Mining DoS attack sequences on Network Traffic using Fuzzy Time Interval

Alpa Reshamwala
Assistant Professor, Computer Engineering Department
MPSTME, SVKM's NMIMS University
Mumbai, India

Sunita Mahajan, PhD
Principal, Institute of Computer Science
M.E.T, Bandra
Mumbai, India

ABSTRACT

Intrusion of network which couldn't be analyzed, detected and prevented may make whole network system paralyze while the abnormal detection can prevent it by detecting the known and unknown character of data. Many intrusions aren't composed by single events, but by a series of attack steps in chronological order. Analyzing the order in which events occur can improve the attack detection accuracy and reduce false alarms. Intrusion is a multi step process in which a number of events must occur sequentially in order to launch a successful attack. Although conventional sequential patterns can reveal the order of attack events, the time between events can also be determined but it causes the sharp boundary problem. That is, when a time interval is near the boundary of two predetermined time ranges, one either ignore or overemphasize it. Therefore, this paper uses the concept of fuzzy sets so that Dos attack sequential patterns are discovered on network traffic in fuzzy time interval. In this paper, an apriori based candidate generation algorithm has been implemented with Fuzzy time intervals to detect Dos attack sequences. The experimental results are also compared with the dataset which is generated by the SPMF sequential dataset generator.

Keywords

Data mining; Fuzzy Set; Sequential pattern; Time interval; Intrusion detection system ; DoS attacks; Apriori.

1. INTRODUCTION

With the rapid development of electronics and information technology, computer network plays a significant role in our daily lives. However, more and more people use it as a tool for crime, and it has brought great losses to many companies and countries. An intrusion into a computer system is any activity that violates system integrity, confidentiality, or data accessibility. In order to meet this challenge, Intrusion Detection System is being designed to protect the availability, confidentiality and integrity of critical networked information systems [1][2][3]. Many intrusions aren't composed by single events, but a series of attack steps in chronological order. Any single message or command of these steps maybe normal and doesn't have obvious signatures of string attacks. But a series of messages or commands in chronological order can constitute an attack. Analyzing the order in which events occur can improve the attack detection accuracy and reduce false alarms. This is because, very often, intrusion is a multi step process in which a number of events must occur sequentially in order to launch a successful attack. Therefore, sequential pattern mining algorithms are applied to intrusion detection to mine the order correlation about time sequential data, and then it can detect this kind of attack.

Sequential pattern mining, one of the most popular data mining methods, receives many attentions in recent years. Sequential pattern mining is one of the most well-known methods and has broad applications including web-log analysis, customer purchase behavior analysis and medical record analysis. For example, the sequential pattern mining can generate a sequential pattern such as on buying an inkjet printer a customer will look for ink cartridge followed by printing□ papers and later for an□ □ ink cartridge refill kit. This pattern shows that customers who buy an inkjet printer will have strong probability to buy an ink cartridge, printing papers, and ink cartridge refill kit in order. With the help of this pattern, the retailer can send an ink cartridge refill kit promotion program to customers after customers bought an inkjet printer, an ink cartridge, and printing papers. However, from these discovered sequential patterns, the time gaps between successive patterns cannot be determined. For example, "inkjet printer□ □ (*one week*) □ □ ink cartridge refill kit" and "inkjet printer □ □ (*one year*) □ □ ink cartridge refill kit" are two sequential patterns with the same order but different time-intervals. It is clear that customers buy the ink cartridge refill kit after they bought inkjet printer one week later and one year should not be viewed as the same behavior. If time-interval information is not considered in the two sequential patterns, managers might consider the two patterns as the same pattern of "inkjet printer: □ □ ink cartridge refill kit".

In the research [4], Chen *et al.* have proposed a generalization of sequential patterns, called time-interval sequential patterns, which reveals not only the order of patterns, but also the time intervals between successive patterns but it causes the sharp boundary problem. Therefore, this paper uses the concept of fuzzy sets to solve the sharp boundary problem. Two efficient algorithms based on fuzzy theory, FTI-Apriori algorithm and the FTI-PrefixSpan algorithm [5]. In [6], the author has contributed to the ongoing research on FTI sequential pattern mining by proposing an algorithm to detect and classify audit sequential patterns in network traffic data. The paper also defines the confidence of the FTI audit sequences, which is not yet defined in the previous researches. In [7], the authors have proposed an algorithm which uses a fuzzy genetic approach to discover optimized sequences in the network traffic data to classify and detect intrusion.

In this paper, the concept of fuzzy set theory is used so that fuzzy time-interval Dos attack sequential patterns are discovered on network traffic. In this paper, an Apriori a candidate generation algorithm based Fuzzy time interval algorithm to detect Dos attack sequences on network traffic data of KDD Cup 1999, 10 percent of training dataset, which is the annual Data Mining and Knowledge Discovery competition organized by ACM Special Interest Group on

Knowledge Discovery and Data Mining, the leading professional organization of data miners.

2. RELATED WORK

The problem of mining sequential patterns was first introduced by Agarwal and Srikant [8] which discovers patterns that occur frequently in a sequence database. A sequence database is formed by a set of data sequences. Each data sequence includes a series of transactions, ordered by transaction times.

After mid 1990's, following Agrawal and Srikant [8], many scholar provided more efficient algorithms [9][10][11][12]. Besides these, works have been done to extend the mining of sequential patterns to other time-related patterns. Existing approaches to find appropriate sequential patterns in time related data are mainly classified into two approaches. In the first approach developed by Agarwal and Srikant [13], the algorithm extends the well-known Apriori algorithm. This type of algorithms is based on the characteristic of Apriori—that any subpattern of a frequent pattern is also frequent [8]. The later, uses a pattern growth approach [9], employs the same idea used by the Prefix-Span algorithm. In the algorithm [11], Shrikant and Agrawal, specified the maximum interval (max-interval), the minimum interval (min-interval) and the sliding time window size (windowsize). Moreover, they cannot find a pattern whose interval between any two sequences is not in the range of the window-size. Agrawal and Srikant [8], introduced mining traditional sequential mining, by ignoring the time interval and including only the temporal order of the patterns.

To address the intervals between successive patterns in sequence database, Chen *et al.* have proposed a generalization of sequential patterns, called time-interval sequential patterns, which reveals not only the order of patterns, but also the time intervals between successive patterns [4]. Chen *et al.* developed algorithms to find sequential patterns using both the approaches [4]. An extension of the algorithm developed by Chen *et al.* [5], to solve the problem of sharp boundaries to provide a smooth transition between members and non-members of a set. The sharp boundary problems can be solved by the concept of fuzzy sets. Two efficient algorithms, the FTI-Apriori algorithm and the FTI-PrefixSpan algorithm, were developed for mining Fuzzy time interval(FTI) sequential patterns. The results shown by Yangdong Ye *et al.* in [14], shows that possibility distribution of train's arrival time can be statistically gained, and then can compute fuzzy time interval based on the theorem of probability-possibility transformations. In [15], Chung-I Chang *et al.*, proposed an algorithm called sequential pattern mining with fuzzy time intervals (SPFTI). Finally, the experimental result verifies that result of the proposed SPFTI algorithm outperforms with the fuzzy sequential patterns mining with fixed time interval. In paper [16], Chung-I Chang *et al.* proposed an algorithm called integrated sequential pattern mining with fuzzy time intervals (ISPFTI). The main idea of ISPFTI algorithm is to use the a priori-like method to mine the frequent sequential patterns of sequence database and use fuzzy theory to mine the time interval between frequent sequences. There are several other reasons that support the use of FTI in place of crisp interval. First, the human knowledge can be easily represented by fuzzy logic. Second, it is widely recognized that many real world situations are intrinsically fuzzy, and the partition of time interval is one of them. Third, FTI is simple and easy for users.

Currently many researchers apply data mining algorithms to intrusion detection, concentrating on association rules, classification, clustering and sequence pattern mining, etc. The research of sequential pattern mining algorithms based intrusion detection is mainly about anomaly detection for call sequences of host-based system and user command sequences. In addition, one can improve the performance of sequential pattern mining algorithms based intrusion detection by adding domain knowledge to improve the efficiency of mining, as well as adding fuzzy logic using constraints to remove unwanted rules and join the concept of time, etc.

Stephanie Forrest research group firstly introduced an intrusion detection method using system call sequences [17], which mainly describes normal operation model of the process by a short fixed-length sequential pattern. But it only considered system call in the order of time, without considering the parameters of calling. Its improved algorithms include sequence analysis of slide window, sequence prediction and HMM (Hidden Markov Models) correlated analysis [18], etc. Wenke Lee *et al.* utilize conditional entropy model to get the optimal length of short sequences [19], however, the conditional entropy model used in the whole dataset isn't reasonable because the process of system call sequence collection passes over a long period of time. Karlton Sequeira and Mohammed Zkai [20] made temporal clustering for the user' s operation command sequences, and generated the method to get users' normal behavior: ADMTI(Anomaly-based Data Mining for Intrusions). Once it finds the user's operation, sequence is incompatible with the normal mode; the sequence is treated as imitating user alarm. ADMTI doesn't depend on the sign of training data, but it still has higher false alarm rate.

Anrong *et al.* [21], addresses application of sequential pattern in intrusion detection by refining the pattern rules and reducing redundant rules. Their work implements PrefixSpan algorithm in the data mining module of network intrusion detection system (NIDS). Fuzzy logic addresses the formal principles of approximate reasoning. It provides a sound foundation to handle imprecision and vagueness as well as mature inference mechanisms by varying degrees of truth. As boundaries are not always clearly defined, fuzzy logic can be used to identify complex pattern or behavior variations. And it can be accomplished by building an intrusion detection system that combines fuzzy logic rules with an expert system in charge of evaluating rule truthfulness.

Shengbin Zhang *et al.* [22], proposed an improved PrefixSpan algorithm to fit into intrusion detection system. Though this strategy avoids the construction of projected databases, it needs to scan the original database many times. SongShi-Jie *et al* [23], introduced an efficient sequential pattern mining algorithm. It uses item transaction vertical data structure to reduce the times of scanning the original database effectively; reuses large itemsets to increase the mining granularity; uses bitmap mode to solve the problem of counting the candidate sequential patterns; and prunes the candidate sequences to improve the speed of sequential pattern mining. Weihong Cai *et al.* [24], proposed a new mode of intrusion detection by using the fuzzy logic with data mining and immune genetic algorithm. It creates respectively the rule collection of natural behavior mode and inspecting behavior mode. Sequential pattern mining based intrusion detection is also applied in the field of macro-area analysis. Duan Yi-feng *et al.* [25], studied the applications of alarm sequential pattern mining, and sequential pattern mining algorithms provided an effective way to find the knowledge of network alarms. According to

the characteristics of network audit data, Hong-liang Xin *et al.* [26], presented a fast sequential pattern mining algorithm. Time and attribute-relative features were utilized to lead the mining process, and strict attribute schemes were used to prune sequential rules. Weijun Zhu *et al.*, in [27], applied Interval temporal logic (ITL) model which reduces the false negative rate of misuse detection for concurrent attacks. Milanesi, G. *et al.* [28], proposed a novel system for indoor video surveillance. It is able to detect and track moving objects even in the presence of significant variations of scene illumination. The system has been tested on a wide variety of situations, proving its effectiveness and robustness. Kai Xing Wu *et al.*, in [29] discusses about Intrusion detection system (IDS) is based on data mining technology. In this paper, a novel framework based on data mining techniques was proposed for designing IDS. In this framework, the classification engine, which is actually the core of the IDS, uses fuzzy association rules for building classifiers. Yongzhong Li *et al* [30], discusses about the excellent performance of the HMM (hidden Markov model), its use in pattern recognition. Due to the high false alarm rate in the classical intrusion detection system (IDS) based on HMM, a fuzzy approach for the HMM, called fuzzy hidden Markov models (FHMM) was proposed. Jianxiong Luo *et al.*, described an extension in [31], that uses fuzzy frequent episodes for near real-time intrusion detection. They first define fuzzy frequent episodes and then describe experiments that explore their applicability for real-time intrusion detection. Ming-Yang Su *et al.* [32], proposed a fast algorithm to generate fuzzy association rules by incremental mining approach, for which the transactions or data records are online instantly collected from live packets. That is, as one data record is collected online, the latest fuzzy rules can be obtained immediately.

Xiaogang Wang *et al.* [33], proposed watermarking schemes for tracing network attack flows to detect stepping-stone intrusion and fight against the abuse of anonymity. They have proposed an efficient sequential watermark detection (ESWD) model for tracing network attack flows. In [6], the author has contributed to the ongoing research on FTI sequential pattern mining by proposing an algorithm to detect and classify audit sequential patterns in network traffic data. The paper also defines the confidence of the FTI audit sequences, which is not yet defined in the previous researches. In [7], the authors have proposed an algorithm which uses a fuzzy genetic approach to discover optimized sequences in the network traffic data to classify and detect intrusion. In [34], the authors have implemented Apriori a candidate generation algorithm and PrefixSpan a pattern growth algorithm on a network intrusion dataset from KDD Cup 1999, 10 percent of training dataset, which is the annual Data Mining and Knowledge Discovery competition organized by ACM Special Interest Group on Knowledge Discovery and Data Mining, the leading professional organization of data miners. To address the absence of timestamp in the dataset, two approaches were considered to generate the sequence database from the dataset. One is by taking service as reference attribute and the other one by taking a timestamp window of size one day (86400 seconds). From the experimental results it can be seen that PrefixSpan for predicting DoS attacks sequences on KDD cup 99 training dataset are efficient. These results are then compared with SPAM (Sequential Pattern Mining) algorithm which uses vertical bitmap data layout allowing for simple, efficient counting. In [35], I-Apriori a candidate generation algorithm and I- PrefixSpan a pattern growth algorithm have been implemented to detect time interval denial of service (DoS) attack sequences on network traffic data of KDD Cup 1999, 10 percent of training dataset. The comparison study is done on the number of patterns and on the average length of patterns obtained by varying the time interval of the sequential patterns.

3. FUZZY LOGIC

Fuzzy logic has been a powerful tool for decision making to handle imprecise and uncertain data. In contrast to classical set, a fuzzy set is a set without crisp boundaries; the transition from "belong to a set" to "not belong to a set" is gradual. Membership function is utilized to reflect a degree of membership and indicated by a value in the range [0.0, 1.0].

A Fuzzy set can be defined as - If *U* is a collection of objects denoted generically by *x*, then a *fuzzy set A* in *U* is defined as a set of ordered pairs:

$A = \{(x, \mu_A(x)) | x \in U\}$,

where, $\mu_A(x)$ is the membership function and *U* is the universe of discourse.

Just like an algebraic variable takes numbers as values, a Linguistic/ Fuzzy Variables *takes words* or sentences as values. The set of values that it can take is called its "*term set"*. Each value in the term set is a *"Fuzzy variable"* defined over a *"Base variable"*. The base variable defines the universe of discourse for all the fuzzy variables in the term set. The fuzzy variables themselves are adjectives that modify the variable (e.g. "large positive" error, "small positive" error, "zero" error, "small negative" error, and "large negative" error). As a minimum, one could simply have "positive", "zero", and "negative" variables for each of the parameters. Additional ranges such as "very large" and "very small" could also be added to extend the responsiveness to exceptional. For example, Fuzzy Linguistic Variables are used to represent qualities spanning a particular spectrum: Temp: {Freezing, Cool, Warm, Hot}, as shown in Figure 1.

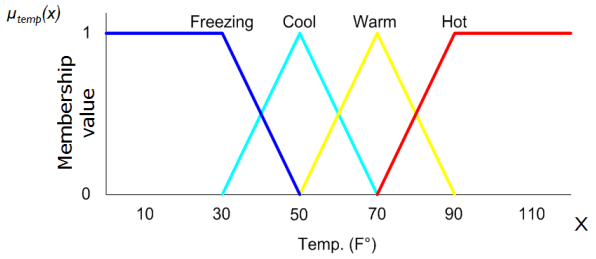

Fig 1: Membership function of Temperature.

4. FUZZY TIME INTERVAL

Two approaches have been used to determine linguistic terms and fuzzy membership functions. The first approach relies on domain experts to specify the functions based on their background knowledge and requirements. The second approach assumes that the functions are obtained by a preprocessing phase that learns the functions from the data, such as learning by neural-network, by GA, by clustering method, and by entropy measure. Since, in the current fuzzy mining researches, the first approach is more popular than the second one.

For example, when taking time interval of six hours in a 24 hour day one can have four fuzzy time-interval FTI - 4 by using four linguistic terms: *Instant* (*I*), *Immediate* (*Im*), *awhile* (*A*) and *Later* (*L*) within a day. Consider a day having 24

hours, *Immediate* linguistic variable can be defined as, if the time interval between two events is less than or equal to 6 hours then they are definitely in sequence and has the membership value as 1 and if time interval is within 12 hours, between 6 and 12 hours, more than 12 hours can be represented by equations (1), (2), (3) and (4) respectively.

The graphical representation is shown in figure 2. Similarly for FTI-8 one can have LT1, LT2, LT3, LT4, LT5, LT6, LT7 and LT8 and their linguistic variable can be *Instant*, *Imediately*, S*ooner*, *veryShort*, *Shortly*, *Considerable*, *Lately*, *tooLate*. Accordingly, FTI-12 and FTI-24 can be defined in the same manner.

$$\mu_{Instant}(ti_j) = \begin{cases} 1 & ; tij = 0 \\ \frac{6-tij}{6}; & 0 < tij \leq 6 \\ 0 & ; tij \geq 6 \end{cases} \quad \textbf{(1)}$$

$$\mu_{Immediate}(ti_j) = \begin{cases} \frac{tij}{6} & ; 0 < tij < 6 \\ \frac{12-tij}{6}; & 6 < tij \leq 12 \\ 0 & ; tij \geq 12 \\ 1 & ; tij = 6 \end{cases} \quad \textbf{(2)}$$

$$\mu_{Awhile}(ti_j) = \begin{cases} \frac{tij-6}{6} & ; 6 < tij < 12 \\ \frac{18-tij}{6}; & 12 < tij \leq 18 \\ 0 & ; tij \leq 6 \text{ or } tij \geq 18 \\ 1 & ; tij = 12 \end{cases} \quad \textbf{(3)}$$

$$\mu_{Later}(ti_j) = \begin{cases} \frac{tij-12}{6} & ; 12 < tij < 18 \\ \frac{24-tij}{6}; & 18 < tij \leq 24 \\ 0 & ; tij \leq 12 \text{ or } tij \geq 24 \\ 1 & ; tij = 18 \end{cases} \quad \textbf{(4)}$$

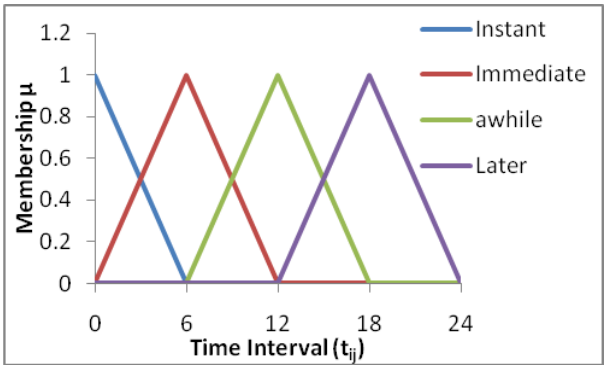

Fig 2: Membership function for time interval

5. SEQUENCE DATABASE GENERATION

When mining the intrusion data using sequential pattern mining algorithms, it needs to preprocess the source data, and then receives training dataset by analyzing the features of data used in intrusion detection. If one has a good algorithm, and not high-quality training data, the detective result will be not good. Different from other application fields, intrusion detection usually uses some artificial intelligence methods, which analyze data by choosing a model. However choosing models always depends on instinct and expert knowledge, and there isn't an objective method to evaluate the data.

Data acquired from the dataset may be not sequential. A sequence is an ordered list of items [8]. Now, consider KDD cup 99 training dataset which is approximately 4,900,000 single connection vectors, each of which contains 41 features and is labelled as either normal or an attack, with exactly one specific attack type. For example, a network connection can be uniquely identified by the combination of its *timestamp* (start time), *src host* (source host), *src port* (source port), *dst host* (destination host), and *service* (destination port). These are the essential attributes when describing network data [36]. To address the absence of timestamp in the dataset, for generating a sequence database one can consider the sequence by taking a timestamp window of size 1 day (86400 seconds). For experimental purpose, DoS attacks have been considered, which are approximately 392000 single connection vectors each labelled as either normal or a DoS attack.

To analyze sequential dataset statistics as well to generate synthetic dataset with the same parameters as DoS attack sequence dataset of KDD cup 99 training data, SPMF (Sequential Pattern Mining Framework) is used which is implemented by Phillipe Fournier-Viguera [37] and available from http://www.philippe-fournier-viger.com/spmf/.

The results for DoS attack sequence dataset of KDD cup 99 training dataset is as shown in Table 1.

Table 1. KDD Cup 99 Dos attack Parameters

	Parameters	**KDD Cup 99**
S	Number of sequences	6
N	Number of distinct items	6
I_S	Average number of itemsets per sequence	15.83
I	Average number of items per itemset	1
T	Average time interval length	11

6. EXPERIMENTAL RESULTS

Common sequential pattern mining algorithms for intrusion detection generally can be divided into two types: one is level-wise and based on Apriori theory, such as AprioriAll, GSP and SPADE, etc. The other is the sequential pattern growth algorithm based on divide and conquers, such as FreeSpan, PrefixSpan, etc. Using the feature of Apriori, Apriori-based algorithms largely reduce searching space, but there are still inherent and inner shortcomings: generating a large number of

candidate sequences, scanning the whole database and it is difficult to mine long sequential patterns. So it has lower space-time efficiency. The main consumption of PrefixSpan based on sequential pattern growth is the construction of projected databases. And its performance of space is better than performance of time. There is room left to improve the time performance. For example, before generating new seed sequences, seed sequences which don't get longer in next process are removed in advance, so that it can reduce time consumption by loop. The time efficiency existing in these two types of algorithms affects the real-time performance of intrusion detection.

In this section a simulation study is performed to compare the performances of the traditional algorithms without time interval: Apriori-based algorithm; Apriori and pattern growth algorithm; PrefixSpan. Also, with time interval: I-Apriori and I- PrefixSpan algorithms [4], and with fuzzy time interval algorithms FTI- Apriori which find sequential patterns with fuzzy time intervals as well as fixed time interval. These algorithms were implemented in Sun Java language and tested on an Intel Core Duo Processor, 2.10 GHz with 2GB main memory under Windows XP operating system.

The comparison study is done on the number of patterns and the average length per sequential patterns by varying the intervals in the fuzzy time-interval sequential pattern. Here, intervals are partitioned in four different ways. Now, consider KDD cup 99 training dataset which is approximately 4,900,000 single connection vectors, each of which contains 41 features and is labeled as either normal or an attack, with exactly one specific attack type. To address the absence of timestamp in the dataset, for generating a sequence database the sequence is considered by taking a timestamp window of size 1 day (86400 seconds). For experimental purpose, DoS attacks are considered, which are approximately 392000 single connection vectors each labeled as either normal or a DoS attack. Now, by taking time intervals of 1 Day, one can get values of one, two, three and six hour partitions of time with equal depth. For example, when taking time interval of six hours in a 24 hour day one can have four fuzzy time-interval sets as FTI - 4 {LT0; LT1; LT2; LT3}, where; the literals can be represented by four linguistic terms: *Instant* (*I*), *Immediate* (*Im*), *awhile* (*A*) and *Later* (*L*). The fuzzy membership function can be as given by equations (1), (2), (3) and (4) respectively. During each test, all parameters are fixed and the minimum support threshold is varied from 20 to 90% as shown in the figures 3 and 4, to determine how the number of patterns varies with the minimum support for both the dataset - KDD cup 99 training dataset and the S6-N6-I_S16-I1-T8 dataset which is generated by the SPMF sequential dataset generator. The parameters taken in S6-N6-I_S16-I1-T8 dataset generation is same as shown in Table I, with 6 as number of sequences, 6 number of distinct items, 16 average number of itemsets per sequence, 1 as the average items per itemset and 8 as the average time interval length.

As shown in Figure 3 and 4, shows the experimental observations on DoS attack sequence dataset of KDD cup 99 training dataset. From the figure 3, it is observed that the traditional sequence pattern mining Apriori based algorithm generates maximum patterns followed by fuzzy time interval algorithm FTI-Apriori followed by PrefixSpan algorithm; and finally the two time interval sequential pattern mining algorithms, I- Apriori and I-PrefixSpan. Observations in Figure 4 indicate that, as the intervals become smaller, the situation becomes worse and the number of patterns decline.

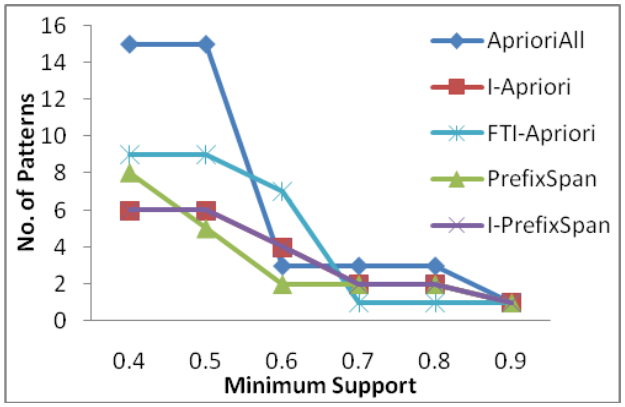

Fig 3: KDDCup 99 DoS attack-No of Patterns

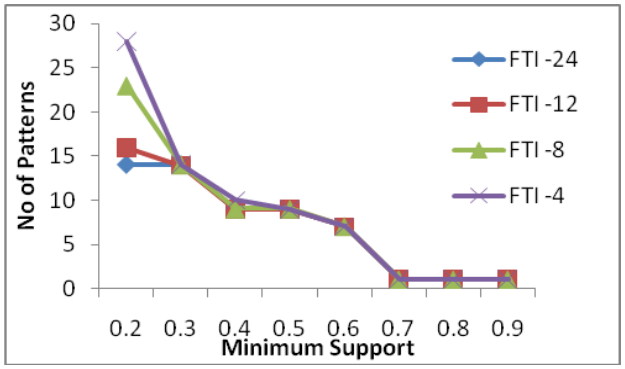

Fig 4: KDDCup99 Dos attack- FTI-Apriori - No. of patterns verses pattern type for varying FTI

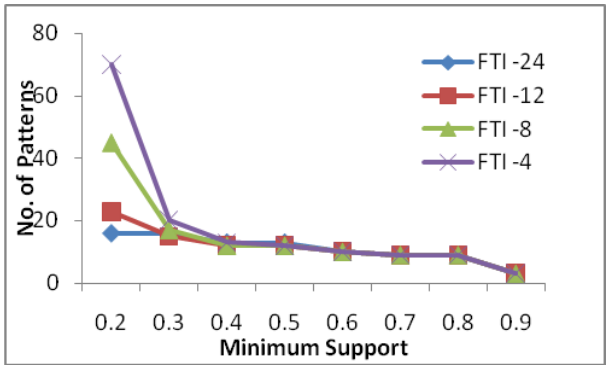

Fig 5: S6-N6-I_S16-I1-T8 - FTI-Apriori - No. of patterns verses pattern type for varying FTI

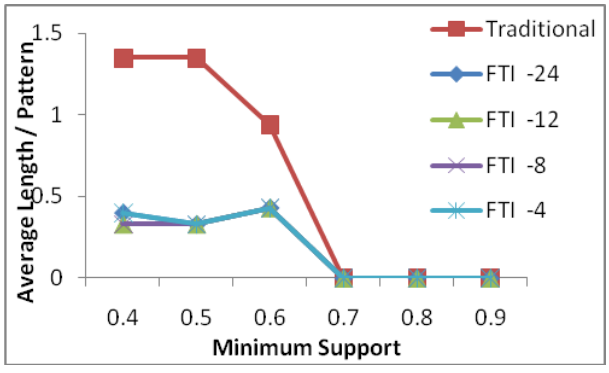

Fig 6: KDDCup99 Dos attack -Average length of patterns

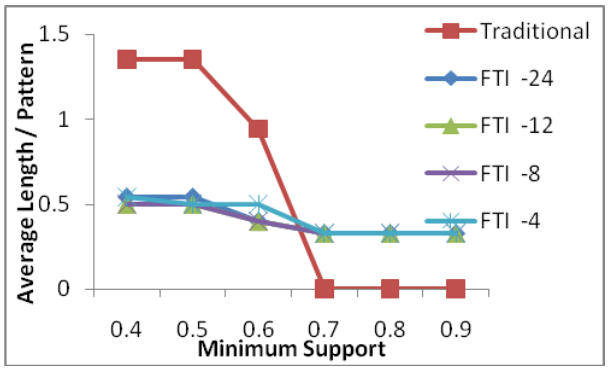

Fig 7: S6-N6-I_S16-I1-T8 -Average length of patterns

In figure 5, it is observed that with the same parameters dataset S6-N6-I_S16-I1-T8, generates maximum patterns compared to KDD Cup 99 dataset for varying FTI. Figure 4 and 5 shows that as the time line is partitioned into fewer intervals, candidates tend to have larger support.

From, figure 6 and 7 it can be observed that though by keeping all the parameters fixed the average length per pattern are fluctuating within the negligible range in dataset S6-N6-I_S16-I1-T8 as compared to KDDCup99 Dos attack dataset.

Observations shows that the average length of the traditional pattern is the longest, followed by the average length of FTI-4 fuzzy time interval patterns, then the FTI-8 fuzzy time interval patterns; and finally the FTI-12 and FTI-24. Hence it can be observed that as the time line is partitioned into fewer intervals, more of the candidate patterns become frequent and as a result the average length of patterns increases. Thus fewer intervals are prone to have a larger number of patterns as well as longer average length.

Also, as the fuzzy time interval patterns are discovered the support of the candidate patterns are spread over several fuzzy intervals, rather than only one interval as the crisp interval. This generates fewer patterns as the support increases and decreases the average length of the patterns.

7. CONCLUSIONS AND FUTURE WORK

Although conventional sequential patterns can reveal the order of attack events, the time between events can also be determined but it causes the sharp boundary problem. That is, when a time interval is near the boundary of two predetermined time ranges, one either ignore or overemphasize it. Therefore, this paper uses the concept of fuzzy sets so that fuzzy time-interval Dos attack sequential patterns are discovered on network traffic.

In this paper I- Apriori a candidate generation algorithm and I- PrefixSpan [4] is implemented, which finds sequential patterns with time intervals which reveal not only the order of patterns, but also the time intervals between successive patterns. Also Apriori a candidate generation algorithm based Fuzzy time interval algorithm FTI- Apriori is implemented on two dataset one, KDD Cup 1999 dataset to detect DoS attack sequences on network traffic data and the S6-N6-I_S16-I1-T8 dataset which is generated by the SPMF sequential dataset generator. Experimental results have shown that, as the time interval set taken for Dos attack is smaller very few attack sequences are detected as compared to when the time interval is increased, more of the candidate patterns become frequent and as a result the average length of patterns increases. Thus fewer intervals are prone to have a larger number of patterns as well as longer average length. Also, as the fuzzy time interval patterns are discovered the support of the candidate patterns is spread over several fuzzy interval, rather than only one interval as the crisp interval.

Also, the average length of the Dos attack detected reduces as the time interval narrows down. Although on keeping all the parameters same of the two different dataset, experimental results have shown that fluctuation of average length patterns are negligible when the time interval of the sequence dataset generated is at a unit difference.

Intrusion detection based on sequential pattern mining is still a research focus with emphasis on improving and optimizing sequential pattern mining algorithm.A possible extension is to apply fuzzy theory or rough set theory to partition the intervals, leading to a situation in which the boundary of an interval is no longer fixed but flexible. Given such an extension, the problem of sharp boundaries can be solved and a smooth transition provided between members and non-members of a set. Also as proposed in [7], the use of fuzzy genetic approach to discover optimized sequences in the network traffic data to classify and detect intrusion can also be implemented. Sequential pattern mining based intrusion detection using the cloud computing may be a trend in the future. The comprehensive analysis for intrusion behaviors from multiple angles by introducing other data mining techniques with the sequential pattern and implementing multi-level mining, so that providing more valuable intrusion information to security administrators and reducing false alarm rate will be also the goal of future research.

Compared intrusion detection with other applications, the background of its domain knowledge plays more prominent role in the rule mining. Especially some behavior signals which are sensitive to security status in the whole sequential data often play the marked role of specific patterns. For this feature, when applying sequential pattern mining to intrusion detection, one should not only consider the universality of current sequential pattern mining algorithms, such as the definition of support, etc, but also consider the inspirational role that domain knowledge(such as the service relation of attack behavior in the logic) plays in the pattern mining. The integration of these two aspects will be research trends for sequential pattern mining algorithms based intrusion detection

8. REFERENCES

[1] Guangjun Song, Zhenlong Sun, Xiaoye Li, "The Research of Association Rules Mining and Application in Intrusion Alerts Analysis", Second International Conference on Innovative Computing, Infonnatio and Control (ICICIC 2007),pp.567, 2007.

[2] Zhan Jiuhua, "Intrusion Detection System Based on Data Mining", First International Workshop on Knowledge Discovery and Data Mining (WKDD 2008), pp.402-405, 2008.

[3] Ya-Li Ding, Lei Li, Hong-Qi Luo, "A NOVEL SIGNATURe SEARCHING FOR INTRUSION DETECTION SYSTEM USING DATA MINING", Machine Learning and Cybernetics, 2009 International Conference on Volume I, pp.l22 - 126, 2009.

[4] Y. L. Chen, M. C. Chiang, and M. T. Ko, "Discovering time-interval sequential patterns in sequence databases", Expert Systems with Applications, Volume 25, Issue 3, October 2003, pp 343–354.

[5] Yen-Liang, Tony Cheng-Kui Huang, "Discovering Fuzzy Time-Interval Sequential Patterns in Sequence Databases", IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, 2005, vol.35, pp.959-972.

[6] Sunita Mahajan and Alpa Reshamwala, "Amalgamation of IDS Classification with Fuzzy techniques for Sequential pattern mining ",IJCA Proceedings on International Conference on Technology Systems and Management - ICTSM 2011, Number 3 - Article 7, pp 9–14.

[7] Sunita Mahajan and Alpa Reshamwala, "An Approach to Optimize Fuzzy Time-Interval Sequential Patterns Using Multi-objective Genetic Algorithm", ICTSM 2011, CCIS 145, pp. 115–120, 2011, Springer-Verlag Berlin Heidelberg 2011.

[8] R. Agrawal and R. Srikant, "Mining sequential patterns", In Proc. Int. Conf. Data Engineering, 1995, pp. 3–14.

[9] Pei, J., Han, J., Pinto, H., Chen, Q., Dayal, U., & Hsu, M.-C., "PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth", Proceedings of 2001 International Conference on Data Engineering, pp. 215–224.

[10] Han, J., Pei, J., Mortazavi-Asl, B., Chen, Q., Dayal, U., & Hsu, M.-C., "FreeSpan: Frequent pattern-projected sequential pattern mining", Proceedings of 2000 International Conference on Knowledge Discovery and Data Mining, pp. 355–359.

[11] Srikant, R., & Agrawal, R., "Mining sequential patterns: Generalizations and performance improvements", Proceedings of the 5th International Conference on Extending Database Technology,1996, pp. 3–17.

[12] Zaki, M. J., "SPADE: An efficient algorithm for mining frequent sequences", volume 42 Issue 1-2, January-February 2001, pp 31–60.

[13] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules", Proceedings of 20th VLDB Conference Santiago, Chile, 1994, pp. 487–499.

[14] Yangdong Ye, Qing Zhou , Xu Wang, Limin Jia, "Analysis of fuzzy time interval in the hybrid Petri net model of train operation system", 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), 2010.Vol. 1 pp. V1-644 - V1-648.

[15] Chung-I Chang, Hao-En Chueh , Lin, N.P., "Sequential Patterns Mining with Fuzzy Time-Intervals", Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 2009. FSKD '09, Vol 3,pp. 165 – 169.

[16] Chung-I Chang, Hao-En Chueh, Yu-Chun Luo, "An integrated sequential patterns mining with fuzzy time-intervals", International Conference on Systems and Informatics (ICSAI), 2012, pp. 2294 - 2298.

[17] S. Hofmeyr, S.Forrest, A.Somayaji. "Intrusion Detection Using Sequences of System Calls," Journal of Computer Security, 1998.Vol. 6 pp.151-180.

[18] Yin, Qing-Bo, Zhang, Ru-Bo, Li, Xue-Yao and Wang, Hui-Qiang "Research on Technology of Intrusion Detection Based on Linear Prediction and Markov M odel," Chinese Journal of Computers, 2005, Vol. 28, no. 5, pp. 900-907.

[19] Lee W, Stolfo L S, Mok K W. "A Data Mining Framework for Adaptive Intrusion Detection," Proceedings of the 1999 IEEE Symposium on Security and Privacy, 1999. pp. 120-132.

[20] Karlton Sequeira, Mohammed Zkai. "ADMIT: anomaly-based data mining for intrusions," Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 2002.

[21] XUE Anrong, HONG Shijie, JU Shiguang, CHEN Weihe, "Application of Sequential Patterns Based on User's Interest in Intrusion Detection", Proceedings of 2008 IEEE International Symposium on IT in Medicine and Education, pp 1089- 1093, 2008.

[22] Zhang Shengbin, XI Hongsheng, WANG Weiping."Computer Intrusion Detection Based on PrefixSpan," Computer Engineering, 2003.

[23] SONGShi-Jie, HUANGZun-Guo, HUHua- Ping, JINShi-Yao. "A Sequential Pattem Mining Algorithm for Misuse Intrusion Deteetion," Workshop: Information Security and Survivability for Grid. Oct.21-24, 2004, Wuhan, China. pp.458-465.

[24] CAI Weihong, LIU Zhen and WANG Meilin, "Intrusion Detection Based on Fuzzy Logic and Immune GA," Computer Engineering, 2006.

[25] Duan Yi-feng , Hu Gu-yu , Ding Li. "An Application of Sequential Pattern Mining in Network Alarm Data Analyses," Journal of Beijing University of Posts and Telecommunications, 2004.12.

[26] XIN Hong-liang, OUYANG Wei-min and ZHU Wan-tao,. "Audit-oriented sequence mining algorithm with strict constraints," Computer Applications, 2006.

[27] Weijun Zhu, Qinglei Zhou, Ping Li, "Intrusion detection based on model checking timed interval temporal logic", IEEE International Conference on Information Theory and Information Security (ICITIS), 2010, pp. 503 – 505.

[28] Milanesi, G., Sarti, A., Tubaro, S., " Robust real-time intrusion detection with fuzzy classification", International Conference on Image Processing. 2002, vol.3,pp. III-437 - III-440.

[29] Kai Xing Wu, Juan Hao, Chunhua Wang, "Application of Fuzzy Association Rules in Intrusion Detection", International Conference on Internet Computing & Information Services (ICICIS), 2011, pp. 269 – 272.

[30] Yongzhong Li, Rushan Wang , Jing Xu, Ge Yang, Bo Zhao, "Intrusion Detection Method Based on Fuzzy Hidden Markov Model", Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 2009. FSKD '09, vol. 3, pp. 470 – 474.

[31] Jianxiong Luo, Bridges, S.M., Vaughn, R.B., Jr., "Fuzzy frequent episodes for real-time intrusion detection", The

10th IEEE International Conference on Fuzzy Systems, 2001, vol. 1, pp. 368 – 371.

[32] Ming-Yang Su, Sheng-Cheng Yeh, Kai-Chi Chang, Hua-Fu Wei, "Using Incremental Mining to Generate Fuzzy Rules for Real-Time Network Intrusion Detection Systems", 22nd International Conference on Advanced Information Networking and Applications - Workshops, 2008. AINAW 2008, pp. 50 – 55.

[33] Xiaogang Wang, Junzhou Luo, Ming Yang, "An efficient sequential watermark detection model for tracing network attack flows", IEEE 16th International Conference on Computer Supported Cooperative Work in Design (CSCWD), 2012, p. 236 – 243.

[34] Reshamwala, A., Mahajan, S., "Prediction of DoS attack sequences", International Conference on Communication, Information & Computing Technology (ICCICT), 2012, pp. 1 – 5.

[35] Reshamwala, A., Mahajan, S., "Detection of DoS attack time interval sequences on network traffic", World Congress on Information and Communication Technologies (WICT), 2012, pp. 739 – 744.

[36] Lee W and Stolfo S J, "Data mining approaches for intrusion detection", Proceedings of the 7th USENIX Security Symposium, :26-29, 1998.

[37] SPMF: "Sequential Pattern Mining Framework".