Sequential Pattern Mining for Intrusion Detection System with Feature Selection on Big Data

Fidalcastro.A[1], Baburaj.E[2]
[1] Research Scholar, Correspondent Author , Sathyabama University
Department of Computer Science and Engineering
Chennai, India.
[E-mail: fidalcastro@gmail.com]
[2] Professor, Sun college of Engineering and Technology
Department of Computer Science and Engineering
Nagercoil, India.
[E-mail: alanchybabu@gmail.com]

Received May 3, 2016; revised November 9, 2016; revised December 29, 2016; revised February 22, 2017; accepted March 9, 2017; published October 31, 2017

Abstract

Big data is an emerging technology which deals with wide range of data sets with sizes beyond the ability to work with software tools which is commonly used for processing of data. When we consider a huge network, we have to process a large amount of network information generated, which consists of both normal and abnormal activity logs in large volume of multi-dimensional data. Intrusion Detection System (IDS) is required to monitor the network and to detect the malicious nodes and activities in the network. Massive amount of data makes it difficult to detect threats and attacks. Sequential Pattern mining may be used to identify the patterns of malicious activities which have been an emerging popular trend due to the consideration of quantities, profits and time orders of item. Here we propose a sequential pattern mining algorithm with fuzzy logic feature selection and fuzzy weighted support for huge volumes of network logs to be implemented in Apache Hadoop YARN, which solves the problem of speed and time constraints. Fuzzy logic feature selection selects important features from the feature set. Fuzzy weighted supports provide weights to the inputs and avoid multiple scans. In our simulation we use the attack log from NS-2 MANET environment and compare the proposed algorithm with the state-of-the-art sequential Pattern Mining algorithm, SPADE and Support Vector Machine with Hadoop environment.

Keywords: Sequential Pattern mining, Intrusion Detection System, Spoofing attack, Flooding of packets, Big Data, Feature Selection, Apache Hadoop YARN

1. Introduction

The malicious activities in the system and network are monitored by a software application or device which is termed as Intrusion Detection System. Intrusion is the process of attempting to break into or misuse your system. Intruders may be the legitimate users of the network or from the outside of the network. Intrusion can be physical, remote or system intrusion [1]. The pattern can be discovered and recognition by intrusion detection system. Techniques for intrusion detection are anomaly-based, misuse-based, specification–based. Anomaly based is the deviation from the normal behavior of the network. Misuse based compares with the known attack signature with the current system activities. Specification based is the runtime violations of specification of the network [2]. The different ways to intrude are buffer overflows, unexpected combinations and unhandily input and race conditions. Denial of Service attack (DOS) and flooding of packets may be due to the growth of data which leads to such kind of threats and attacks.

Data mining is a process for computing the database which is used for discovering patterns in large set of data concerning several methods in database system. Data mining is the analysis of observational datasets to find unsuspected relationships. [7] Data sets may be in the mold of some set of measurement taken from some process or environment. A data mining algorithm is a well defined procedure that takes data as input and generates output in the form of models and patterns. [8]

Sequential pattern mining is used to detect the attack in the large datasets. Sequential pattern mining is a topic of data mining concerned with finding statistically relevant patterns generated by the input data where the values are delivered in sequence [12]. The time regularity of items with time stamps in the database is denoted by sequential pattern mining. To find the huge number of possible sequential patterns is hidden in databases is a great challenge of sequential pattern mining. Some of attacks leaves signature as sequences that can be found through sequential pattern mining algorithms.

Feature selection is the process of selecting the minimum number of features from the ‘m’ number of features in the given set of data. According to some selection criteria the feature selection method is used to select the significant subset of the given attributes[15]. In pattern recognition and Data mining groups Feature Selection is one of the active research areas. Feature election can be done in two different ways 1) Supervised Learning and 2) Unsupervised Learning. For high dimensional data, Un-supervised learning is done because it contains large amount of redundant data. [17]

Fuzzy logic is one of the feature selection techniques for the extraction of features in the input log. Since 90’s Fuzzy logic is used with Intrusion Detection System because which has the ability to deal with complexity and uncertainty. Fuzzy logic is used in simple micro controller to large control systems as problem solving methodology. [17]

Traditional rules are derived from frequent item sets, which only consider the occurrence of items but do not reflect any other factors. In Weighted sequence rule mining transactions are attached with weighted values according to some criteria. Some interesting rules may not be found by standard mining. To find out different kinds of interesting patterns from a set of data with item weight or transaction weight. The weights are calculated for different ways for different application. Here we use the fuzzy logic for future selection so fuzzy weighted rule mining to optimize the rules.

Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, cu ration, search, sharing,

storage, transfer, visualization, and information privacy. In recent years big data which is popularly known for its wide range of growth in data. “SIZE” is a constantly moving target, as of 2011 ranging from a few dozen tera bytes to petta bytes of data. In future the growth of data may range upon zetta bytes of data. Big data is mainly based on three characteristics (3V’s): 1) Growth of data size-Volume (terabytes to zetta bytes of data), 2) Increase in speed- Velocity (batch to streaming data) and 3) Types of data- Variety (structured, unstructured and semi-structured data). The world is moving fast so now the fourth V is emerging which is termed as veracity meant for the doubtful of data. [25]

Hadoop is a commodity hardware which is an open-source software framework with consists of distributed file system for storage and it is used for large scale processing of datasets. Hadoop is Apache foundation software. Hadoop Map-Reduce is a software framework for easily writing application which process vast amount of data (multi-tera byte datasets) in parallel on large clusters (thousand of nodes) of commodity hardware in a reliable, fault-tolerant manner. Hadoop consists of Job Tracker and Task Tracker. Only one Name node is present and several Data nodes. Hadoop Map-Reduce concept is used to reduce the time and space complexity and give greater throughput and scalability. [26]

Here the contributions to improve speed and reduce time and space consumption. At the first stage find features from the input log by fuzzy logic feature selection which is related to attacks. The second stage setting up Big data environment and here the input log is into Hadoop Map-Reduce environment. At the third stage a novel sequential pattern mining algorithm applied to Hadoop Map-Reduce environment for attack pattern detection. In fourth stage novel algorithm compared with Big Data environment for speed and accuracy for Intrusion detection.

2. Related Work

The Intrusion Detection System (IDS) detects anomalies and attacks in the network using semi supervised approach, varying HOPERAA algorithm and Hybrid IDS model. IDS are integrated with Data mining to identify the hidden data of interest and it is used to detect denial of service attack in the processing set of data. Due to Denial of Service attack the packets may flood in the network and flooding of packets may occur because the attackers may hold the network as the legitimate users [1].

Markovian IDS an intrusion detection system to protect nodes from malicious attacks. The Markovian IDS engender attack-pattern mining algorithm to forecast the future attack patterns and to put off the node from malicious attacks. Markovian IDS have greater success when compared with game theory. Flooding may suggest itself in the transport layer which leads to power supply failure due to Denial of Service attacks [2].

Fuzzy logic based on intrusion detection system is used for the analysis and selection of features that are generated from the input log. Most of the fuzzy based ids consist of very limited features for data collection in dynamic environment [3].

SPADE algorithm is used for fast discovery of sequential patterns. In SPADE the sequences are generated only in three database scans this avoids the problem of repeated database scans and simple join operation are also use to solve the problem of repeated data scans . This algorithm generates the patterns for detecting the Denial of Service attacks and BFS concept is used for horizontal and vertical scanning of databases [7].

Here proposed system act as a genetic feature selection wrapper to search for an optimal feature subset for dimensionality reduction and evaluate classification and feature selection performance and compared with some well-known classifiers as well as feature selection wrappers and filters[14].Sequential pattern mining is popular research topic. A subsequence is

called sequential pattern or frequent sequence. Frequent sequence frequently appears in a sequence database, and its frequency must be less than user-specified minimum support threshold [7] [10][11].

The vertical algorithms to perform well on datasets having long and dense sequences and to have excellent overall performance compared with algorithms using the horizontal format [8][9][11][12]. The papers talk about fuzzy logic and how it incorporated with intrusion on filter methods [17] [18] [19]. Mining Weighted Association Rules, Sequence rules and ranking methods discussed briefly [20][21][22][23].

The network intrusion traffic focuses on the specific problem of big data classification. Hadoop distributed file system HDFS combined with cloud computing and Hive database solve the problem of big data classification. The extraction of useful information from the data is obtained through training and validation by traditional machine learning techniques [25]. The new opportunities and challenges of big data mining are discussed to overcome the problem due to space, speed and time complexity. The concept of parallel processing is implemented in Hadoop Map Reduce. [25] [26].

The focus of big data analytics is mainly to gain on valuable and valid insights from big data [27]. DIKW stands for Data, Information, Knowledge, and Wisdom. Wisdom denotes non-probabilistic, non-deterministic and exploration data process [28]. NoSQL (Not Only SQL) and Hadoop Distributed File System (HDFS) is used for storing unstructured data in large scale data analytics using Map Reduce framework. Massive scalability, high agility, deep analytic and low latency are the concepts of big data analytics [29].

There are number of tools proposed for Batch processing but Hadoop is the best among all the tools for Batch processing which is done on the top of Hadoop framework [30]. For a large scale organization an emerging approach to detect intrusion detection is big data analytics. Security analytics sources (SASs) helps to prevent the messages from tampering by the intruders. The concept of pillar Box enforces integrity and stealth against tampering of data from attackers who controls the network [31].

This paper they identified problem with this process is that the accuracy algorithm which is used may not identify entire patterns in the logs. This type of challenges can affect in two ways. 1) Missing with regular patterns. 2) The detection neglects some new patterns. They are using novel data mining method by using new Modified Apriori Algorithm [34]. In this paper the authors introduced a novel sequential pattern mining technique called Rare Sequential Pattern Mining (RSPM) technique. The algorithm is useful in SCADA to find anomalies.

3. Proposed Work

Fuzzy logic with sequential data mining is used to detect the attacks in our intrusion detection system. Feature selection is used to avoid over-fitting of patterns and reduce the computational overhead. To find the minimum subset of features from the input log features by fuzzy logic as a filter. The selected features are only input to the first frequent pattern and second frequent set selected through fuzzy weighted minimum support in FL_SPADE algorithm, the patterns are generated to detect the attacks in network log. The development environment is Hadoop version 2.6 (YARN) which is a programming framework used for Map-Reduce function. By using Map-Reduce, the Space and Time complexity can be reduced.

3.1. Proposed System Architecture:

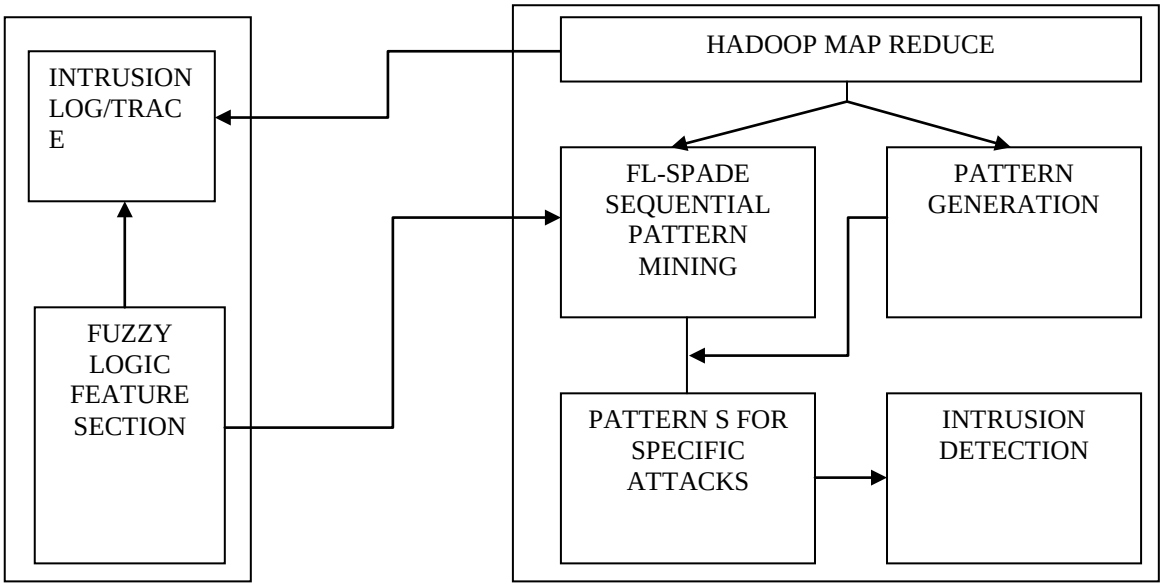

Fig. 3.1. System Model for Attack Detection

STEP 3: Network data save into Map-Reduce environment which consist of Hadoop distributed file system and cluster resource management maps and reduce the task .The Job tracker assigns the task to the Task tracker. Shuffle and Sort is an intermediate function between maps and reduce.

Then SPADE and FL_SPADE applied to get patterns. Hadoop version 2.0 consists of cluster resource management and Hadoop distributed file system which is redundant and reliable storage. Map Reduce is used for data processing and HDFS creates multiple replicas of data blocks for reliability. Hadoop works on the basis of key-value pairs.
The Map instance is denoted by

Map (key, value) -> (key', value')

The Reduce function is denoted as,

Reduce (key', value') -> (key'', value'')

3.2 Fuzzy Feature Selection

Feature selection is used to avoid over-fitting of patterns and reduce the computational overhead. Fuzzy logic technique is used to select the specific features from the input log. The Fuzzy logic is used to select the features by fixing the value ranging from 0 to 1. Fuzzy Rules are used in Fuzzy logic which is one of the important applications of Fuzzy theory. Fuzzy sets are sets without any fixed boundaries. Fuzzy Interface system can be created and edited by Fuzzy logic tool box. Fuzzy set is a pair which can be denoted by (V, M). Where V is a set, M: V-> [0, 1] [15] .

3.2.1 Filter-based approaches

The feature selection experiments are conducted using filter-based approaches on the training data. The objective was filter can achieve the best performance for intrusion detection data, and also suggests a good feature subset that contains relevant features with relative order of importance for baseline reference. Various filtering criteria applied in this experiment to measure the relevance of features from training data. [14]

Information gain (IG) It is known as mutual information, before and after observing features, it measures the expected reduction in entropy of class. Selects features by Larger difference. IG is measured as

$$\text{InfoGain}(S, F) = \text{Entropy}(s) - \sum_{v \in V(F)} \frac{|Sv|}{|S|} \text{Entrophy}(Sv) \quad \ldots.(1)$$

where S is the pattern set, Sv is the subset of S where feature F has value v, |S| is the number of samples in S, and v is value of feature F. The entropy of class before observing features is defined as

$$\text{Entrophy}(S) = \sum_{c=C}^{n} -\frac{|Sc|}{|S|} \log_2 \frac{|Sc|}{|S|} \quad \ldots.(2)$$

where Sc is the subset of S belonging to class c. C is the class set and IG is the fastest and simplest ranking method. [14]

Gain ratio (GR) normalizes the IG by dividing it by the entropy of S with respect to feature F, in order to discourage the selection of features with many uniformly distributed values. GR is measured as

$$\text{GainRatio}(S,F) = \text{InfoGain}(S,F)/\text{SplitInfo}(S,F) \quad \ldots.(3)$$

$$\text{SplitInfo}(S, F) = \sum_{i=1}^{n} -\frac{|Si|}{|S|} \log_2 \frac{|Si|}{|S|} \quad \ldots..(4)$$

where Si is the subset of S where feature F has its i'th possible value, and n is the number of subclasses split by feature F. [14]

Chi-square (CS) measures the chi square statistics of each individual feature with respect to the classes. According to the descending order of their chi square values the features are ranked. Strong correlation with the classes are the large chi square values for the features. The feature F is defined as

$$\text{ChiSquare}(F) = \sum_{i=1}^{m} \sum_{j=1}^{k} (Aij - Eij)/Eij \quad \ldots.(5)$$

$$Eij = (Ri.Cj)/|S| \quad \ldots.(6)$$

k is the number of classes, Ri is the number of samples in the i'th interval and Cj is the number of samples in the jth class. m is the number of intervals discretized from the numerical

values of F, Aij is the number of samples in the i'th interval with j'th class, and Eij is the expected occurrence of Aij. [14]
The features are selected and then given as input to the sequential pattern mining algorithm. The total number of features are shown below.

Table 1. Total Number of Features

Serial Number	Features of a Node
1	Number of neighbours
2	Number of added neighbours
3	Number of removed neighbours
4	Number of active routes
5	Number of invalidated routes
6	Number of added routes by route discovery mechanism
7	Number of updated routes
8	Number of added routes under repair
9	Number of received route request packets destined to this node
10	Number of received route request packets to be forwarded by this node
11	Number of broadcasted route request packets from this node
12	Number of forwarded route request packets from this node
13	Number of received route reply packets destined to this node
14	Number of received route reply packets to be forwarded by this node
15	Number of received broadcast route error packets (to be forwarded or not)
16	Number of broadcasted route error packets from this node
17	Number of received total routing protocol packets
18	Number of routes under repair
19	Number of added routes by overhearing
20	Number of invalidated routes due to expiry
21	Number of invalidated routes due to other reasons
22	Number of received route reply packets to be forwarded by this node
23	Number of initiated route reply packets from this node
24	Number of received total routing protocol packets to be forwarded
25	Number of initiated total routing protocol packets from this node
26	Number of forwarded total routing protocol packets by this node

In **Table 1** shows the total number of features available in Mobile ad hoc networks.The features which appeared in the extracted rules are selected as the relevant and important features for Intrusion detection. All the other features are considered as irrelevant. Both an increase in detection rate and a decrease in false positive rate are expected. The features selected here by Fuzzy Logic is as follows.

Table 2. Selected Features

S.No	Filter Based Approaches	List of Features
1	Information Gain	9,10,11,12,13,14
2	Gain Ratio	1,2,7,9,10,12,14
3	Chi Square Measures	1,2,9,11,12,13

Table 2 shows list of filter based approaches used and the outcomes are listed

3.1.2. ALGORITHM FL_SPADE

FL-Spade uses selected features from fuzzy feature selection for Frequent 1 sequences and weighted fuzzy rule mining was used to generate Frequent 2 sequences. The below definitions are for weighted fuzzy rule mining and the FL-SPADE algorithm.

In Weighted Association rule mining transactions are attached with weighted values according to some criteria. Some interesting rules may not be found by standard mining. To find out different kinds of interesting patterns from a set of data with item weight or transaction weight. [21]

Fuzzy Item Weight FIW is a value attached with each fuzzy set. It is a non-negative real number value range [0..1] with respect to some degree of importance. Weight of a fuzzy set for an item i_j is denoted as i_j $[l_k$ [w]]. [21]

Fuzzy Itemset Transaction Weight FITW is the aggregated weights of all the fuzzy sets associated to items in the item set present in a single transaction. Fuzzy Item set transaction weight for an item set (X, A) is calculated as vote for t_i' satisfying

$$X = \prod_{k=1}^{|X|} (\forall [i[w]] \in X) t_i [i_k[w]] \quad \text{.....(7)}$$

Fuzzy Weighted Support FWS is the aggregated sum of FITW of all the transactions item set is present, divided by the total number of transactions. It is denoted as: [21]
FWS(X)=Sum of votes satisfying X/Number of records in T

$$\sum_{i=1}^{n} \prod_{K=1}^{|X|} (\forall [i[w]] \in X) t_i [i_k[w]] / n \quad \text{.....(8)}$$

Fuzzy Weighted Confidence FWC is the ratio of sum of votes satisfying both X ∪ Y to the sum of votes satisfying X with Z = X ∪ Y . It is formulated as: [21]

$$FWC (X \rightarrow Y) = FWS (Z) / FWS (X) \quad \text{.....(9)}$$

$$\sum_{j=1}^{n} \frac{\prod_{k=1}^{|Z|}\left(\forall[i[w]] \in Z\right) t_i[Z_k[w]]}{\prod_{k=1}^{|X|}\left(\forall[i[w]] \in X\right) t_i[x_k[w]]} \quad \ldots.(10)$$

FL_SPADE Algorithm:

FL_SPADE (FWS, D):

```
f1 = {frequent items or 1-sequences - fuzzy logic feature selection};
f2 = {frequent 2-sequences – fuzzy logic weights FITW,FWS,FWC };
e = {equivalence classes[X]};
For all [X] belongs to e do Enumerate-frequent-Seq-fuz[X]
Enumerate-frequent-seq-fuz(S, FWS)
for all Ai  in S
        Ti  <- {}
            for all Aj  in S, with j≥i
                R<- Ai v Aj   (join)
            if(Prune(R)==FALSE) then
                L(R)=L(Ai) intersects L(Aj)
                if R satisfies FWS then
                        Ti  <- Ti U {R}
            end
Enumerate_frequent_seq (Ti, FWS) //……DFS
end
For all non-empty Ti
Enumerate_frequent_seq (Ti, FWS) //……BFS
```

The main steps include the computation of the frequent 1-sequences and 2-sequences, the decomposition into prefix-based parent equivalence classes, and the enumeration of all other frequent sequences via BFS or DFS search within each class.

Computing frequent 1-sequences and 2-sequences

In the horizontal format the database consists of a set of input-sequences. Each input-sequence has a set of events, along with the items contained in the event. In contrast our algorithm uses a vertical database format, where we maintain a disk-based id-list for each item. Each entry of the id-list is a sid, eid pair where the item occurs. This enables us to check support via simple id-list joins.

Computing F1: Given the vertical id-list database, all frequent 1-sequences can be computed by the fuzzy logic feature set. For each database item, we read its id-list from the disk into memory only where the features are present in feature set. We then scan the id-list, incrementing the support for each new sid encountered.

Computing F2: Let N D j F1j be the number of frequent items, and A the average id-list size in bytes .

1. Use a preprocessing step to gather the counts of all 2-sequences above only with the specific features using fuzzy weighted support(FWS) instead of minimum support. Fuzzy

weighted support gives weight age to each item and generated sequences must be optimum and more important sequences. Since this information is invariant, it has to be computed once, and the cost can be amortized over the number of times the data is mined.

2. Perform a vertical-to-horizontal transformation on-the-fly. For each item i , we scan its id-list into memory.

3.1.3 Support Vector Machine

Support Vector Machine classifies and constructs a set of hyperplanes or a hyperplane, which uses Lagrangian methods to minimize a regularized function of the empirical classification error. The SVM algorithm finds a linear hyperplane separation with a maximal margin in this hyperspace. Support vectors are the points that are lying on the margin. To get a good separation, the hyperplane's distance should be maximized to the nearest training point of any class. The good classification is achieved when the larger the distance of the hyperplane's. Here , we use libSVM library. (24)

For evolution two metrics are used detection rate and false positive rate. Here we use C-SVC algorithm in libSVM. S_SVC is evaluated with all the features and selected features from fuzzy logic. The performance and efficiency improved with fuzzy features.

4. Experimental Results and Analysis

The Simulation environment is carried out in NS-2 simulator installed in Linux Operating System. The Scenario consists of 25 wireless nodes. For routing the data AODV routing protocol is used. After creation of ns-2 simulation environment, the scenario files to be generated for mobile node movement and CBR traffic pattern.

Fig. 4.1. Sample NS2 Multiple Spoofing nodes present in the Network

Fig. 4.1 shows the screen shot of NS2 with spoofing node and attack present in the network and logs are collected and stored in hadoop environment for applying spade and fl-spade. It shows the screen shot of spoofing attack and logs are collected for applying sequential mining algorithms.

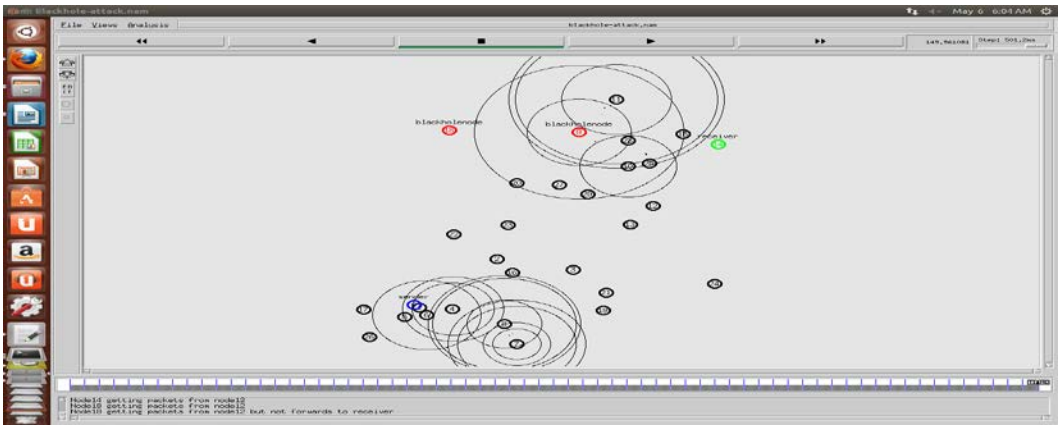

Fig. 4.2. Sample Blackhole nodes during transmission in the Network

Fig. 4.2 shows the screen shot of NS2 with black hole node and attack present in the network and logs are collected and stored in hadoop environment for applying spade and fl-spade. It shows the screen shot of Black hole attack and logs are collected for applying sequential mining algorithms.

The preprocessed NS2 trace file is the input to fuzzy logic to find features and applied to SPADE. FL-SPADE algorithm is used the selected features and fuzzy weighted support to find optimized rules to detect attacks. The above techniques are developed using Java. SPADE, FL-SPADE algorithms are compared with preprocessed synthetic attack log generated by the above mentioned scenario. The same algorithms are modified to fit with Hadoop environment and performance and accuracy are compared. The Map-Reduce, Job tracker, Shuffle and Sort are created accordingly.

5. Performance Evaluation

In this section, the implementation of AODV protocol is performed. Node movement generation and traffic pattern file is generated for node communication in the network. The SPADE, SVM and FL-SPADE are applied and compared. The Performance and Accuracy improved significantly for FL-SPADE.

Table 3. SPADE Performance on Attacks

Simulation	Detection Rate	False Positive Rate
medium traffic, low mobility	93.36%	0.8%
high traffic, low mobility	92.46%	2.5%
medium traffic, medium mobility	95.36%	1.45%
high traffic, medium mobility	94.45%	1.4%
medium traffic ,high mobility	91.38%	1.0%
high traffic ,high mobility	89.86%	1.8%

Table 4. SVM performance on Attacks

Simulation	Detection Rate	False Positive Rate
medium traffic, low mobility	97.63%	0.35%
high traffic, low mobility	94.90%	1.9%
medium traffic, medium mobility	97.50%	1.02%
high traffic, medium mobility	96.34%	1.1%
medium traffic ,high mobility	96.67%	0.75%
high traffic ,high mobility	92.86%	1.34%

Table 5. FL-SPADE Performance on Attacks

Simulation	Detection Rate	False Positive Rate
medium traffic, low mobility	98.76%	0.26%
high traffic, low mobility	96.01%	1.1%
medium traffic, medium mobility	98.40%	0.90%
high traffic, medium mobility	97.45%	0.98%
medium traffic ,high mobility	97.60%	0.30%
high traffic ,high mobility	96.75%	1.0%

Table 2, 3, 4 shows that State of art algorithm SPADE, SVM are applied and compared with FL-SPADE. Results confirm that accuracy and performance improved significantly in FL-SPADE.

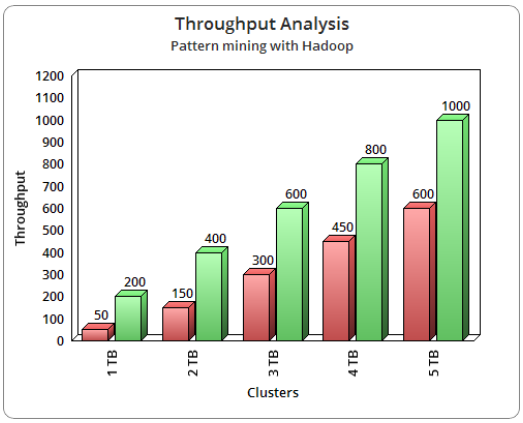

Fig. 5.1. Throughput Analysis

The above graph(**Fig 5. 1**) shows the performance of the network measured against intrusion in the MANET. The red line defines the detection of presence of attack in the network using SPADE in Hadoop Environment, and the green line depicts the FL-SPADE algorithm detects the attack and improves the performance.

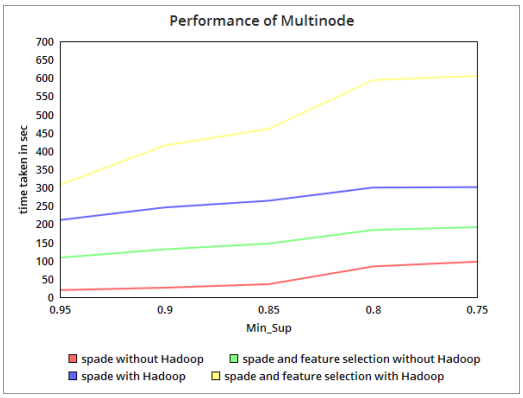

Fig. 5.2. Performance of Multinode

The above graph(**fig 5. 2**) shows the performance of the network measured against intrusion in the MANET. The red line defines the detection of presence of attack in the network using SPADE without Hadoop and green line shows Fuzzy future selection with FL-SPADE without Hadoop, blue line shows SPADE with Hadoop, yellow line shows fuzzy logic feature FL-SPADE with Hadoop it shows the detection was improved in every phase and improves the performance

6. Conclusion

The detection of attacks in Mobile Ad hoc networks with Fuzzy logic based algorithms is a novel kind of approach in wireless networks, and the approach helps in detecting attacks by generating optimized sequential rules based on the selection of relevant features and fuzzy weighted rule mining. The performance of mining algorithm has increased with the reduced feature set and optimized rules. Both increase in detection rate and decrease in false positive rate are observed. The generated patterns are applied on Hadoop map-reduce. Apache Hadoop YARN is used for map and reduces the patterns. Relevant patterns are grouped together and irrelevant patterns are grouped together. The proposed approach uses Apache Hadoop YARN, which solves the problem of repeated pattern by mapping the similar pattern together. Experimental results show that the algorithm has higher scalability and good performance, which makes it suitable to work with Big Data that is an advantageous to sequential pattern mining algorithms and optimized to detect the attacks. Detection rate of Intrusion detection system and throughput are increased.

References

[1] G.V Nadiammai, M. Hemalatha, “Effective approach toward Intrusion Detection System using Data Mining Techniques,” *Egyptian Informatics Journal*, December 2013. Article (CrossRef Link)

[2] Jen-Yan Hang, I-En Liao, Yu-Fang Chung, Kuen-Tzung Chen, *“Shielding Wireless Sensor Network using Markovian Intrusion Detection System with Attack Pattern Mining,” Information Sciences*, 29,March 2011. Article (CrossRef Link)

[3] A. Chaudhary, V.N.Tiwari and A. Kumar, “Analysis of Fuzzy Logic Based Intrusion Detection System in Mobile Ad Hoc Networks,” *BIJIT-BVICAM’S International Journal of Information Technology* 2014. Article (CrossRef Link)

[4] Portnoy, L., Eskin, E. And Stolfo, S., "Intrusion detection with unlabeled data using clustering," in *Proc. of the Workshop on Data Mining for Security Applications*, November 2001. Article (CrossRef Link)
[5] Y. Zhao, W. Liu, W. Lou, and Y. Fang, "Securing mobile ad hoc networks with certificate less public keys," *IEEE Trans. Dependable Secure Comput.*, vol. 3, no. 4, pp. 386–399, Oct.–Dec. 2006. Article (CrossRef Link)
[6] Agrawal, R., Ramakrishnan, S., "Mining sequential patterns," in *Proc. of 11th International Conference on Data Engineering*, pp. 3–14. IEEE, 1995. Article (CrossRef Link)
[7] Mohammed J. Zaki, "SPADE: An Efficient Algorithm For Mining Frequent Sequences," *Machine Learning, 42*, 31-60, Kluwer Academic Publishers, 2001.
Article (CrossRef Link)
[8] Aseervatham, S., Osmani, A., Viennet, E., bitSPADE, "A Lattice-based Sequential Pattern Mining Algorithm Using Bitmap Representation," in *Proc. of 6th Intern.Conf. Data Mining*, pp. 792–797. IEEE, 2006. Article (CrossRef Link)
[9] Ayres, J., Flannick, J., Gehrke, J., Yiu, T., "Sequential pattern mining using a bitmap representation," in *Proc. of 8th ACM SIGKDD International Conference on Knowledge Dis-covery and Data Mining*, pp. 429–435. ACM, 2002. Article (CrossRef Link)
[10] Unil Yun, "A new framework for detecting weighted sequential patterns in large sequence databases," *Science direct Knowledge-Based Systems 21*, 110–122, 2008.
Article (CrossRef Link)
Fayyad, U.M., and Irani, K.B., "The attribute selection problem in decision tree generation," in *Proc. of AAAI-92, Proceedings of the Ninth National Conference on Artificial Intelligence*, AAAI Press/The MIT Press, 104–110, 1992. Article (CrossRef Link)
[11] Sun, Ron, and C. Lee Giles, "Sequence learning: Paradigms, algorithms, and applications," Springer Science & Business Media, Vol. 1828, 2001. Article (CrossRef Link)
[12] Philippe Fournier-Viger, Antonio Gomariz , Michal Sebek, Martin Hlosta, "VGEN: Fast Vertical Mining of Sequential Generator Patterns." Article (CrossRef Link)
[13] C.-H. Tsang et al., "Genetic-fuzzy rule mining approach and evaluation of feature selection techniques for anomaly intrusion detection," *Pattern Recognition 40*, 2373 – 2391, 2007. Article (CrossRef Link)
[14] M. Ramze Rezaee, B. Goedhart, B.P.F. Lelieveldt, J.H.C. Reiber, *"Fuzzy feature selection," Pattern Recognition 32*, 1999. Article (CrossRef Link)
[15] Marion Leleu, Christophe Rigotti , Jean-Fran ̧cois Boulicaut, and Guillaume, "GO-SPADE: Mining Sequential Patterns over Datasets with Consecutive Repetitions," *MLDM 2003*, LNAI 2734, pp. 293–306, 2003. Article (CrossRef Link)
[16] Liu, H., and Setiono, R., "A probabilistic approach to feature selection - a filter solution," in *Proc. of International Conference on Machine Learning (ICML-96)*, July 3-6, 1996, Bari, Italy, San Francisco: Morgan Kaufmann Publishers, CA, 319–327, 1996.
Article (CrossRef Link)
[17] Siedlecki, W., and Sklansky, J,. "On automatic feature selection," *International Journal of Pattern Recognition and Artificial Intelligence*, 2, 197–220, 1998.
Article (CrossRef Link)
[18] M. Setnes, et al., "Similarity measures in fuzzy rule base simplification," *IEEE Trans. Syst. Man Cybernetics.—Part B: Cybernetics 28 (3) 376–386, 1998.*
Article (CrossRef Link)
[19] Ke Sun and Fengshan Bai, "Mining Weighted Association Rules without Preassigned Weights," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, VOL. 20, NO. 4, APRIL 2008. Article (CrossRef Link)
[20] Maybin Muyeba, M. Sulaiman Khan, Frans Coenen "Fuzzy Weighted Association Rule Mining with Weighted Support and Confidence Framework." Article (CrossRef Link)

[21] F. Tao, "Weighted association rule mining using weighted support and significant framework," in *Proc. of the Ninth ACMSIGKDD International Conference on Knowledge Discovery and Data Mining*, August, pp. 661–666, 2003.
Article (CrossRef Link)
[22] W. Wang, J. Yang, and P.S. Yu, "Efficient Mining of Weighted Association Rules (WAR)," in *Proc. of* ACM SIGKDD '00, pp.270-274, 2000. Article (CrossRef Link)
[23] Siedlecki, W., and Sklansky, J., "On automatic feature selection," *International Journal of Pattern Recognition and Artificial Intelligence, 2, 197–220, 1998.*
Article (CrossRef Link)
[24] Shan Suthaharan, "Big Data Classification: problems and challenges in Network Intrusion Prediction with Machine Learning," *University of North Carolina*, Greensboro, NC 27402, USA. Article (CrossRef Link)
[25] Dunren Che, Mejdl Safran, and Zhiyong Peng, "From Big Data to Big Data Mining: Challenges, Issues' and Opportunities," *Suthern LLinois University*, 62901, USA.
Article (CrossRef Link)
[26] Amir Gandomi, Murtaza Haider, "Beyond the Hype: Big Data Concepts, Methods and analytics," *International Journal of Information Management*, 2014.
Article (CrossRef Link)
[27] Gu Jifa, Zhang Lingling, "Data, DIKW, Big Data and Data Science," in *Proc. of 2nd International conference on Information Technology and quantitative Management, ITQM*, 2014. Article (CrossRef Link)
[28] Samson Oluwaseun Fadiya, Serdar Saydam, Vanduhe Vany Zira, "advancing big data for Humanitarian needs," *Humanitarian Technology :science ,Systems and Global Impact 2014*, HumTech 2014. Article (CrossRef Link)
[29] Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE, Gong-Qing Wu, and Wei Ding, Senior Member, IEEE, "Data mining with Big Data," *IEEE transactions on Knowledge and Data Engineering*, vol 26,no.1 January 2014. Article (CrossRef Link)
[30] Karthik Kambatlaa, Giorgos Kollias b, Vipin Kumarc, Ananth Gramaa, "Trends in Big Data analytics," *J.Parallel Distrib. Comput*, 2014. Article (CrossRef Link)
[31] Kevin D. Bowers, Catherine Hart, Ari Juels, "Securing the Data in Big Data Security Analytics." Article (CrossRef Link)
[32] Shruti Karde1 , Mettu Govind Rao2 , Rajesh Bhise3 "INTRUSION DETECTION AND ANOMALY DETECTION SYSTEM USING SEQUENTIAL PATTERN MINING," *IJRET: International Journal of Research in Engineering and Technology*, eISSN: 2319-1163 | pISSN: 2321-7308. Article (CrossRef Link)
[33] Anisur Rahman, Yue Xu, Kenneth Radke, Ernest Foo, "Finding Anomalies in SCADA Logs Using Rare Sequential Pattern Mining," *Network and System Security*, Volume 9955 of the series Lecture Notes in Computer Science pp 499-5. Article (CrossRef Link)

A. Fidalcastro completed B.E in CSE form Bharathidasan University in 2000 and M.E in CSE form Anna University, in 2004. Currently pursuing his Ph.D. from Sathyabama University, since 2010. He has published several papers in International conferences and journals. His Area of research interest includes Network Security and Dataming.
E-mail: fidalcastro@gmail.com

Dr. E. Baburaj completed B.E in CSE form Madurai Kamaraj Univeristy in 1992 and M.E in CSE from Madurai Kamaraj Univeristy, in 2002. He completed his Ph.D in the faculty of Information and Communication Engineering from, Anna University, Chennai in 2009. He is in teaching profession for the past twenty years. He has published twenty papers in International conferences and journals. He is also the consultant in Oasis Grace LLC, Oman. He is a life member in professional bodies like ISTE, CSI and UPA his research interest in the area of Wireless Sensor Networks and Cloud Computing.
E-mail: alanchybabu@gmail.com