A Comparative Study of Machine Learning Methods for Generation of Digital Forensic Validated Data

Nandan Kumar
Department of Computer Science and Engineering
National Institute of Technology, Sikkim-737139, India
Email: nandankumar23598nk@gmail.com

Pankaj Kumar Keserwani
Department of Computer Science and Engineering
National Institute of Technology, Sikkim-737139, India
Email: pankaj.keserwani@gmail.com

Shefalika Ghosh Samaddar
Department of Computer Science and Engineering
National Institute of Technology, Sikkim-737139, India
Email: shefalika99@yahoo.com

***Abstract*—A number of machine learning algorithms are avail-able for detection of different types of network anomalies. How-ever, network anomalies vary in their requirement for detection and there is no general method or algorithm that is applicable to all types of anomalies which may occur in a network. The paper divulges the application of different available methods in comparison mode and analyses the rationale of the algorithms. Freely available databases are used for study and used as a benchmark for testing different methods of network anomalies. The benchmarks are used for testing accuracy. Such accuracy threshold indicates the required data validation. Validated data may be used forensic purpose. The data for Digital forensic must pass a threshold value from the test data. Determination of efficiency of methods is additional analytical aspect of the paper that is achieved in Python coding and in well formulated steps.The paper gives an overview of the process for finding network anomalies and data accuracy for digital forensic use.**

Keywords—Data Mining, Digital Forensic, Anomaly Detection, Intrusion Detection System (IDS), Dataset, Network, Vulnerability.

I. INTRODUCTION

Anomaly detection is one of the required parameters of distributable change detection in network system either at the host or at the service provider side. The parameter of anomaly detection is not very specific to the field of computer networking, rather it is a general parameter that can be applied depending upon the datasets to be used. The datasets can be related to climatic data, sequential anomaly patterns, a protection sequence, image anomalies in a collection of time related fMRI and many other including marketing and e-Commerce data. The training data set searches the benchmark data and detects noise data or outliers to divulge a valid pattern of possible breach or anomaly on the field of application. Such detection of anomalies when applied to networking, may provide a sound system of intrusion detection. Anomaly detection in networking has been tried using statistical method to imply vulnerability of intrusion methods. The classification mechanism using a number of data miningtools over a network traffic database is another familiar technique for detection of anomalies in network. A survey has been provided for different machine learning methods keeping aside the methods which has been obtained using certain data mining tools. Also deep learning, deep neural network methods foranalyzing network traffic classification are popular approach that is able to provide network flow prediction, probability prediction, occurrence of outliers, noise accumulation of a particular node etc. [1].

The paper is concerned with only one aspect of method for detection of network anomalies that is machine learning. The anomaly is detected based on the wide variety of change detection even if there is some tolerance of noise or outliers. There is anomaly in the dynamic stream of data as well. The methods of machine learning or knowledge intensive learning by its various applications and this notification for selection of machine learning algorithms, may be considered while deciding the other methods at anomaly and change detection

[2]. The validated datasets are used in various digital forensic applications such as medical forensic, marketing prediction intrusion forensic etc.

II. BACKGROUND STUDY AND RELATED LITERATURE SURVEY

Network anomaly can not be defined in a single word. Any kind of disturbance or attack or abbreviation of major for local, spatial outliers, other change detection or hindrance in traffic due to attack and intrusion, are all can be considered as network anomaly depending on these circumstances under which the network is transmitting the traffic from one peer to another. A number of methods are available to detect and dislocate the above anomalies by various specialised analytical measures, for example, SLOM provides a measure for local spatial outliers [3]. Even an active intrusion detection system (IDS) can be used to block suspected or doubtful attacks with full automation. The active and the passive intrusion detection system (IDS) may differ in the way of response to intrusion. An active IDS automatically limits the state of attacked system or immediately isolating it from the network using an identified proxy server assigned for the purpose. The other one, passive IDS detects the attack successfully and the attacking systems are remained under observation till they try to penetrate the system in anunauthorized manner. Thus, a passive IDS system is mostly categorized to monitor and analyse network traffic activity and only an alert warning to the operator may be provided including the potential

vulnerabilities and attacks. The role of passive IDS is only limited to identification of breaching and attack, preparation of log file for use in digital forensic system and a clear notification to the administrator about the observation in a formatted report. The inefficiency of passive IDS remain in their provision of manual intervention rather than to auto-handle the intrusion.

Similar to IDS in a system, there is Network Intrusion Detection System (NIDS). NIDS is usually connected to a sensor network or some network appliances that is able to provide aggregation information from the sensor network through a NIC operating in promiscuous mode and a separate management interface that is compatible with the sensor node such as IDS. IDS is put at some monitoring device in a network segment, at the boundary of autonomous system or traffic exchange mode (higher level router) on different network segments. NIDS has its able counterpart in the turn of a Host Intrusion Detection System (HIDS). Such implementation of software application is installed on workstations capable of monitoring hardware traffic for any anomaly. The agent monitors the operating system of the host and create log file and/or appropriate alarm through HIDS. HIDS can monitor the hosts on which it is installed and analyse each and every packet (incoming and outgoing) from nominated devices. The devices are selected on the basis of passive data analysis. The other peers having past records of intrusion attempts are usually observed for anomaly creation in a network.

A signature based IDS is equipped with a database having a previous attack signature and identified network vulnerabilities, such IDS is also known as knowledge based and behavioural based IDS. The separator database of signatures for different hosts connected to their network is monitored using the attacked footprints or intrusion specifications. Such specification or attack footprints are also included in the signature and such database can be used for tracing of identification and prevention of the similar attacks, when it occurs on an anomaly in network afterwards. Knowledge-based IDS successfully identifies known intrusion attempts. However, a behaviour based or anomaly based IDS system is capable of learning a pattern of normal system activity and distinguishing it from the intrusion attempts even if the outlier or the noise is very much near to the baseline categorized on the basis of the training of typical machine learning. Any deviation from the pattern cause an alarm to be triggered or accommodating such deviation into automatic database.

On the basis of the data captured over network traffic and identification of anomalies, there are various classification of anomalies over a network and the databases are designed in order to identify anomalies easily of a particular class. Broadly the anomalies are classified as point anomalies, context-based anomalies and collective anomalies. Point anomalies occur when individual data points differ in a significant manner to be noticed from the rest of the data points or normal data points. Point anomalies are easy to detect and datasets can be treated with software for visualization of such anomalies. The cause of point anomaly generation may not be a intrusion or attack always. Such anomalies can be generated even by a powersurge while creating the database, generating the data set from dynamic data flow between various nodes of a network or some error in identification of the proper format for the data to be classified.

Context based anomalies are typically context dependent. The data therefore, captured along with their context attributes and behavioural attributes. There can be some contradiction in classification between context based anomalies and point anomalies. Context based anomaly may have some outliers as normal attributes, which are beyond any consideration in case of point anomalies.

Collective anomalies are actually a collection of instances when treated against the entire dataset. Such data instances as a collection are treated as anomalies. A number of outliers explore the area of collective anomalies as the data instances are able to provide analytical results through various tools [4] [3] [5] [6].

A. Classification method and their comparative efficiency

There are a number of classification methods which are in the realm of data mining and can well be obtained by combining a number of classification and clustering techniques.

1)*Support Vector Machine (SVM):* Support Vector Machine (SVM) is a state of art algorithm based on both linear and nonlinear regression. Typically support vector machine (SVM) is able to act as an hyperplane classified based on linear separability. The best solution obtained for classifying the data set and identifying the best hyperplane is the requirement of this algorithm for classification. There are other development such as a Robust Support Vector Machine (RSVM) which are made robust to the presence of the anomalies through the tracing data set [7]. RSVM can even capable of ID [8]. SVM has inherent capability to work fast compared to neural network when applied in real datasets but any normal dataset having a mixture of normal anomaly data, the performance of both are comparable [9]. SVM in its various form can achieve higher accuracy than radial basis function neural network [10].

2)Regression Tree: A complex network data can be analysed through a suitable regression tree by applying robust analytical method. Regression tree can be idle with non linear relationship between defects, high order interaction and missing data values from a fixed data format. Regression tree is used when response variable is continuous in nature and is subject to some kind of quantization leading to its numeric value. The target variable usually does not belong to any class and a regression model treats the target variable using each of the independent variable. Such treatment leads to splitting data at several split points where the errors or anomalies may occur. At each of this split point the error between the predicted value and actual value is determined. A sum square error (SSE) is determined at the split point acrossthe variables. The split point errors are then compared and the split points or the variable having the lowest SSE is selected asroot node or split point. The processes continued recursively till there is network traffic running over the network and anomaly detection is required.

3)*Classification Tree:* Classification tree is mainly used for categorical data. It is used to separate the dataset into classes belonging to the response variable. If there is homogeneity of data then the dataset is split on the basis of the response variable leading to major classifications.

B. Cluster Analysis

Cluster analysis is partition of the datasets or instances into a number of subclasses depending on their similarities and deceived dissimilarities of various data instance collected for the training set and dynamically using it for target dataset. Clustering is an unsupervised learning that does not require any predefined classes. The process of cluster analysis involved dependency upon the similarity measure and dissimilarity nearness distance that can be defined on the basis of the dataset to be considered. A number of classes will be created from such measure and distance analysis. There are finite set of categories or cluster for which the dataset will belong to [11]. Even if no previous cluster exists a new cluster formulation may be obtained after dynamic rearrangement of data. Following techniques of clustering are prevalent:

1)*Partitional Clustering:* A partitional clustering algorithm divides a given set of data objects into a number of disjoint cluster such that each data instance will be contained in one and only one cluster. There are two types of partitioning methods: centroid based and medoid based.

In centroid based partitional clustering, each cluster is represented using the centroid or mean of the data instances. In medoid based clustering each cluster is represented by mean of an data instance that is closest to the mean value of the data set captured. The well-known k-mean algorithm is a centroid based algorithm, it partitions the datasets into k-subsets such that all points of a given a subset are closed or in the nearest neighbourhood to the centroid.

2)*Hierarchical Clustering:* A nested sequence of partition that are represented graphically by a dendrogram is provided by hierarchical clustering. Each node in the tree represents cluster except the leaf node. Each cluster in the tree other than the leaf node is the union of its sub clusters from all children nodes. The root of the tree is the clustering containing all the objects. Sometime a leaf node may contain a single object and is called singleton cluster. Hierarchical clustering are divided into two major sub categories: agglomerative method and divisive method.

Agglomerative method initiate cluster formulation in a bottom up approach starting with an separate/ individual object in a separate cluster and recursively tries to merge them until all the data instances captured belongs to the same cluster. The second method divisive method which initiate cluster identification by splitting up the data set into several small cluster in a top down approach until each cluster becomes a singleton cluster.

C. Neural Networks

Neural network is designed on the basis of behavioural techniques of human brain. Any artificial neural network consists of different neurons and synapses that are analyticallyesigned in a network system similar to the human brain function. Classification technique for anomaly detection is heavily dependent on neural network based algorithm to build classifiers in multi-class as well as one class setting neural network is used in multilayer perception, Neural Trees Auto-associative Networks, Adaptive Resonance Theory Based, Radial Basis Function Based, Hopeld Networks, Oscillatory Networks [12] neural trees are the other prevalent methods of neural network. Neural network operates in two steps:

1) A training on normal training data, learning different normal classes and

2) A testing instance for an input in a neural networks acceptance of the test input indicate that it is normal dataset and rejection indicate an anomaly. Replicator neural network have been used for one class anomaly detection [13]. A multilayer feed toward neural network is designed having the same number of input and output neurones corresponding to the features of the data. The training process typically involves compressing data into 3 hidden layers.

The testing process involves reconstruction of each data instance x_i using the already trained network to obtain reconstructed output o_i. The reconstruction error $_i$ for text input instance x_i is computed as:

$$\delta i = i/nPn\ j=1(xij - oij\ )\ 2$$

where n is the number of features over which the data is defined. The reconstruction error $_i$ is directly used as an anomaly score for the test instance.

Neural network works on the basis of the symmetry in the model and sometime, addition of bias helps to build up such symmetry. Each neurone is interlinked with other nodes or neurones and well devised weight is associated with each neurone. The neural behaviour is mimiced through the artificial perception behaviour.

Neural networks perform better than support vector machine under certain specified circumstances. For example, if the data instances are having a number of features or data is required to be classified into multiple classes; the performance is comparable.

Comparison between support vector machine and neural network have been made earlier [14] over different datasets. The other studies suggest how using the neural network based approach enhances the load forecasting for anomalies, load condition [15]. An architecture based neural network instead of using multilayer perceptron for finding the anomalies condition has been suggested as a solution. Fuzzy based approach in neural network has also been used [16] that increases the efficiency of model by using a differentfuzzy logic methods.

D. K-nearest Neighbourhood Method

K-nearest neighbourhood method is a proven anomaly detection technique. It is based on the assumption that normal data instances tries dense population neighbourhood while anomalies are outliers in the real sense [17]. The K-nearest

neighbourhood method can be divided into three broad categories:

1) The first one is distance based neighbour. Distance based neighbour differentiate anomalies from other based on the number of instances accumulated in the neighbour-hood [18].

2) The second method of distribution based method finds probabilistic data model using the given data in case of failure of these two methods. It is the data instance that is treated by an outlier [19].

3) The third category density based approach detect local anomalies based on local density of an object neighbour-hood [20].

K-nearest neighbourhood approach is applicable for both categorical and continuous attributes. Categorical attributes are used to obtain simple matching cofficient in order to find dissimilarity where as in case of continuous attributes distance between instances are used for separating normal instance from outliers.

The greatest disadvantage of nearest neighbourhood approach is that it is not scalable in case of increase in the number of attributes. The computational complexity $O(n^2)$ is considered high compared to other available methods. The one that is used for general purpose anomaly detection is relative density based approach. Varying density in dataset creates problem in detection of outlier in case of nearest neighbourhood algorithm. Linear time performance can be achieved when the data is in random order using a simple nested loop algorithm [21]. Outlier detection using in degree N (OODEN) algorithm is a graph based detection method which gives good result in small synthetic dataset [22].

E.Data for Digital Forensic Use

There are a number of dataset that can be used for digital forensic. The main categories are:

Organisational data

Network communication data including anomalies

Architectural data (infrastructure dependent)

Legal dataset of rule base

All of the above dataset must satisfy the criteria of evidence building, namely, accuracy, validity and dataset desired or generated must be verifiable. The present paper takes care of only one criteria of dataset for use in digital forensic i.e. accuracy.

III. Dataset Description in Use

Different anomaly detection techniques are suggested and analysed on the basis of either technique used or applicability over datasets. Anomaly in case of the network is detected in some points of outliers. These outliers are detected either by set based outlier detection technique or special set based outlier detection considering statistical distribution of attribute values with complete disregard for special relationship among items of multidimensional databases. The main idea of the problem is to consider the applicability of machine learning methods for detection of anomalies in network traffic and transmission. The problem make use of some well known database having inbuilt training sets that are collected over a period of time or built synthetically having log information and related methods as their basis. A test based comparison has been made to show the efficacy of the method considered in case of anomaly detection in network.

Any statistical problem over a spatial, analytical statics of attributes is generally avoided as the methods chosen are not stochastic in nature. The neighbourhood aggregate static value such as mean, standard deviation (SD) etc. are considered as a routine affair of the algorithm in their inbuilt capacity for computation of such values. Experimental layout is mainly based on the stepwise feature based derivation of results.

IV. Experimental Steps To Improve The Efficiency Of Different Algorithms

The parameters used in the dataset are duration, protocol-type, service, flag, source bytes, destination bytes, urgent packets etc.

Whole experimentation is done in Anaconda's python environment. Several python libraries are used such as Numpy for manipulating array, Pandas for handling large chunk of data etc. Throughout the process the sklearn library has been used to perform data mining tasks.

Steps of experimentation are provided below:

1) The dataset contains 41 different features such as:

"duration","src_bytes","dst_bytes", "land", "wrong_fragment","urgent", "hot", "num_failed_logins", "logged_in", "num_compromised","root_shell", "su_attempted","num_root", "num_file_creations", "num_shells", "num_access_files","num_outbound_cmds", "is_host_login","is_guest_login","count", "srv_count", "serror_rate", "srv_serror_rate","rerror_rate", "srv_rerror_rate","same_srv_rate", "diff_srv_rate","srv_diff_host_rate", "dst_host_count","dst_host_srv_count", "dst_host_same_srv_rate", "dst_host_diff_srv_rate", "dst_host_same_src_port_rate", "dst_host_srv_diff_host_rate", "dst_host_serror_rate", "dst_host_srv_serror_rate", "dst_host_rerror_rate", "dst_host_srv_rerror_rate"

2) KDDCup is the training dataset which has been used with Pandas, since Pandas can handle large chunk of data. Data can be received from any methods as read-csv, read-html or read-txt depending on file format.

3) A variable 'labeling' has been initialized to contain the target attribute data instances in float.

4) The data has been classified into two groups: normal or attack.

5) The data have been normalized and standardized so that the data range can be decided. and then the data has been pre-processed keep the model unbiased.

6) In the model clf, the parameters are specified depending on dataset and computation time has been noted to train the model.

7) Lastly, the test data has been run to predict the label through different model and the accuracy has been noted.

V. Results And Discussion

The algorithm considered are tested for their effectiveness of training time and accuracy test data available on link http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html. The algorithms are tested within the same time slot and evaluated in a comparative mode. Several such studies can be generated for evaluation of accuracy and detection mechanism of temporal outliers, spatial temporal outliers and other anomalies that occurs in a continuous data such as network traffic with a limited bandwidth. Such comparative study can be used to predict traffic volume distribution in a network over a network segment that is considered as neighbourhood for all practical purposes of applicability. The comparative study divulges datasets of high order accuracy to be used for the purpose of evidence building in case of digital forensic application. If a threshold 0:9000 of accuracy it is assigned for valid evidence then it may be interpreted that only three out of six methods can provide forensically validated data of high order accuracy. However, it may be noted that a digital forensic in a system has its more complicated avatar of big data forensic and/or cloud forensic. In that case, the total system design for aforesaid criteria should take an integrated approach rather than taking one criteria at a time and finding a suitable machine learning method of data churning having forensic elements in place. Fig. 1 indicates only a methodology; not an integrated approach. The limitation of the study is that all the methods considered are old enough to get a further lease of life so that it can be applied on big data. The experiment conducted is such that, it pertains to smaller dataset such as the given in this section. The methods considered can be selected for updation of appropriateness in a big data. The cost effectiveness of the methods will be realised only when a similar result of comparison can be generated for big data.

Classifier	Training Time	Accuracy	DF Evidence Data
K Nearest Neighbour	264.544	0.9252	Yes
SVM(rbf)	37705.71	0.7914	No
SVM(default)	46616.73	0.8999	No
Naive bayes	1.487	0.8758	No
Decision Tree	1.376	0.9294	Yes
Random Forest	2.196	0.9258	Yes

Fig. 1. Output of different classifiers on training time and efficiency with DF evidence on test data

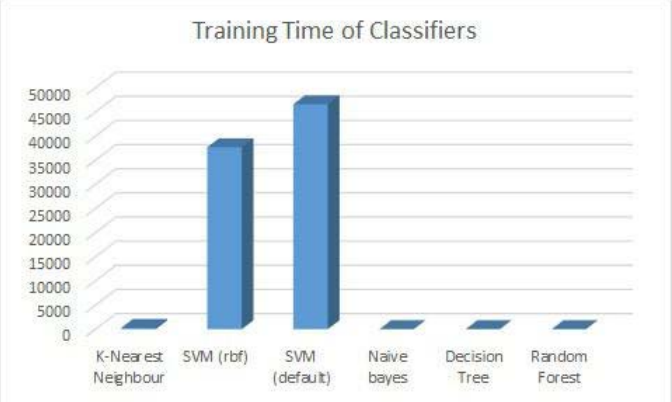

Fig. 2. Comparative Analysis of Classifiers for Training Time

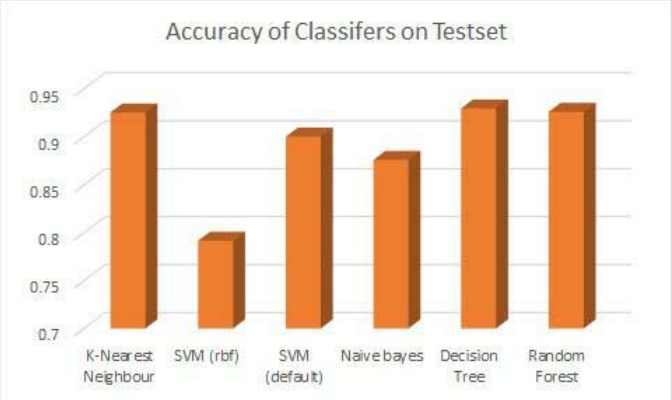

Fig. 3. Comparative Analysis of Classifiers efficiencies on test data

VI. Conclusion

The comparative study conducted is the beginning of a set of experimentation for detecting anomalies in a network that deals with big data. The focus of the paper is on neighbourhood outliers that are applicable in a network segment. The designs of method are compared and analysed for accuracy and the time taken for training the data models. Comparison of the performance is evident from the result and shows its effectiveness in small data segment. Such study satisfy neighbourhood outlier query processing, model reconstruction for specific data anomaly, random verification of test data, root outlier detection in a network segment etc. Data accuracy offorensic capability is obtained in a synthetic approach though not an integrated approach for evidence building in digital forensic. It gives a way for determination and testing of evidences satisfying a particular criteria.

The future direction of work will grow towards detection of appropriate method applicable in case of big data analysis.

Acknowledgment

The authors express deep sense of gratitude to Param Kanchenjunga High Performance Computing Centre, National Institute of Technology, Sikkim India, where the work has been carried out.

REFERENCES

[1] Z.Fadlullah, F. Tang, B. Mao, N. Kato, O. Akashi, T. Inoue, and K. Mizutani, "State-of-the-art deep learning: Evolving machine intelligence toward tomorrows intelligent network traffic control systems," IEEE Communications Surveys & Tutorials, 2017.

[2] J. W. Kappler, N. Roehm, and P. Marrack, "T cell tolerance by clonal elimination in the thymus," Cell, vol. 49, no. 2, pp. 273–280, 1987.

[3] P. Sun, S. Chawla, and B. Arunasalam, "Mining for outliers in sequential databases," in Proceedings of the 2006 SIAM International Conference on Data Mining. SIAM, 2006, pp. 94–105.

[4] P. D'haeseleer, S. Forrest, and P. Helman, "An immunological approach to change detection: Algorithms, analysis and implications," in Security and Privacy, 1996. Proceedings., 1996 IEEE Symposium on. IEEE, 1996, pp. 110–119.

[5] C. C. Noble and D. J. Cook, "Graph-based anomaly detection," in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003, pp. 631–636.

[6] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM computing surveys (CSUR), vol. 41, no. 3, p. 15, 2009.

[7] Q. Song, W. Hu, and W. Xie, "Robust support vector machine with bullet hole image classification," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 32, no. 4, pp. 440–448, 2002.

[8] W. Hu, Y. Liao, and V. R. Vemuri, "Robust anomaly detection using support vector machines," in Proceedings of the international conference on machine learning, 2003, pp. 282–289.

[9] S. Mukkamala, G. Janoski, and A. Sung, "Monitoring system security using neural networks and support vector machines," in Hybrid Information Systems. Springer, 2002, pp. 121–137.

[10] G. C. Tsang, P. P. Chan, D. S. Yeung, and E. C. Tsang, "Denial of service detection by support vector machines and radial-basis function neural network," in Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on, vol. 7. IEEE, 2004, pp. 4263–4268.

[11] B. Depaire, G. Wets, and K. Vanhoof, "Traffic accident segmentation by means of latent class clustering," Accident Analysis & Prevention, vol. 40, no. 4, pp. 1257–1266, 2008.

[12] M. Augusteijn and B. Folkert, "Neural network classification and novelty detection," International Journal of Remote Sensing, vol. 23, no. 14, pp. 2891–2902, 2002.

[13] S. Hawkins, H. He, G. Williams, and R. Baxter, "Outlier detection using replicator neural networks," in DaWaK, vol. 2454. Springer, 2002, pp. 170–180.

[14] S. Mukkamala, G. Janoski, and A. Sung, "Intrusion detection using neural networks and support vector machines," in Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on, vol. 2. IEEE, 2002, pp. 1702–1707.

[15] R. Lamedica, A. Prudenzi, M. Sforna, M. Caciotta, and V. O. Cencellli, "A neural network based technique for short-term forecasting of anomalous load periods," IEEE Transactions on Power Systems, vol. 11, no. 4, pp. 1749–1756, 1996.

[16] J. Asmuss and G. Lauks, "Network traffic classification for anomaly detection fuzzy clustering based approach," in Fuzzy Systems and Knowledge Discovery (FSKD), 2015 12th International Conference on. IEEE, 2015, pp. 313–318.

[17] Z. I. Ali, "A comprehensive survey of link mining and anomalies detection."

[18] T. Hu and S. Y. Sung, "Detecting pattern-based outliers," Pattern Recognition Letters, vol. 24, no. 16, pp. 3059–3068, 2003.

[19] M. Petrovskiy, "Outlier detection algorithms in data mining systems," Programming and Computer Software, vol. 29, no. 4, pp. 228–237, 2003.

[20] W. Jin, A. K. Tung, and J. Han, "Mining top-n local outliers in large databases," in Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2001, pp. 293–298.

[21] S. D. Bay and M. Schwabacher, "Mining distance-based outliers in near linear time with randomization and a simple pruning rule," in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003, pp. 29–38.

[22] V. Hautamaki, I. Karkkainen, and P. Franti, "Outlier detection using k-nearest neighbour graph," in Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, vol. 3. IEEE, 2004, pp. 430–433.