Intrusion Detection System Using K-SVMeans Clustering Algorithm

[1]Jaisankar N, [2]Swetha Balaji, [3]Lalita S, [4]Sruthi D,

Department of Computer Science and Engineering,
Misrimal Navajee Munoth Jain Engineering College,
Chennai, India

Abstract

In recent years, as the usage of internet increases, new types of attack on network information keep increasing. Intrusion Detection System (IDS) is an important tool to identify various attacks to secure the networks. Traditional clustering algorithms work on “flat” data, making the assumption that the data instances are homogeneous in nature. Many real world data, however, is heterogeneous in nature. To handle the heterogeneity of the data, in this paper, we propose IDS using K-SVMeans clustering algorithm and other classification techniques. Here we use KDD-Cup99 dataset as simulation dataset for our experiment. The experimental results show that the proposed model reduces the time complexity, improves overall detection accuracy and minimizes the false alarm rate considerably.

Keywords – IDS; Fuzzy logic; spatial outlier detection; K- SVMeans Clustering; EC4.5; SVM; KDD-Cup99 Dataset.

I. Introduction

Recently, security of computer network has gained high precedence in everyone’s lives. Intrusion detection refers to all processes used in discovering unauthorized users of network or computer devices. This is achieved through specifically designed software with a sole purpose of detecting unusual or abnormal activity. In 1980, James P. Anderson [8] published a study outlining ways to improve computer security auditing and surveillance at customer sites. The original idea behind automated IDS is often credited to him for his paper on “How to use accounting audit files to detect unauthorized access”. This study paved the way as a form of misuse detection for mainframe systems.

The goal of an Intrusion Detection System (IDS) is to provide a layer of defence against malicious users of computer systems by sensing a misuse and alerting operators to on-going attacks. Most real-world data, especially data available on the web, possess rich structural relationships. Most of the clustering algorithms neglect the structural relationships between the individual data types. We present K-SVMeans clustering, which integrates two sources of information into a single clustering framework.

The behaviour of the intruder differs from that of a legitimate user in ways that can be qualified [9]. This type of behaviour can be identified using a data mining technique called spatial outlier detection and classification techniques.

A spatial outlier is a spatially referenced object whose non-spatial attribute values are significantly different from the values of its neighbourhood [4]. The value of the outlier is replaced by the average of its neighbour.

Decision trees are trees that classify instances by sorting them based on feature values. Each node in a decision tree represents a feature in an instance to be classified and, each branch represents a value that the node can assume.

A. Preliminaries

1) Fuzzy Decision Tree

Fuzzy decision tree induction has two major components [5]: fuzzy decision tree building and inference procedure for decision making. Inferences obtained from the tree can be written as a set of rules, which can then be used for developing systems. Fuzzy decision tree also allows data to be passed down multiple branches. The fuzzy decision tree construction involves the following: attribute value space partitioning methods, selecting of branching attribute, branching test method to decide which data follows down branches and leaf node labelling methods. In this paper, we have used spatial outlier detection algorithm as first level detection and in the second level we have proposed a new fuzzy based efficient C4.5 algorithm which reduces the time complexity and false alarm rates greatly and also improves detection accuracy.

This paper is organised as follows. Section II in brief describes a number of related studies in the areas of concern. The proposed system design is discussed in section III. The comparison of results and their discussions are explained in section IV. Finally, the conclusion and suggestions for future work are provided in section V.

II. Literature Survey

Thair Nu Phyu [1] examined the different classification techniques and as a result of the study, the decision tree induction method was found to be a more efficient way for classifying datasets. One of the most useful characteristics of decision trees is their comprehensibility, i.e. Decision trees can be easily understood by all. They also tend to perform better with discrete/categorical features.

Chih-Fong Tsai and Chia-YingLin [2] have proposed a hybrid machine learning model based on combining clustering and classification technique to detect attacks. K-means clustering is used for extracting the cluster centres and these cluster centres are used along with a data point in TANN approach for classifying malicious users.

Chang-Tien Lu et al. [7] presented different algorithms for spatial outlier detection to overcome the drawback of false detection by object which contain true spatial outliers in their neighbourhood.

Levent Bolleli et al. [13] proposed a novel clustering algorithm called K-SVMeans which handles the heterogeneity of the data and shows significant improvements over the traditional clustering methods.

S. Shekhar et al. [11] proposed methods to find out spatial outliers and Jisu Oh and Shan Huang [4] have proposed methods to eliminate the outliers by replacing their values with an average of their neighbours. Tien-Chin Wang and Hsien Da Lee [5] present a fuzzy decision tree model based on fuzzy set theory and information theory.

Jaisankar N et al. [15] proposed a Fuzzy Efficient C4.5 classification algorithm which is more efficient than SVM classifier in terms of detection accuracy and false alarm rate.

R. Shanmugavadivu and Dr.N.Nagarajan [6] have used automated strategy for generation of fuzzy rules, which are obtained from the definite rules using frequent items. They developed an anomaly based intrusion detection system in detecting the intrusion behavior within a network. A fuzzy decision-making module was designed to build the system more accurate for attack detection, using the fuzzy inference approach.

Salvatore Ruggieri [3] in his work describes a decision tree induction technique called EC4.5 which adopts 3 strategies over the C4.5 algorithm which helps attains a performance gain of up to 5 better than that of C4.5. All the above survey shows that the current data mining techniques have been used effectively in the current IDS but, still there is a possibility to improve the speed of the algorithm, to improve the detection rate and also to reduce the false alarm rate.

III. System Design

The overall system design of the proposed work is shown in Fig. 1. It consists of four modules: clustering module, outlier detection module, classification module and decision module.

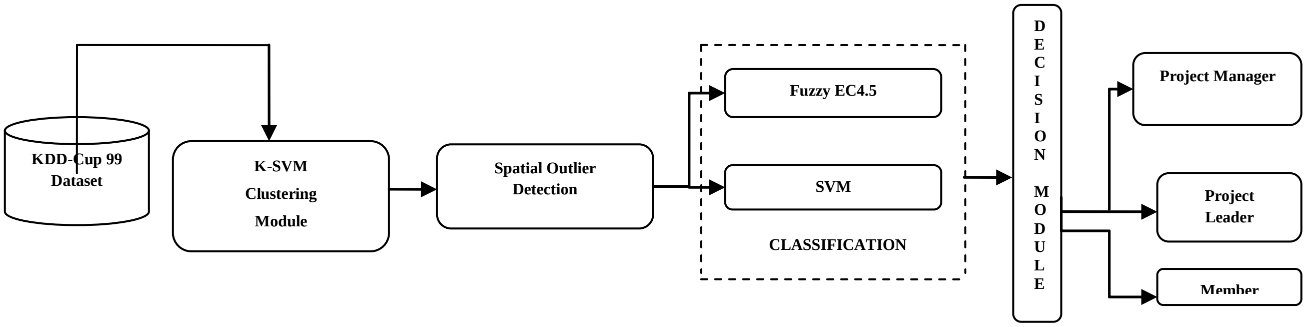

Fig. 1: Overall System Design

A. KDD-Cup Dataset

In our proposed system, we use KDDCup99 Dataset as input. There are 41 different features in a KDDCup99 dataset. This feature is classified into three different groups. They are Basic Type, Content Type and Traffic Type. In these features, 9 features are of basic type, 13 features are of content type and 19 features are of traffic type. There are 23 types of

attacks contained in training information and 37 types of attacks contained in test information, 14 types of attacks more than training information.

The simulated attacks fall in one of the following four categories:

- Denial of Service Attack (DoS): Perpetrators project illusions of some computing or memory resource fully engaged or unavailable to authorized users.
- User to Root Attack (U2R): is a class of exploit in which the attacker starts out with access to a normal user account on the system (perhaps gained by sniffing passwords, a dictionary attack, or social engineering) and is able to exploit to gain root access to the system.
- Remote to Local Attack (R2L): occurs when an attacker who has the ability to send packets to a machine over a network, but does not have an account on that machine, exploits some vulnerability to gain local access as a user of that machine.
- Probing Attack: is an attempt to gather information about a network of computers for the apparent purpose of circumventing its security controls.

B. Clustering Module

K-SVMeans, integrates the well-known K-Means clustering with the highly popular Support Vector Machines (SVM). The clustering along the main data type of interest is performed using the K-Means algorithm and relational similarity in the additional dimension is learned through on- line Support Vector Machines [14]. During K-SVMeans clustering process, an SVM is trained for each cluster on the additional (secondary) dimension of the data. K-SVMeans can be run in multiple iterations where the initialization of the SVM learner is performed by using the clustering solution generated in the previous run. In the first iteration, we run standard K-Means algorithm to yield a clustering based on the primary data type X. In the beginning of an iteration t + 1, K-SVMeans looks at each cluster generated in the previous run and selects m objects closest to the centroid of that cluster.

C. Spatial Outlier Detection

A spatially referenced object whose non-spatial attribute values are significantly different from its neighbours is termed as a spatial outlier. The algorithm is divided into two subparts: (a) Model construction and (b) Outlier detection. The average attribute value of the neighbours is first calculated. Secondly, objects with significant difference in values are spotted and detected as outlier and these values are substituted by the average value of its neighbours.

D. Classification Module

The KDD-Cup99 dataset is taken for analysis and out of 41 attributes, 34 continuous valued attributes are selected. 34 attributes are grouped into 5 different classes which form the fuzzy data. Definite and indefinite rules are generated using the maximum and minimum deviation value of normal and attack attributes. Then we calculate the information gain for each attribute, among which the attribute with the highest information gain is chosen as the splitting criterion for the decision tree construction. The Efficient C4.5 (EC4.5) algorithm is used for classification, which adopts the best among three strategies of computing the information gain of continuous attributes. All the strategies adopt a binary search of the threshold in the whole training set starting from the local threshold computed at a node. The first strategy computes the local threshold using C4.5 algorithm which uses quick sort method to sort the cases. The second strategy also uses the same C4.5 algorithm but uses counting sort to sort the cases. The third strategy calculates the local threshold using main memory version of rain forest algorithm which does not need sorting. Support Vector Machines (SVM) are Supervised learning machines that plot the training vectors in high dimensional feature space, labelling each vector by its class. The data is also linearly separable. The linear SVM searches for a hyper plane with the largest margin. Computing such a hyper plane to separate the data points leads to a quadratic optimization problem. There are two main reasons why we have used SVM in this paper for intrusion detection. The first reason is its performance in terms of execution speed and the second reason is its scalability. SVM does not depend on dimensionality of the feature space.

E. Decision Module

The authorized and malicious users were successfully identified. The results are then sent to the Administrator, who then forwards it to the Project Manager, Project Leader and Members.

IV. Results and Discusssion

A. Performance Analysis

The KDD-Cup99 dataset is used for our experiment. The dataset has 41 attributes for each connection record plus one class label. Here we have used 6890 records of training data for our experiment. R2L and U2R attacks do not have any sequential patterns like DoS and probe. When using training set for experiment it constructs a model to give maximum detection accuracy. As the dataset has 5 different major classes we perform a 5 class classification. Although decision tree algorithm is capable of handling 5 class classification problems, we have used fuzzy efficient C4.5 classifier in this work so that the comparison with SVM classifier will make sense.

The table 1, table 2, Fig 2 and Fig 3 summarise the results and experiments. The experimental results show that the time consumed by fuzzy EC4.5 is much better than C4.5 classifier.

B. Comparison Analysis

Initially, there is no significant difference between accuracy of the two methods, but C4.5 is slightly better when proportion of normal information is small. When the proportion of normal information is high (>75%), SVM is better. On the average, C4.5 is slightly better than SVM. In comparison of false alarm rate, Fuzzy Efficient C4.5 is better than SVM when the proportion of normal information is 30%, 40%, 60%, and for other percentage of data set SVM is better.

After applying spatial outlier detection, same experiments are conducted. In detection rate analysis, fuzzy efficient C4.5 algorithm detects better than SVM. In case of false alarm rate analysis SVM further reduces the false detection greatly. These are shown in table 2 and Fig 3.

TABLE 1: Detection Accuracy Analysis between Fuzzy C4.5 and Sum before Outlier Detection

% of Dataset	Fuzzy EC4.	SVM
10	73.40	67.34
20	71.95	64.75
30	74.86	59.00
40	66.85	52.90
50	71.65	67.35
60	71.25	48.00
70	68.90	41.60
80	66.35	49.25
90	66.50	59.00
AVERAGE	70.19	56.5767

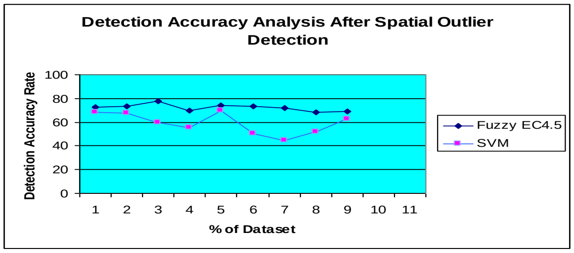

Fig 2: Detection Accuracy Analysis after Spatial Outlier Detection

TABLE 2: False Alarm Rate Analysis before Spatial Outlier Detection

% of Dataset	Fuzzy EC4.5	SVM
10	2.90	2.10
20	2.00	1.75
30	1.95	2.25
40	1.40	0.92
50	1.45	1.30
60	0.85	1.50
70	1.05	0.65
80	1.55	0.25
90	1.20	0.70
AVERAGE:	1.5944	1.2689

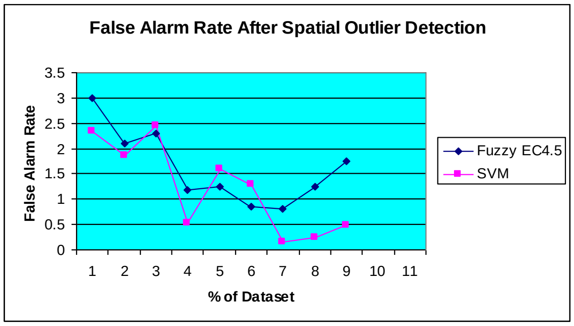

Fig 3: False Alarm Rate after Spatial Outlier Detection

V. Conclusion

We have developed an intrusion detection system for detecting the intrusion behaviour within a network. A K-SVMeans clustering algorithm was used which handles the heterogeneity of the data more efficiently. A spatial outlier detection algorithm was adopted for our experiment, to detect outliers which simplify the data for further classification. The comparison analysis shows that the fuzzy based efficient C4.5 classifier detects the attacks better than SVM, and after spatial outlier detection not only detection accuracy was improved, but the false alarm rate and time complexity were also greatly reduced. Further work in this direction can be carried out by improving EC4.5 and SVM classifiers using concepts of rough set theory and fuzzy rough set theory.

References

[1] Thair Nu Phyu, "Survey of Classification Techniques in Data Mining", Proceedings of the International Multi Conference of Engineers and Computer Scientists 2009 Vol I, IMECS 2009, March 18 - 20, 2009, Hong Kong.

[2] Chih-Fong Tsai and Chia-YingLin, "A triangle area based nearest neighbours approach to intrusion detection", Pattern Recognition 43 (2010) 222 – 229.

[3] Salvatore Ruggieri ,"Efficient C4.5", IEEE Transaction on knowledge and data engineering, Vol. 14, No. 2, March/April 2002.

[4] Jisu Oh and Shan Huang, "Spatial Outlier Detection and Implementation in WEKA", 2004

[5] Tien-Chin Wang and Hsien-Da Lee, "Constructing a Fuzzy Decision Tree by Integrating Fuzzy Sets and Entropy", 2005

[6] R. Shanmugavadivu and Dr.N.Nagarajan, "Network Intrusion Detection System Using Fuzzy Logic", Indian Journal of Computer Science and Engineering (IJCSE), 2011

[7] Chang-Tien Lu, Dechang Chen, and Yufeng Kou, "Algorithms for Spatial Outlier Detection", Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03) 0-7695-1978-4/03 $ 17.00 © 2003 IEEE.

[8] Anrong Xue, Lin Yao, Shiguang Ju, Weihe Chen, and Handa Ma, "Algorithm for Fast Spatial Outlier Detection" Young Computer Scientists, ICYCS 2008. The 9th International Conference, pp 1872-1877, 2008.

[9] Pamula.R, Deka.J.K, and Nandi.S, "An Outlier Detection Method Based on Clustering", Emerging Applications of Information Technology (EAIT), 2011 Second International Conference, pp 253-256, Feb. 2011.

[10] Fu Xiao and Xie Li, "Using Outlier Detection to Reduce False Positives in Intrusion Detection", Network and Parallel Computing, 2008. NPC 2008. IFIP International Conference, pp 26-33, Oct 2008.

[11] S. Shekhar, C.-T. Lu, and P. Zhang, "A Unified Approach to Spatial Outliers Detection", GeoInformatica, An International Journal on Advances of Computer Science for Geographic Information System, 7(2), June 2003.

[12] Su-Yun Wu and Ester Yen, "Data mining-based intrusion detectors", Expert systems with applications 36(2009) 5605-5612.

[13] Levent Bolelli, Seyda Ertekin, Ding Zhou and C. Lee Giles, "A Clustering Method For Web Data With Multi-Type Interrelated Components", May 8–12, 2007.

[14] A. Bordes, S. Ertekin, J. Weston, and L. Bottou, "Fast kernel classifiers with online and active learning", Journal of Machine Learning Research, 6:1579–1619, September 2005.

[15] Jaisankar N, Lalita S, Sruthi D and Swetha Balaji, "Intelligent Intrusion Detection System Using Fuzzy Based Efficient C4.5 Algorithm", Proceedings of the International Conference on Computer Science and Informatics (ICCSI), March 2012.