the more stratified sampling will provide a gain in precision compared to simple random sampling. Stratified sampling, therefore, is preferred to simple random sampling.

5.VARIOUS STAGES OF STRATIFIED RANDOM SAMPLING

It is where the researcher draws simple random samples from successively more

homogenous groups ("strata") until the individual subject level is reached. For instance, the researcher may sample occupations, then sample companies within occupations, and then sample individual workers. The purpose of stratified random sampling is to increase research precision by ensuring that key populations of subjects are represented in the sample (ex., people in certain job categories). The greater the heterogeneity of the strata and the finer the stratification (that is, the smaller the strata involved) depending on the topic of study. The more the precision of the results. For instance, stratifying by gender at the highest level might well introduce bias in measuring opinions about an item known to be gender-related, whereas stratifying by state would be less likely to introduce a bias since there are more categories (more states than genders) and there is less likely to be a correlation with the opinion item [11]. At each stage, stratified sampling is used to further increase precision. Because the variance of individuals from their group mean in each strata is less than the population variance, standard errors are reduced. This means conventional significance tests, based on population variances, will be too conservative – there will be too many Type I errors, where the researcher wrongly accepts the null hypothesis.

6.AN APPROPRIATENESS OF STRATIFIED HIERARCHIC CLUSTER TECHNIQUE

There are many other hierarchic cluster methods, to name but a few: complete-link, average-link, etc. My concern here is to indicate their appropriateness for data retrieval. It is as well to realize that the kind of retrieval intended is one in which the entire cluster is retrieved without any further subsequent processing of the data in the cluster. This is in contrast with the methods proposed by Rocchio, Litofsky, and Crouch who use clustering purely to help limit the extent of a linear search.

Stratified systems of clusters are appropriate because the level of a cluster can be used in retrieval strategies as a parameter analogous to rank position or matching function threshold in a linear search. Retrieval of a cluster which is a good match for a request at a low level in the hierarchy tends to produce high precision but low recall; just as a cut-off at a low rank position in a linear search tends to yield high precision but low recall [7,12]. Similarly, retrieval of a cluster which is a good match for a request at a high level in the hierarchy tends to produce high recall but low precision [13]. Hierarchic systems of clusters are appropriate for three reasons. First, very efficient strategies can be devised to search a hierarchic clustering. Secondly, construction of a hierarchic system is much faster than construction of a non-hierarchic (that is, stratified but overlapping) system of clusters [14]. Thirdly, the storage requirements for a hierarchic structure are considerably less than for a non-hierarchic structure, particularly during the classification phase.

7.PROPORTIONATE STRATIFICATION

The essence of stratification is the classification of the population into subpopulation, or strata, based on some supplementary information, and then the selection of separate samples from each of the strata. The benefits of stratification derive from the fact that the sample sizes in the strata are controlled by the sampler, rather than being randomly determined by the sampling process [15]. The sample size of each stratum or its allocation is very important to reduce the sampling error. Often the strata sample sizes are made proportional to the strata population sizes; in other words, a uniform sampling fraction is used. This is known as proportionate stratification. In case of proportional stratification, we can treat it in estimation as if it were a SRS [16]. However, sampling error of a sample by proportionate stratification never exceed that of SRS in terms of variance.

Since

$$f = f_1 = f_2 = \Lambda = f_h,$$

$$\hat{\bar{Y}} = \sum_h \left(W_h \times \hat{\bar{Y}}_h \right) = \sum_h \left(\frac{N_h}{N} \times \frac{1}{n_h} \sum_{i=1}^{n_h} y_{h,i} \right) = \sum_h \left(\frac{1}{Nf} \times \sum_{i=1}^{n_h} y_{h,i} \right) = \frac{1}{n} \times \sum_h \sum_{i=1}^{n_h} y_{h,i}$$

$$\hat{V}(\hat{\bar{Y}}) = \sum_h \left(W_h^2 \times \hat{V}(\hat{\bar{Y}}_h) \right) = (1-f) \times \sum_h \frac{W_h \times s_h^2}{n}$$

It may be noted that the variance of the mean based on a proportionate stratified sample is similar to that of a mean based on an SRS [17]. The only difference is that the population element variance $2S$ in the SRS formula is replaced by the weighted average within stratum

$$S_W^2 = \sum_h \left(W_h \times S_h^2 \right)$$

variance in the proportionate stratified formula. As an approximation with large , it can be shown by an analysis of variance

$$S^2 = S_W^2 + \sum_h \left(W_h \times (\bar{Y}_h - \bar{Y})^2 \right)$$

Since the last term in this formula is a nonnegative quantity (a sum of square term), it follows that ; in other words, a proportionate stratified sample cannot be less precise than SRS of the same size. $S^2 \geq S_W^2$; For a given total variability in the population, the gain in precision arising from employing a proportionate stratified sample rather than an SRS is greater the more heterogeneous are the strata means or, equivalently, the more homogeneous are the element values within the strata.

8.CONCLUSION

Proportionate stratification is much used because it produces simple estimators and because it guarantees that the estimators are no less precise than those obtained from an SRS of the same size. The members of the population are grouped into strata, and a sample is taken from each stratum separately at a prescribed rate. Controlling the sample sizes in the strata by the sampler, stratified sampling enhances the representativeness of the sample, thus increases the overall precision of the survey results.

REFERENCE

1. Bartlett, J. E., II, Kotrlik, J. W., & Higgins, C. (2001). Organizational research: Determining appropriate sample size for survey research. Information Technology, Learning, and Performance Journal, 19(1) 43-50.
2. Shoushan LUO. Intrusion detection [M], Beijing University of post and telecommunications Press, 2003.
3. Xiangfeng ZHANG, Yufang SUN. a survey on the development of intrusion detection system[J].computer science.2003,8(30), pp.45-50.
4. A.K. JAIN, M.N. MURTY, P.J. FLYNN. Data Clustering: A Review. ACM Computing Surveys.1999, 3(31), pp.264-270.
5. Dongqi LI.research on clustering algorithm [D].southwest jiaotong university.2007.
6. KDD99CUP Dataset. Http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.
7. J. Brentano S. Snapp and G. Dias et al. Dids (distributed intrusion detection system) motivation, architecture, and an early prototype. In Fourteenth National Computer Security Conference, Washington, DC, October 1991.
8. Eugene H. Spafford and Diego Zamboni. Intrusion detection using autonomous agents. Computer Networks, 34(4):547{570, October 2000.
9. Mark Slagell. The design and implementation of MAIDS (mobile agent intrusion detection system) [R]. Technical Report TR01-07, Iowa State University Department of Computer Science, Ames, IA, USA, 2001.
10. Major Dennis J. Ingram, H Steven Kremer, and Neil C. Rowe. Distributed Intrusion Detection for Computer Systems Using Communicating Agents.
11. Eleazar Eskin, Andrew Arnold, Michael Prerau, Leonid Portnoy and Salvatore Stolfo. A Geometric Framework for Unsupervised Anomaly Detection: Detecting Intrusions in Unlabeled Data. Applications of Data Mining in Computer Security. Kluwer 2002.
12. J. Han and M. Kamber. Data Mining, Concepts and Technique. Morgan Kaufmann, San Francisco, 2001.
13. The third international knowledge discovery and data mining tools competition dataset KDD99-Cup http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, 1999.
14. Foster Provost, Tom Fawcett, and Ron Kohavi. The case against accuracy estimation for comparing induction algorithms. In proceeding of the Fifteenth International Conference on Machine Learning, July 1998.
15. L.Portnoy, E.Eskin and S.Stolfo. "Intrusion detection with unlabeled data using clustering ". Proceedings of ACM CSS Workshop on Data Mining Applied to Security. Philadephia, PA, 2001.
16. S.R.Snapp,J.Brentano,G.V.Dias,T.L.Goan,L.T.Heberlein,C.Ho,K.N.Levitt, B.Mukherjee,S.E.Smaha,T.Grance,D.M.Teal, "DIDS(Distribution Intrusion Detection System) Motivation, Architecture, and An Early Prototype. " Proc. 14th National Computer Security Conference, 1991.
17. B.Scholkopf, J.C.Platt, A.J.Smola. "Estimating the Support of a High- Dimensional Distribution." Neural Computation, 13(7), 2001, pp.1443-1472.1447.

BIOGRAPHY

Prof Subbu Reddiar Ramamoorthy received his bachelors degree in Electrical and Electronics Engineering from PSG College of Technology, Madras University. He obtained his Masters Degree in Industrial Engineering from Indian Institute of Technology, Delhi. He has served in Indian Air Force for 15 years and G.E.C Brunei for 3 years as project engineer. He has more than 2 decades of teaching experience in different engineering colleges and he is currently heading information technology department of Vels Srinivasa College of Engineering and Technology. His research interest includes network security and ethical hacking.

Dr.V.Shanthi is a post graduate Bharadidasan University and Doctorate from Madras University. She has published nearly 15 papers in national and international journals. Also she is a supervisor for research scholars in many universities. She is also a member of board of evaluation and question paper setting. She is in the teaching profession for the last two decades. Her area of interest includes artificial intelligence, data mining and network security. She is also acted as Chairman, Board Evaluation for Anna University and Sastra University.