International Journal of Communication Networks and Security

Volume 2 | Issue 1 Article 9

January 2013

INTRUSION DETECTION FOR DISCRETE SEQUENCES

MRS. M. VIJAYALAKSHMI
Department, Intell Engineering College, Affiliated to JNTU ANANTAPUR, India.,
VIJAYALAKSHMI@gmail.com

MR. K . JANARDHAN
Department, Intell Engineering College, Affiliated to JNTU ANANTAPUR, India., JANARDHAN@gmail.com

INTRUSION DETECTION FOR DISCRETE SEQUENCES

MRS. M. VIJAYALAKSHMI[1], MR. K .JANARDHAN[2]

[1]M.Tech, [2]M.Tech. Asst. Professor Computer Science & Engineering Department, Intell Engineering College, Affiliated to JNTU ANANTAPUR, India.

Abstract- Global understanding of the sequence anomaly detection problem and how techniques proposed for different domains relate to each other. Our specific contributions are as follows: We identify three distinct formulations of the anomaly detection problem, and review techniques from many disparate and disconnected domains that address each of these formulations. Within each problem formulation, we group techniques into categories based on the nature of the underlying algorithm. For each category, we provide a basic anomaly detection technique, and show how the existing techniques are variants of the basic technique. This approach shows how different techniques within a category are related or different from each other. Our categorization reveals new variants and combinations that have not been investigated before for anomaly detection. We also provide a discussion of relative strengths and weaknesses of different techniques. We show how techniques developed for one problem formulation can be adapted to solve a different formulation; thereby providing several novel adaptations to solve the different problem formulations. We highlight the applicability of the techniques that handle discrete sequences to other related areas such as online anomaly detection and time series anomaly detection.

1. INTRODUCTION

Sequence data is found in a wide variety of application domains such as intrusion detection, bio-informatics, weather prediction, system health management, etc. Hence anomaly detection for sequence data is an important topic of research. There is extensive work on anomaly detection techniques that look for individual objects that are different from normal objects. These techniques do not take the sequence structure of the data into consideration. For example, consider the set of user command sequence. While sequences S1–S4 represent normal daily profiles of a user, the sequence S5 is possibly an attempt to break into a computer by trying different passwords. Though the sequence S5 is anomalous, each command in the sequence by itself is normal. A sequence is an ordered series of events. Sequences can be of different types, such as binary, discrete, and continuous, depending on the type of events that form the sequences. Discrete and continuous sequences (or time series) are two most important form of sequences encountered in real life. In this survey we will focus on discrete sequences. Discrete or symbolic sequences are ordered sets of events such that the events are symbols

S1 login, passwd,mail, ssh, . . . ,mail,web, logout
S2 login, passwd,mail,web, . . . ,web,web,web, logout
S3 login, passwd,mail, ssh, . . . ,mail,web,web, logout
S4 login, passwd,web,mail, ssh, . . . ,web,mail, logout
S5 login, passwd, login, passwd, login, passwd, . . . , logout

Sequences of User Commands belonging to a finite unordered alphabet. For example, a text document is a sequence of words, a computer program is executed as a sequence of system calls, a gene is a sequence of nucleic acids. Often, one is interested in detecting anomalies in discrete sequences to find possible intrusions, frauds, faults, contaminations. For example, the sequences of commands issued by computer users (as shown in Table) are collected to detect possible intrusive activities. Anomaly detection for discrete sequences is a challenging task, since it involves exploiting the sequential nature of data to detect anomalies Some of the specific challenges are as follows: • Multiple definitions of anomalies exist for sequences; an event within a sequence maybe anomalous, a subsequence within a sequence maybe anomalous, or the entire sequence maybe anomalous. Each definition needs to be handled differently. For example, given a data base of protein sequences, one might be interested in sequences that are anomalous. On the other hand, given a long sequence of system calls executed in a computer system, one might be interested in detecting subsequences when the computer system was under a virus attack. A technique that detects anomalous events within a sequence might not be directly applicable to detecting anomalies that are caused by a subsequence of events occurring together.

The length of the anomaly within a sequence is usually not known and varies significantly across different application domains. Techniques typically have to rely on user-defined lengths, which may or may not be optimal• Computational complexity is a significant issue for sequence data, especially since sequences can be very long and the alphabet size can be large.

II. IMPLEMENTATION

To detect anomalies in discrete sequences, a deeper analysis reveals that different techniques actually address different problem formulations and An anomalous sequence from a database of unlabeled

sequences and Detect an anomalous subsequence within a long sequence of events To provide a comprehensive and structured overview of the existing research in this context We provide a basic anomaly detection technique, and show how the existing techniques are variants of the basic technique This approach shows how different techniques within a category are related or different from each other.

1. Discrete Sequences

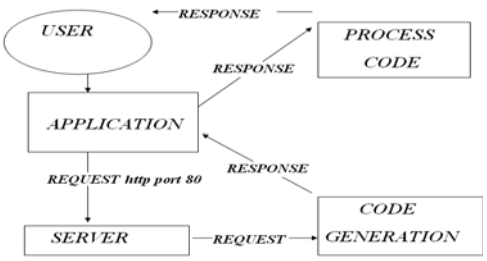
Discrete or symbolic sequences are ordered sets of events such that the events are symbols. Sequences of transactions from online banking, customer purchases, and other retail commerce domains. The "symbols" in such sequences are "actions" by customers and the alphabet size for such sequences can also be large. Anomalies in such data sets correspond to irregular or abnormal behavior by customers

2. Anomaly Detection

Anomaly detection for discrete sequences is a challenging task, since it involves exploiting the sequential nature of data to detect anomalies. A basic anomaly detection technique can be described as follows: First, all k-length windows are extracted from the given sequence T and stored as a database of fixed length windows, denoted as Tk. Each window is assigned an anomaly score by comparing it with rest of the windows in Tk. The windows with anomaly scores above a user defined threshold are chosen as the top discords. Since the length of the "true" discord is not known a priori, the techniques that solve this problem generally assume that the discords are contained within a subsequence (or a window) of fixed length k.

3. Window Scoring Techniques

Another possible technique for scoring the windows would be that the anomaly score of any window is calculated as equal to its distance to its mth nearest neighbor in Tk. One such variant, called HOTSAX , The anomaly score of a window is equal to its distance to its nearest neighbor in Tk (only the non-self matches are considered). One drawback of the nearest neighbor based technique is that they involve an additional parameter, m, which needs to be set carefully, though approaches such as using the weighted sum of distance to the m nearest neighbors to compute the anomaly score can be used to reduce the sensitivity on m.



The basic technique and its variations have following two key issues:

1. Computational Complexity

For each query pattern, the time required to compute its anomaly score is linear in the length of S and the length and number of sequences in S. If there are multiple query patterns, e.g., all short subsequences that occur in S, the total time required to score all query patterns adds up to a high value. To address this issue, Keogh etal proposed a technique called TARZAN which uses suffix trees to efficiently compute the frequency of occurrence of a query pattern in a given sequence. Suffix trees are created for S and for each sequence in S1. Only two suffix trees are required, one for the sequences in S and for the sequence S, and can be constructed with complexity linear in the length of the sequences. The counts for a query pattern s, can be obtained with complexity linear in the length of s.

2. Interpolated Markov Models (IMM) To efficiently find the number of windows extracted from a sequence that contain the query pattern or its permutations, as a subsequence. the basic technique that solves problem formulation 3 compares this value with the normalized frequency of occurrence of the window in the test sequence to compute its anomaly score Nevertheless, techniques for problem formulation 3 can also be used to find certain types of anomalies in the context of problem formulation 1. An alternate cause of anomaly could be that a test sequence is anomalous because it contains one or more patterns whose frequency of occurrence in the test sequence is significantly different from their frequency of occurrence in the training sequences. Such anomalous sequences can be detected as follows: For the given test sequence, fixed length windows are extracted. Each window is assigned an anomaly score using the basic technique that solves problem formulation 3. The anomaly scores of all windows are aggregated to obtain an overall anomaly score for the test sequence.

ONLINE ANOMALY DETECTION

Several application domains collect sequence data in a streaming fashion , e.g., system call data collected by a computer system, data generated by an aircraft during its flight, etc. Such domains, often require the anomalies to be detected in such sequences in an online fashion, i.e., as soon as they occur. Online anomaly detection has the advantage that it can allow analysts to undertake preventive or corrective measures as soon as the anomaly is manifested in the sequence data. For example, in aircraft health monitoring, the current flight sequence for an aircraft is tested if it is anomalous or not, with respect to a database of historical normal flight sequences of the aircraft. Determining that the current flight sequence

has an anomaly, as soon as it occurs (even before the entire flight sequence is collected) might help the health monitoring system to raise an early alarm.

CONCLUSION

The survey covers techniques developed independently for different domains and thus opens up the possibility of applying techniques developed within one domain to a completely different setting.
In this survey we have discussed three different problem formulations that are relevant in varied application domains. For each problem formulation we have identified distinct groups of techniques that use a specific approach to detect anomalies. Within each of these groups we have provided a basic technique and shown how different existing techniques are variations of the corresponding basic technique. We also provide the motivation behind the different variations and the strengths and weaknesses associated with the different categories of techniques. This results in a better understanding of the current state of research as well as allows future researchers to develop novel variations

FUTURE SCOPE

Kernel Based and Window based technique for problem formulation1 as well as basic technique for problem formulation2 can easily extend to multivariate sequence as long as a similarity measures can be developed to compare two multivariate discrete sequence. such techniques have not been tried yet but need to be investigate in future.

REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection – a survey,” ACM Computing Surveys (To Appear), 2008.

[2] V. Hodge and J. Austin, “A survey of outlier detection methodologies,” Artificial Intelligence Review, vol. 22, no. 2, pp. 85–126, 2004.

[3] A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur, and J. Srivastava, “A comparative study of anomaly detection schemes in network intrusion detection,” in Proceedings of SIAM International Conference on Data Mining. SIAM, May 2003.

[4] S. Forrest, C.Warrender, and B. Pearlmutter, “Detecting intrusions using system calls: Alternate data models,” in Proceedings of the 1999 IEEE ISRSP. Washington, DC, USA: IEEE Computer Society, 1999, pp. 133–145.

[5] S. A. Hofmeyr, S. Forrest, and A. Somayaji, “Intrusion detection using sequences of system calls,” Journal of Computer Security, vol. 6, no. 3, pp. 151–180, 1998. [Online]. Available: citeseer.ist.psu.edu/ hofmeyr98intrusion.html

[6] C. C. Michael and A. Ghosh, “Two state-based approaches to program-based anomaly detection,” in Proceedings of the 16th Annual Computer Security Applications Conference. IEEE Computer Society, 2000, p. 21.