International Journal of Advanced Computer Science and Information Technology (IJACSIT)
Vol. 2, No. 3, 2013, Page: 17-24, ISSN: 2296-1739

www.elvedit.com

Outlier Detection in Stream Data by Machine Learning and Feature Selection Methods

Authors

Hossein Moradi Koupaie *moradyhsnm@yahoo.com*
Advanced Informatics School, Universiti Teknologi Malaysia *Kuala Lumpur, Malaysia*

Suhaimi Ibrahim *suhaimiibrahim@utm.my*
Advanced Informatics School, Universiti Teknologi Malaysia *Kuala Lumpur, Malaysia*

Javad Hosseinkhani *jhkhani@gmail.com*
Department of Computer Engineering/ Islamic Azad University, Zahedan Branch *Zahedan, Iran*

Abstract

In recent years, intrusion detection has emerged as an important technique for network security. Machine learning techniques have been applied to the field of intrusion detection. They can learn normal and anomalous patterns from training data and via Feature selection improving classification by searching for the subset of features which best classifies the training data to detect attacks on computer system. The quality of features directly affects the performance of classification. Many feature selection methods introduced to remove redundant and irrelevant features, because raw features may reduce accuracy or robustness of classification. Outlier detection in stream data is an important and active research issue in anomaly detection. Most of the existing outlier detection algorithms has less accurate because use some clustering method. Some data are so essential and secretary. Therefore, it needs to mine carefully even if spend cost. This paper presents a framework to detect outlier in stream data by machine learning method. Moreover, it is considered if data was high dimensional. This method is more accurate from other preferred models, because machine learning method is more accurate of other methods.

Key Words

Outlier Detection, Stream Data, Framework, Support Vector Machine.

I. INTRODUCTION

Outlier detection is the process of finding data objects with behaviors that are very different from expectation. Such objects are called outliers or anomalies. Outlier detection is important in many applications in addition to fraud detection such as medical care, public safety and security, industry damage detection, image processing, sensor/video network surveillance and intrusion detection [1, 2, and 10].

Intrusion detection systems (IDS) have become important security tools applied in many contemporary network environments. They gather and analyze information from various sources on hosts and networks in order to identify suspicious activities and generate alerts for an operator. The task of intrusion detection is often analyzed as a pattern recognition problem-an IDS has to tell normal from abnormal behavior. It is also of interest to further classify abnormal behavior in order to undertake adequate counter-measures. An IDS can be modeled in various ways. A model of this kind usually includes the representation algorithm (for representing incoming data in the space of selected features) and the classification algorithm (for mapping the feature vector representation of the incoming data to elements of a certain set of values, e.g. normal or abnormal etc.). Some IDS, like models presented in [20], also include the feature selection algorithm, which determines the features to be used by the representation algorithm. Even if the feature selection algorithm is not included in the model directly, it is always assumed that such an algorithm is run before the very intrusion detection process [19].

The quality of the feature selection algorithm is one of the most important factors that affect the effectiveness of an IDS. The goal of the algorithm is to determine the most relevant features of the incoming traffic, whose monitoring would ensure reliable detection of abnormal behavior. Since the effectiveness of the classification algorithm heavily depends on the number of features, it is of interest to minimize the cardinality of the set of selected features, without dropping potential indicators of abnormal behavior. Obviously, determining a good set of features is not an easy task. The most of the work in practice is still done manually and the feature selection algorithm depends too much on expert knowledge. Automatic feature selection for intrusion detection remains therefore a great research challenge [19].

There are many outlier directional methods in the literature and in practice. They categorize to 4 groups: Statistical Methods, Proximity-Based Methods, Clustering-Based Methods, and Classification methods. Statistical methods (also known as model-based methods) make assumptions of data normality. They assume that a statistical (stochastic) model generates normal data objects, and that data not following the model are outliers. Proximity-based methods assume that an object is an outlier if the nearest neighbors of the object are far away in feature space, that is, the proximity of the object to its neighbors significantly deviates from the proximity of most of the other objects to their neighbors in the same data set. Clustering-based methods assume that the normal data objects belong to large and dense clusters, whereas outliers belong to small or sparse clusters, or do not belong to any clusters. In Classification methods Outlier detection can be treated as a classification problem if a training data set with

class labels is available. The general idea of classification-based outlier detection methods is to train a classification model that can distinguish normal data from outliers [3, 5, and 7].

Classification method is more accurate than other method. There is some algorithm in classification method, but SVM method is better because although the training time of even the fastest SVMs can be extremely slow, they are highly accurate, owing to their ability to model complex nonlinear decision boundaries. They are much less prone to over fitting than other methods. The support vectors found also provide a compact description of the learned model. SVMs can be used for numeric prediction as well as classification [4, 6].

II. RELATED WORKS

Two main approaches to intrusion detection system are used namely misuse and anomaly detection. Misuse detection is based on a description of known malicious activities. This description is often modeled as a set of rules referred to as attack signatures. An anomaly detection IDS looks for anomalies, meaning it thinks outside of the ordinary. It uses rules or predefined concepts about “normal” and “abnormal” system activity (called heuristics) to distinguish anomalies from normal system behavior and to monitor report on, or block anomalies as they occur.

Various artificial intelligence techniques have been utilized in IDS. Machine learning as an efficient technique has an inherent capacity to provide decision aids for the analysts and which automatically generate rules to be used for computer network intrusion detection. Feature selection, is a preprocessing step to machine learning of selecting a subset of relevant features for building robust learning models.

To enhance the learning capabilities and reduce the computational intensity of competitive learning, different feature reduction techniques based on Machine Learning have been proposed. We compared three feature reduction techniques based on DT, FNT and PSO on KDD99 dataset.

A feature selection method finds smallest number of features that maximize the performance of the pattern recognition system. There are three categories of feature selection methods, depending on how they interact with the classifier: the wrapper, the filter and hybrid models [21], [22]. The wrapper model uses learning algorithms performance in assessing and selecting features [21], [22]. The filter model considers statistical characteristics of a data set directly without involving any learning algorithm [21], [22]. The wrapper and the filter models have their own advantages and disadvantages. The features selected by applying the wrapper model are usually better adapted to a machine learning algorithm, which is chosen in advance for the feature selection process. However, the wrapper model also requires more computational resources than the filter model. When the number of features becomes very large, such as intrusion detection systems, the filter model is more appropriate due to its computational efficiency. In order to combine the advantages of both models, the hybrid model was proposed

[21], [22]. In this research, we proposed a framework to outlier detection in stream data by classification method. Classification method is more accurate that other methods.

III. PROPOSED FRAMEWORK

In proposed framework at first it is received stream data to Feature selection segment. This segment considers data. If stream data is high dimensional, this segment select important attributed from them and send to classification stream data segment. If data is not high dimensional, it send data to classification stream data segment without doing Feature selection algorithm on them. And, in next segment classify data by improved incremental SVM algorithm. This algorithm supports two linearly separable and linearly inseparable stream data. At last in extract outlier class segment. It is extracted outlier. Because stream data come continually, this segment every period time gives a report about outliers.

A. Feature Selection Segment

There are five fetcher selection methods to apply:

Stepwise forward selection: The procedure starts with an empty set of attributes as the reduced set. The best of the original attributes is determined and added to the reduced set. At each subsequent iteration or step, the best of the remaining original attributes is added to the set [11].

Stepwise backward elimination: The procedure starts with the full set of attributes. At each step, it removes the worst attribute remaining in the set [11, 12].

Combination of forward selection and backward elimination: The stepwise forward selection and backward elimination methods can be combined so that, at each step, the procedure selects the best attribute and removes the worst from among the remaining attributes [13].

Decision tree induction: Decision tree algorithms were originally intended for classification. Decision tree induction constructs a flowchart like structure where each internal (nonleaf) node denotes a test on an attribute, each branch corresponds to an outcome of the test, and each external (leaf) node denotes a class prediction. At each node, the algorithm chooses the “best” attribute to partition the data into individual classes. When decision tree induction is used for attribute subset selection, a tree is constructed from the given data. All attributes that do not appear in the tree are assumed to be irrelevant. The set of attributes appearing in the tree form the reduced subset of attributes [13].

Because method of Decision tree induction is more accurate of other methods, it is used of this method for fetcher selection.

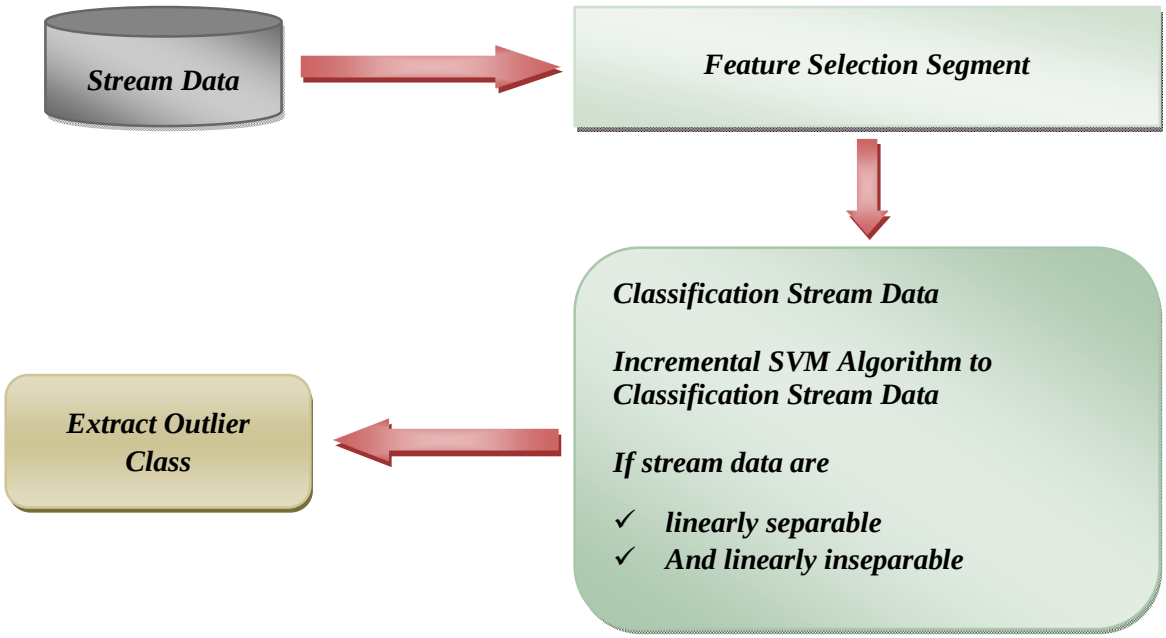

FIGURE 1: FRAMEWORK FOR OUTLIER DETECTION IN STREAM DATA

B. Classification Stream Data by Incremental SVM

A method of ordering linear and nonlinear data is Support vector machines (SVMs). In a case, SVM is an algorithm and the function of it is as follows. To change the original training data into a higher dimension, it applies a nonlinear mapping. It seeks for the linear ideal separating hyper plane through this new dimension. A hyper plane can always separate the data into two classes with a suitable nonlinear mapping to an appropriately high dimension. The SVM discovers this hyper plane utilizing support vectors that is “essential” training tuples and margins which is explained by the support vectors. “I’ve heard that SVMs have attracted a great deal of attention lately. Why?” Vladimir Vapnik and colleagues Isabelle Guyon and Bernhard Boser (1992) have done the first research on support vector machines since the groundwork for SVMs has been around since the 1960s.

Even though the training time of SVMs is very extremely slow, they are very precise and can to model compound nonlinear decision limitations. In compare to other methods, they are much less predisposed to over fitting. The provision vectors also are a compressed explanation of the trained model. SVMs also are able to utilize for numeric calculation along with classification. They have been used for many areas such as object recognition, handwritten digit recognition, and speaker identification.[14,15,16].

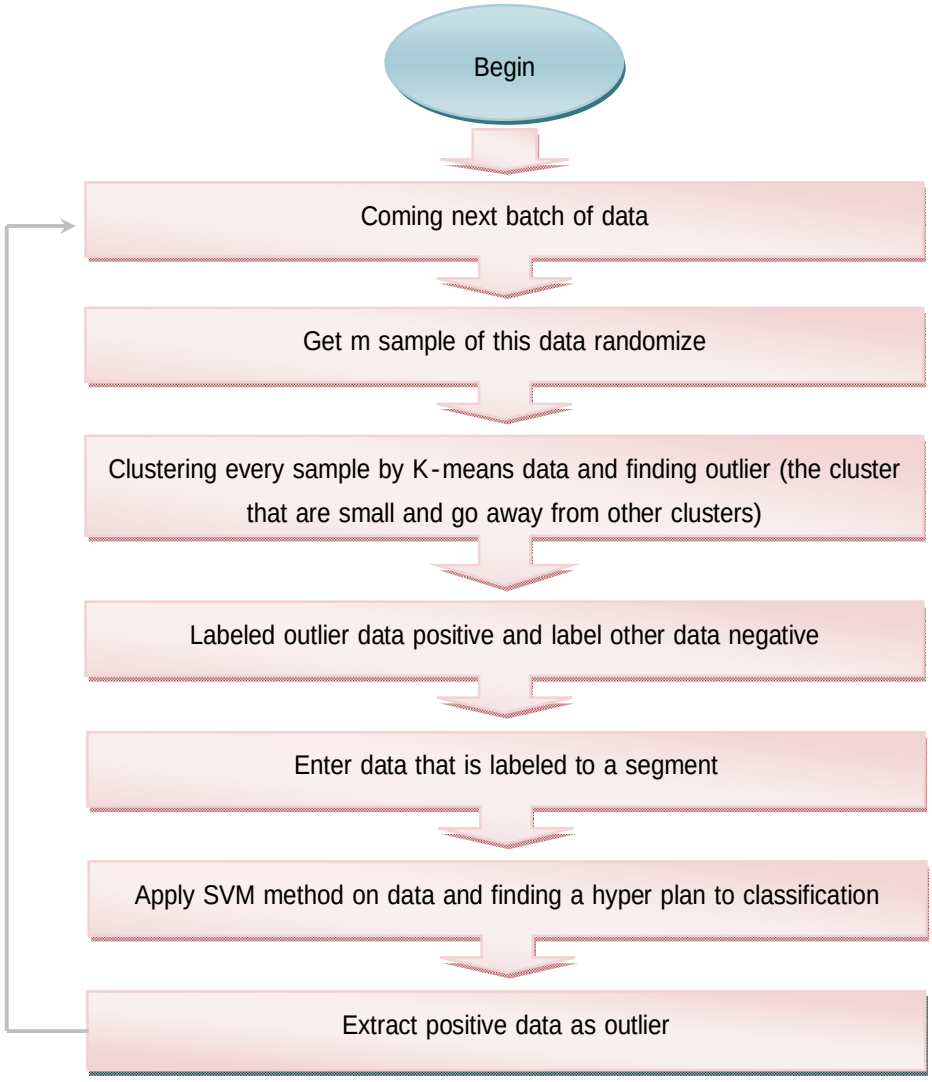

FIGURE 2: ALGORITHMS FOR OUTLIER DETECTION IN STREAM DATA

In this study, it is proposed a method to finding outlier in stream data by incremental support vector machine. The figure 2 shows this method. At first data enter in some batches sequentially. Some sample select randomly of every batches data and clustering by k-mean method. Some cluster is small and go away from other clusters are outliers. Outlier labeled by positive and other data labeled by negative. In all of sample do this method. And all of data that are labeled collect to a segment. Now on these data apply SVM method to finding suitable hyper plane for classification. Next extract data that are positive label as outlier. This process applies for next batch data that is entered.

IV. CONCLUSION

Outlier detection is the process of finding data objects with behaviors that are very different from expectation. Such objects are called outliers or anomalies. Outlier detection is important in many applications in addition to fraud detection such as medical care, public safety and security, industry damage detection, image processing, sensor/video network surveillance and intrusion detection. Intrusion detection systems (IDS) have become important security tools applied in many contemporary network environments. They gather and analyze information from various sources on hosts and networks in order to identify suspicious activities and generate alerts for an operator.

We have proposed a framework to outlier detection in stream data by classification method. Classification method is more accurate that other methods. But it has more cost than other method. Detecting outlier in stream data when data are important is so essential. In future work we can optimize incremental SVM algorithm and prefer an effective feature selection algorithm for more accurate and better time.

REFERENCES

[1] Jin, W., Tung, K.H. and Han,J., (2001). Mining top-*n* local outliers in large databases.In *Proc.2001 ACMSIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'01)*, pp. 293–298,San Fransisco, CA, Aug. 2001.

[2] Babu, S. and Widom,J., (2001). Continuous queries over data streams.*SIGMOD Record*, 30:109–120..

[3] Charu C. Aggarwal, Philip S. Yu, (2001). Outlier detection for high dimensional data, Proc. of the 2001 ACM SIGMOD int. conf. on Management of data, p.37-46, May 21-24, 2001, Santa Barbara, California, United States

[4] Babcock, B., Babu, S. , Datar, M. , Motwani, R., & Widom,J. (2002). Models and issues in data stream systems. In *Proc. 2002 ACM Symp. Principles of Database Systems (PODS'02)*, pages 1–16, Madison, WI, June 2002.

[5] Gibbons, P.B., & Matias, Y.(1998). New sampling-based summary statistics for improving approximate query answers. In *Proc. 1998 ACM-SIGMOD Int. Conf.ManagementofData (SIGMOD'98)*, pages 331–342, Seattle,WA, June 1998.

[6] Knorr, E., & Ng, R, (1997). A unified notion of outliers: Properties and computation. In *Proc. 1997 Int. Conf. Knowledge Discovery and Data Mining (KDD'97)*, pp. 219–222, Newport Beach, CA, Aug. 1997.

[7] Chandola, V., Banerjee, A., & Kumar, V., (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41:1–58.

[8] Chandrasekaran, S., & Franklin, M., (2002). Streaming queries over streaming data.In *Proc.2002 Int. Conf. Very Large Data Bases (VLDB'02)*, pages 203–214, Hong Kong, China, Aug. 2002.

[9] Babcock, B., Babu, S., Datar, M., Motwani, R. & Widom, J. (2002). Models and issues in data stream systems. In *Proc. 2002 ACM Symp. Principles of Database Systems (PODS'02)*, pages 1–16, Madison, WI, June 2002.

[10] Muthukrishnan, S. (2003). Data streams: algorithms and applications. In *Proc. 2003 Annual ACM-SIAM Symp. Discrete Algorithms (SODA'03)*, pages 413–413, Baltimore, MD,Jan. 2003.

[11] R. Kohavi, G.H. John. (1997). Wrappers for feature subset selection, Artificial Intelligence 97 (1–2), 273–324.

[12] R.S. Kulkarni, G. Lugosi, V.S. Santosh. (1998). Learning pattern classification—a survey, IEEE Transaction on Information Theory 44 (6), 2178–2206.

[13] R.S. Kulkarni, M. Vidyasagar. (1997). Learning decision rules for pattern classification under a family of probability measures, IEEE Transactions on Information Theory 43 (1) 154–166.

[14] Liang, Z., Lia, Y. (2009). Incremental support vector machine learning in the primal and applications. Neurocomputing 72(10-12), 2249–2258.

[15] Zheng, J., Yu, H., Shen, F., Zhao, J. (2010). An Online Incremental Learning Support Vector Machine for Large-scale Data. In: Diamantaras, K., Duch, W., Iliadis, L.S. (eds.) ICANN 2010. LNCS, vol. 6353, pp. 76–81. Springer, Heidelberg.

[16] Liu, X., Zhang, G., Zhan, Y., Zhu, E. (2008). An Incremental Feature Learning Algorithm Based on Least Square Support Vector Machine. In: Preparata, F.P., Wu, X., Yin, J. (eds.) FAW 2008. LNCS, vol. 5059, pp. 330–338. Springer, Heidelberg.

[17] Vapnik, V. (1999). The nature of statistical learning theory. Springer, New York.

[18] Ruping, S. (2002): Incremental learning with support vector machines. Technical Report TR-18, Universitat Dortmund, SFB475.

[19] Nguyen, H., Franke, K., & Petrovic, S. (2010, February). Improving effectiveness of intrusion detection by correlation feature selection. In*Availability, Reliability, and Security, 2010. ARES'10 International Conference on* (pp. 17-24). IEEE.

[20] G. Gu, P. Fogla, D. Dagon, W. Lee, and B. Skoric. (2006). Towards an information-theoretic framework for analyzing intrusion detection systems. In Proceedings of the 11th European Symposium on Research in Computer Security (ESORICS'06), September.

[21] I. Guyon, S. Gunn, M. Nikravesh and L.A. Zadeh. (2005). Feature Extraction: Foundations and Applications. Series Studies in Fuzziness and Soft Computing, Springer.

[22] H. Liu, H. Motoda. (2008). Computational Methods of Feature Selection.
Chapman & Hall/CRC.