ANAMOLY DETECTION TECHNIQUES USING SEQUENCES DATA

HARISHBABU.KALIDASU[1]
Asst.Professor, St.Mary's Group of Institutions, Guntur, Andhra Pradesh, INDIA.

CH.VENKATESWARAO[2]
Asst.Professor, HMKS & MGS College of Engineering, Kangala, Guntur, Andhra Pradesh, INIDA.

***Abstract*—Anomaly detection has traditionally dealt with record or transaction type data sets. But in many real domains, data naturally occurs as sequences, and there for e the desire of studying anomaly detection techniques in sequential data sets. The problem of detecting anomalies in sequence data sets is related to but different from the traditional anomaly detection problem, because the nature of data and anomalies are different than those found in record data sets. While there are many surveys and comparative evaluations for traditional anomaly detection, similar studies are not done for sequence anomaly detection. We investigate a broad spectrum of anomaly detection techniques for symbolic sequences, proposed in diverse application domains. Our hypothesis is that symbolic sequences from different domains have distinct characteristics in terms of the nature of sequences as well as the nature of anomalies which makes it important to investigate how different techniques behave for different types of sequence data. Such a study is critical to understand the relative strengths and weaknesses of different techniques. Our paper is one such attempt where we have comparatively evaluated 7 anomaly detection techniques on 10 public data sets, collected from three diverse application domains. T o gain further understanding in the performance of the techniques, we present a novel way to generate sequence data with desired characteristics. The results on the artificially generated data sets help us in experimentally verifying our hypothesis regarding different techniques.**

Introduction

Anomaly detection has traditionally dealt with record or transaction type data sets [5]. But in many real domains, data naturally occurs as sequences, and therefore the desire of studying anomaly detection techniques in sequential data sets. The problem of detecting anomalies in sequence data sets is related to but different from the traditional anomaly detection problem, because the nature of data and anomalies are different than those found in record data sets. While there are many surveys and comparative evaluations for traditional anomaly detection, for e.g., [5, 13, 17], similar studies are not done for sequence anomaly detection. We found only one work comparing the evaluation of four techniques on system call intrusion detection data sets. that compared the performance of four anomaly detection techniques; namely STIDE, t-STIDE (a threshold based variant of STIDE), HMM based, and RIPPER based; on 6 different data sets from system call intrusion detection domain. However, this comparison limits the evaluation to just four techniques, using only system call intrusion detection data sets.

In this paper we investigate a variety of anomaly detection techniques that have been proposed to detect anomalies in symbolic sequences. We classify such techniques as: kernel based, window based, and Markovian techniques. Kernel based techniques use a similarity measure to com-pute similarity between sequences. Window based tech-niques extract fixed length windows from a sequence and assign an anomaly score to each window. Markovian tech-niques assign a probabilistic anomaly score to each event.conditioned on its history , using modeling techniques such as Finite State Automata (FSA), Hidden Markov Models (HMM), and Probabilistic Suffix Trees (PST). We evaluate different anomaly detection techniques and their variations, belonging to the above mentioned classes on a variety of data sets, collected from the domains of pro-teomics [1], system call intrusion detection ,and net-work intrusion detection . Through careful experimentation, we illustrate that the performance of different tech-niques is dependent on the nature of sequences, and the na-ture of anomalies in the sequences. T o further explain the strengths and weaknesses of various techniques, we present a novel method to generate artificial sequences to evaluate anomaly detection techniques, and present further insights using the results on our artificially generated data sets.

II. Problem Statement

The objective of the techniques evaluated in this paper can be stated as follows: Definition 1 Given a set of n training sequences, S, and a set of m test sequences S^T , find the anomaly score A(Sq) for each test sequence Sq ∈ S^T , with respect to S. All sequences consist of events that correspond to a finite alphabet, Σ. The length of sequences in S and sequences in S^T might or might not be equal in length. The training database S is assumed to contain only normal sequences, and hence the techniques operate in a semi-supervised set-ting. we discuss how the techniques canbe extended to unsupervised setting, where S can contain both normal and anomalous sequences.

III. Anomaly Detection T echniques for Se-quences

Most of the existing techniques can be grouped into three categories:

1. Kernel based techniques compute similarity between sequences and then apply a similarity based traditiona lanomaly detection technique.

2. Window based techniques analyze a short windowof events within the test sequence at a time. Thus such techniques treat a subsequence within the test sequence as a unit element for analysis. Such techniques require an additional step in which the anomalous nature of the entire test sequence is determined, based on the scores of each subsequence.

3. Markovian techniques assign a probability to each event of the test sequence based on the previous obser-vations in the sequence. Such techniques exploit the Markovian dependencies in the sequence. In the following subsections we describe several tech-niques that are instantiations of the above three category of anomaly detection techniques.

A. Kernel Based T echniques

Kernel based techniques make use of pair wise similarity between sequences. In the problem formulation stated in Definition 1 the sequences can be of different lengths, hence simple measures such as Hamming Distance cannot be used. One possible measure is the normalized length of longest common subsequence between a pair of sequences.

This similarity between two sequences Si and Sj , is com-puted as: nLC S(Si, Sj) = |LCS(Si ,Sj)|

P |Si||Sj |(1) Since the value computed above is between 0 and 1, nLC S(Si , Sj) can be used to represent distance between Si and Sj .Other similarity measures can be used as well, for e.g., the spectrum kernel . We use nLC S in our experimental study , since it was used in [4] in detectinganomalies in sequences and appears promising.

(a) Nearest Neighbors Based (kNN)

In the nearest neighbor scheme (kNN), for each test se-quence Sq $\in S^T$, the distance to its kth nearest neighbor in the training set S is computed. This distance becomes the anomaly score A(Sq) .A key parameter in the algorithm is k. In our experi-ments we observe that the performance of kNN techniquedoes not change much for $1 \leq k \leq 8$, but the performance degrades gradually for larger values of k.

(b) Clustering Based (CLUSTER)

This technique clusters the sequences in S into a fixed num-ber of clusters, c, using CLARA [16] k-medoids algorithm.The test phase in volves measuring the distance of every test sequence, Sq $\in S^T$, with the medoid of each cluster . The distance to the medoid of the closest cluster becomes the anomaly score A(Sq). The number of clusters, c, is a parameter for this tech-nique. In our experiments we observed that the performance of CLUSTER improved when c was increased, but stabi-lized after a certain value. As c is increased, the number of sequences per cluster become fewer and fewer, thus making the CLUSTER technique closer to kNN technique.

(c) Sparse Markovian T echnique (RIPPER)

The variable Markovian techniques described above allow an event s_{qi} of a test sequence Sq to be analyzed with re-spect to a history that could be of different lengths for dif-ferent events; but they still choose contagious and imme-diately preceding events to s_{qi} in Sq . Sparse Markovian techniques are more flexible in the sense that they estimate the conditional probability of s_{qi} based on events within the previous k events, which are not necessarily contagious or immediately preceding to s_{qi} . In other words the events are conditioned on a sparse history . We evaluate an interesting technique in this category ,that uses a classification algorithm (RIPPER) to build sparse models. In this approach, a sliding window is applied to the training data S to obtain k length windows. The first $k - 1$ positions of these windows are treated as $k - 1$ cate-gorical attributes, and the kth position is treated as a target class. The authors use a well-known algorithm RIPPER to learn rules that can predict the k^{th} event given the first $k - 1$ events. T o ensure that there is no symbol that occurs very rarely as the target class, the authors replicate all training sequences 12 times.

IV. Conclusions and Future Work

Our experimental evaluation has not only provided us in-sights into strengths and weaknesses of the techniques, but have also allowed us to proposed a variant of FSA technique, which is shown to perform better than the original technique. We investigated kernel based techniques and found that they perform well for real data sets, and for artificial data sets with large number of anomalies, but they perform poorly when the sequences have very few anomalies. This can be attributed to the similarity measure. As a future work we would like to in vestigate other similarity measures that would be able to capture the difference between sequences which are minor deviations of each other. We characterize normal as well as anomalous test sequences in terms of the type of patterns (seen-frequent, seen-rare, and unseen) contained in them. We argue that the different window based and Markovian techniques handle the three patterns differently . Since different data sets have different composition of normal and anomalous sequences, the performance of different techniques varies for different data set. Through our experiments we have identified relative strengths and weaknesses of window based as well as Markovian techniques. We have shown that while t-STIDE, FSA, and FSA-z, perform consistently well for most of the data sets, they all have weaknesses in handling certain type of data and anomalies. The original FSA technique proposed estimating likelihoods of more than one event at a time. Using such higher-order Markovian models might change the performance of the Markovian techniques and will be in vestigated in future.

References

[1] A. Bateman, E. Birney , R. Durbin, S. R. Eddy, K. L. Howe, and E. L. Sonnhammer. The pfam protein families database. Nucleic Acids Res., 28:263–266, 2000.

[2] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximiza-tion technique occuring in the statistical analysis of proba-bilistic functions of markov chains. In Annals of Mathemat-ical Statistics, volume 41(1), pages 164–171, 1970.

[3] G. Bejerano and G. Y ona. V ariations on probabilistic suffix trees: statistical modeling and prediction of protein families . Bioinformatics, 17(1):23–43, 2001.

[4] S. Budalakoti, A. Srivastava, R. Akella, and E. Turkov . Anomaly detection in large sets of high-dimensional sym-bol sequences. Technical Report NASA TM-2006-214553, NASA Ames Research Center, 2006.

[5] V . Chandola, A. Banerjee, and V . Kumar. Anomaly detection – a survey . ACM Computing Surveys (T o Appear), 2008.

[6] W. W. Cohen. Fast effective rule induction. In A. Priedi-tis and S. Russell, editors, Proceedings of the 12th Inter-national Conference on Machine Learning, pages 115–123, T ahoe City, CA, jul 1995. Morgan Kaufmann.

[7] W. L. E. Eskin and S. Stolfo. Modeling system call for intru-sion detection using dynamic window sizes. In Proceedings of DARP A Information Survivability Conference and Expo-sition, 2001.

[8] J. Forney , G.D. The viterbi algorithm. Proceedings of the IEEE, 61(3):268–278, March 1973.

[9] S. Forrest, S. A. Hofmeyr, A. Somayaji, and T. A. Longstaff. A sense of self for unix processes. In Proceedinges of the 1996 IEEE Symposium on Research in Security and Privacy, pages 120–128. IEEE Computer Society Press, 1996.

[10] S. Forrest, C. Warrender, and B. Pearlmutter. Detecting in-trusions using system calls: Alternate data models. In Pro-ceedings of the 1999 IEEE Symposium on Security and Pri-vacy, ages 133–145, Washington, DC, USA, 1999. IEEE Computer Society.

About the Auhtors:

HarishBabu.Kalidasu working as Asst.Professor, Dept. of CSE, St.Mary’s Group of Institutions, Guntur. He is completed his B.Tech (IT) at Andhra University, M.Tech (CSE) at JNTU-K. He has 4+ years of experience in teaching field, 2+ years of experience in research work. He is the author & co-author of 12 papers, his research works includes data mining, Networks.

Ch.Venkateswarao ,working as Asst.Professor, Dept of CSE, HMKS & MGS College of Engineering, Kangala, Guntur. He completed M.Tech (CSE) at JNTU-K, He has 4 years of experience in Teaching field and 1 year of experience in Research work. His areas interests image processing, data mining.