Cluster Computing (2019) 22:399–408
https://doi.org/10.1007/s10586-018-2835-2

Strategies for data stream mining method applied in anomaly detection

Ruxia Sun[1] · Sun Zhang[1] · Chunyong Yin[1] · Jin Wang[2,3] · Seungwook Min[4]

Received: 5 September 2017 / Revised: 4 May 2018 / Accepted: 10 August 2018 / Published online: 25 August 2018

Abstract
Anomaly detection, which is a method of intrusion detection, detects anomaly behaviors and protects network security. Data mining technology has been integrated to improve the performance of anomaly detection and some algorithms have been improved for anomaly detection field. We think that most data mining algorithms are analyzed on static data sets and ignore the influence of dynamic data streams. Data stream is the potentially unbounded, ordered sequence of data objects which arrive over time. The entire data objects cannot be stored and they need to be handled in one-time scanning. The data distribution of data stream may change over time and this phenomenon is called concept drift. The properties of data stream make analysis method different from the method based on data set and the analysis model is required to be updated immediately when concept drift occurs. In this paper, we summarize the characteristics of data stream, compare the difference between data stream and data set, discuss the problems of data stream mining and propose some corresponding strategies.

Keywords Anomaly detection · Data stream · Clustering · Concept drift

1 Introduction

Intrusion detection is proposed by Lee [1] in 1998 and provides important protection for network security. In general, intrusion detection is categorized into two types: misuse detection and anomaly detection [2]. The former recognizes the pattern of known attack behaviors and establishes the rule base. The attack behavior is detected if it matches the rules. This method can detect known attack types efficiently, but it cannot detect unknown attack behaviors that result in lower detection rate. The latter learns the pattern of normal behaviors and considers anomaly behavior deviates from normal pattern. It has the ability of detecting unknown attack behaviors, but it could cause higher false alarm rate.

The integration of data mining technology solves the problem of higher false alarm rate [3]. It can mine the potential patterns of normal behaviors from samples and train detecting model automatically. The methods of data mining can be categorized into several types by diverse intentions, such as clustering method, classification method and regression method [4]. The combination of data mining and intrusion detection has made great progress. Wang [5] proposes the approach based on artificial neuron network (ANN) and fuzzy clustering to solve the problem of low detection precision in the respect of low-frequent attacks. Lin [6] proposes the improved k-nearest neighbor (KNN) combined with cluster centers. The experimental results show it performs better than or similar to KNN and support vector machines (SVM) [7–10]. Hoz [11] proposes the multi-objective approach for feature selection and applies it to self-organizing maps (SOM) [12].

✉ Seungwook Min
swmin@smu.ac.kr

[1] School of Computer and Software, Jiangsu Engineering Center of Network Monitoring, Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, Nanjing, China

[2] Key Lab of Broadband Wireless Communication and Sensor Network Technology (Nanjing University of Posts and Telecommunications), Ministry of Education, Nanjing, China

[3] College of Information Engineering, Yangzhou University, Yangzhou, China

[4] Department of Computer Science, Sangmyung University, Seoul, Korea

Most anomaly detection methods are based on static data sets and they ignore the influence of dynamic data stream. Data set is static and the analysis model is established after scanning entire data set for several times. The definition of data stream is different, but commonly data stream can be thought as the sequence of infinite data objects that make it impossible to store entire data objects. The treatment of data stream should be fast enough to prevent the loss of critical data which is a challenge for designing algorithm. The model needs to be updated when the distribution of data object changes. Therefore, the detection of data distribution changing is the other challenge for data stream mining [13–15].

Main applications of data stream mining can be categorized into three fields: data stream clustering, data stream classification and frequent pattern mining [16–18]. There are some classical researches about data stream mining. Besides, the classification methods can be divided into two types according to the number of classifiers: single classifier learning and ensemble learning. Ensemble learning has been proved to be an efficient method of improving predictive accuracy or/and decomposing a complex, difficult learning problem into easier sub-problems [19]. Very fast decision tree (VFDT) [20] is proposed in 2000 for data stream classification and it constructs decision tree based on Hoeffding inequality. Han [21] proposes FP-tree for frequent pattern mining. It directly stores the scan results of the database into frequent pattern tree rather than using candidate itemset. Czarnowski [22] proposes ensemble online classifier which concerns mining data stream with concept drift and the one-class base classifiers can be updated by incoming chunks of data objects.

Data stream clustering methods are mostly transplanted from clustering methods based on static data set. Clustering methods based on data set can be categorized as partitional methods, density-based methods and grid-based methods [23]. Guha [24] proposes the improved data stream clustering algorithm STREAM based on k-means in 2003. In the same year, Aggarwal [25] proposes the framework CluStream for evolving data stream. They firstly propose that data stream should be treated as infinite data objects and the processing over whole data stream is not appropriate. D-Stream [26] is proposed by Chen in 2007 and it is improved according to the density and grid based algorithm which is applied for high dimensional data stream.

There are two other techniques which are necessary in data stream mining: concept drift detection and sliding window technique. The phenomenon of conceptual drift is inevitable in data stream and its detection is vital for updating the model. The detection of concept drift can be a tough problem, because the type of concept drift is diverse. Sergio [27] proposes there are six types of drifts with respect to the ratio of changes. Sliding window is considered as one of the basic technologies which can solve the problems of discovering knowledge in dynamic data stream. It solves the problems of knowledge discovery in potential infinite data stream by superimposed processing of data windows with finite capacity. It also helps data mining methods in static data sets transplanted to dynamic data stream.

The application of data stream mining in anomaly detection is proposed by Oh [28] in 2005. They exploit data stream clustering method and model various statistics of objects as profile to improve the performance of anomaly detection. The clusters can be split and merged to fit the change of data stream. Big data stream also become new branch of data mining and it is based on data stream mining. Guerrieri et al. proposes a distributed data stream mining algorithm DS-means [29]. The mining work consists of three steps: local clustering, model transmission and global clustering which is a typical hierarchical distributed data stream mining framework.

The main contributions of this paper are the following:

1. We introduce the research status of data stream classification, clustering and concept drift. The conventional methods of data stream mining are reviewed and discussed.
2. We summarize the characteristics of data stream and discuss the differences between dynamic data stream and static data set. Based on the comparison, the problems existing in data stream mining are discussed and some corresponding strategies are proposed.
3. The improved anomaly detection model based on data stream mining is introduced in the end of the paper. We briefly introduce the idea of improved model and the experimental performance could be found in our previous studies [30, 31].

The remainder of the paper is organized as five sections. Several main technologies are reviewed in Sect. 2. In Sect. 3, the characteristics of data stream are discussed. Some problems of data stream mining are discussed and strategies are proposed. An improved anomaly detection model based on data stream mining is proposed in Sect. 4. Section 5 concludes the paper and proposes the plans for future research.

2 Related work

In this section, we will review main technology of data stream mining. Data stream classification and clustering methods are fundamental technologies in data stream mining and concept drift detection play an important role in updating classification and clustering model. Therefore, we will review the research works of data stream from

three aspects. The classification methods can be categorized as single classifier learning and ensemble learning. Similarly, the clustering methods also can be divided into several types.

2.1 Data stream classification

Single classifier learning method maintains and incrementally updates single classifier. It can response to concept drift and the classifier can select from ANN, SVM and decision tree. Because of the characteristics of data stream, incremental updating is the main method to solve the change of data stream. Incremental updating refers to updating the model after the processing of instances which is one by one (or batch by batch) sequentially. Ensemble learning method exploits several basic classifiers and the updating method of this classifier is easier than single classifier. Considering the influence of concept drift, the performances of ensemble learning method are better than single classifier learning method. Ensemble learning method is more efficient for being extended and parallelized. It can be rapidly adapted to concept drift by pruning and obtain more accurate description of the concept. Moreover, the training speed of basic classifier is faster than that of single classifier and it is more suitable for dealing with high-speed data stream.

Decision tree models are widely applied to construct classifiers for processing data stream, because decision tree model is similar to human reasoning and is easy to understand. The decision tree based on Hoeffding inequality is most popular for data stream classification, such as VFDT [20], Concept-adapting Very Fast Decision Tree (CVFDT) [32], VFDTc [33].

VFDT is a method of constructing decision trees based on Hoeffding bound for data stream mining environment. It is generated by constantly replacing leaf nodes with branch nodes. Each node retains an important statistic and it accepts the splitting test when the statistics of the node reach a threshold. The most important innovation is to apply Hoeffding inequality to determine the number of samples and the splitting attributes which are need to split leave nodes. The algorithm only needs to scan the data stream once, so it has high temporal and spatial efficiency, and the performance of the classifier is similar to that of the traditional algorithm. However, VFDT ignores the influence of concept drift. CVFDT is the extension of VFDT and it solves the problem of concept drift. The core idea is to replace the historical subtree with new subtree when new subtree is more accurate. It maintains a sliding training window and updates the generated decision trees when the sample flows into or out of the window, keeping generated decision trees adapt to the distribution of samples in the training window.

Ensemble learning makes the classifier more accurate and adapt to the change of concept [34, 35]. It reduces the impact of concept drift by using decision combination function as Fig. 1 shows. The commonly used ensemble methods are boosting and bagging. Bagging is to generate T subset by randomly select from original data set T times and the subset has same size with original data set. Then it trains T base classifiers and combines them as an ensemble classifier. Boosting also obtains several classifiers by resampling from original data set, and finally constructs an ensemble classifier. The difference is that bagging combines classifiers with weights.

There are two approaches to design ensemble classifiers: coverage optimization and decision optimization [19]. The former focuses on the generation of a set of mutually complementary classifiers and they are combined to achieve optimal accuracy with a fixed decision combination function. The latter focuses on designing and training an appropriate decision combination function and the classifiers are given in advance. The selection of classifiers is also an important problem. Both of high diversity and accuracy should be considered.

2.2 Data stream clustering

Clustering method can be considered as the process of dividing data sets into subsets which consist of several similar objects and these subsets are called clusters or classes [36]. Clustering method aims to make objects in same cluster as similar as possible and different from those of other clusters. Since the data objects in clustering do not need labels, clustering method is an unsupervised learning process. Clustering analysis is an important and basic method of data mining. It can be applied to analyze the distribution of the data by dividing data objects into

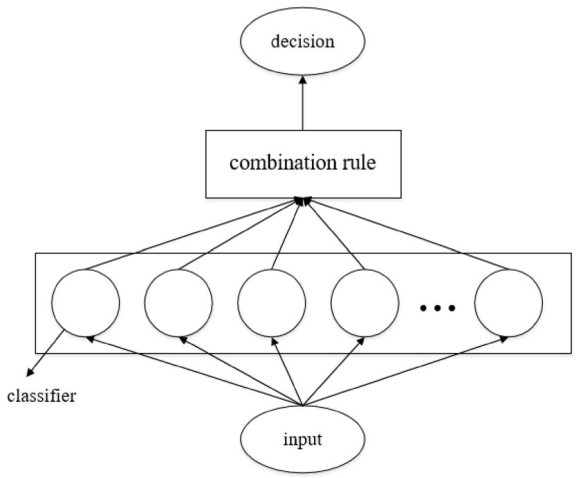

Fig. 1 The diagram of ensemble learning

different clusters and the characteristics of each cluster can be observed. Clustering can also be treated as a preprocessing technique for other data processing methods, such as generating category labels, providing support for classification, extracting features to support correlation analysis, mining frequent item, detecting outliers.

The characteristics of data stream make the data labeling very costly that new arrived data cannot be applied for training classifier. Clustering method is a kind of unsupervised method and it does not require the labeling recourse. Therefore, clustering method is appropriate for data stream and it can be exploited as the preprocessing of classification method. The application of clustering method in anomaly detection is based on two hypotheses:

(1) The number of normal behaviors is much greater than that of abnormal behaviors;
(2) There are obvious differences between normal and abnormal behaviors.

Data stream is infinite and temporal that requires the adaptive updating of analysis model. Incremental learning method means the model can learn knowledge from increasing data and it is thought to solve the problem of concept drift. However, incremental learning method is not appropriate for data stream, as it analyzes each incoming data object that does not meet the need of processing speed in data stream. The technology of data window is utilized in batch learning method. It exploits finite window to analyze infinite data stream. The data objects in data window can be treated as static data set and the samples of data stream. In this way, traditional data mining methods can be transplanted from static data set into dynamic data stream.

Most clustering methods for data stream are based on traditional clustering method of data set. The classical data stream clustering methods include STREAM [24], CluStream [25], DenStream [37], E-Stream [38] and D-Stream [26]. STREAM, which is based on K-means algorithm, is proposed by Guha in 2003. They employ the idea of batch learning method and the summary statistics are stored to represent generated clusters. The processing size of data objects is limited to meet the need of memory and the updating algorithm is shown in Fig. 2.

However, STREAM algorithm executes clustering process in entire data stream and it does not consider the impact of concept drift. Aggarwal proposes CluStream algorithm in 2003 which is based on STREAM algorithm. They make a breakthrough opinion that data stream is an infinite process and newer data is more valuable for data stream. The process of CluStream has two stages: online stage and offline stage. In online stage, generated clusters are updated by incoming data objects to fit the change of data stream. In offline stage, historical clusters saved in pyramidal time frame are clustered according to query requests.

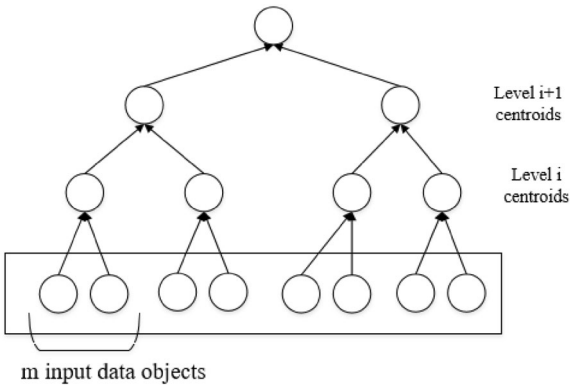

Fig. 2 The diagram of STREAM algorithm

STREAM and CluStream are based on K-means algorithm which belongs to partitional method. Den-Stream and D-Stream belong to density-based method. Cao proposes Den-Stream in 2007 aimed for eliminating the effects of noise and DBSCAN algorithm is applied in offline stage. D-Stream algorithm is proposed by Chen in 2007 and it is based on density and grid. High dimension data object is mapped into grid with low dimension and the outliers are detected by grid density.

2.3 Concept drift

One of the characteristic in data stream is that potential distribution of data stream may change over time and the concept reflected by data distribution also changes that this phenomenon is known as concept drift. Especially in real life data stream, concept drift may happen in these situations [19]:

(1) In computer or telecommunication systems, attack behaviors or abnormal behaviors could make concept drift happens and behavior pattern changes;
(2) Traffic patterns may change over time in traffic monitoring system;
(3) The concept may change in weather prediction system that means climate change or natural disasters.

To tackle the influence of concept drift, the analysis model should be able to detect the change of concept, and quickly adjust the model according to incoming data objects. Concept drift cannot be predicted, but it could be observed in the context.

Given target variable y and condition variable X, data instance can be denoted as (X, y). In time t, every instance is generated from data source with a joint probability distribution $P^t(X, y)$. Concept drift can be observed [39] in t_1

when $P^{t_0}(X, y) \neq P^{t_1}(X, y)$. Besides, when concept drift occurs, either one of or all the following changes: prior probabilities of classes $P(y)$, conditional probabilities of classes $P(X|y)$ and posterior probabilities of classes $P(y|X)$.

Concept drift is distinguished as two types: real concept drift and virtual concept drift. Real concept drift means $P(y|x)$ will change not matter whether the change of $P(x)$ and it can affect the decision boundary. Virtual concept drift means $P(y|x)$ will not change, although $P(x)$ has changed. The decision boundaries will not be affected by virtual concept drift, and analysis model does not need to be updated. However, virtual concept drift should also be detected. The visualization of two types is show in Fig. 3.

In the paper [27], Sergio distinguishes concept drift as six types considering its rapidness, including sudden, gradual, incremental, recurring, blips and noise. They also propose three solutions for learning from data stream with concept drift:

(1) The analysis model should be retrained for every time a new instance or chunk arrives;
(2) The analysis model will be retrained when the degree of detected concept drift is significant. In this way, the concept drift detector can tolerate noise;
(3) The analysis model can update itself adaptively and follow the shifts and drifts of data stream.

Sliding window is applied to store finite incoming data objects from data stream and it also can be utilized to detect concept drift by compare the distribution in historical and new sliding window. However, the size of sliding window is a crucial issue for its performance. A small sliding window could detect small and rapid changes, but it may cause the problem of overfitting. Conversely, large sliding window could store more information, but it may ignore small and rapid changes. Therefore, the best choice is to adaptatively modify the size of sliding window. For example, the window size could be increased if concept drift does not occur, but it will be narrowed when concept drift is detected.

3 Analysis and discussion

In this section, we will analyze data stream and compare it with static data set. The characteristics of data stream determine that mining method for data stream is different from that of static data set. Besides, we will also discuss the problems of data mining algorithm and give several strategies.

3.1 Analysis of data stream

The definition of data stream is described in several papers [19, 24, 36] and it can be summarized as: a data stream is the potentially unbounded, ordered sequence of data objects which arrive over time.

Definition 1 Given data stream S which consists of data object o, it can be denoted as $S = (o_1, o_2, \ldots, o_h)$. Each data object o_i is consist of m features and o_i^m represents the value of m-th feature.

Definition 2 Data window is a common technique for data stream mining and data objects will be divided into different data window according to its arrival order. Data window is denoted as $B_1, B_2, \ldots, B_n$ and the number of data objects in B_i is denoted as N_i. Through the finite size of data window, data object can be treated as data set and exploit traditional mining method.

Data window can be distinguished as three types: landmark window, sliding window and damped window. Landmark window considers the entire data stream and can obtain global frequent pattern by analyzing whole historical data. Sliding window focuses on recent transactions and it is introduced in Sect. 2.3. It is easy for understanding and designing that makes it applied widely in data stream mining. In damped window model, each transaction has a corresponding weight, and the weight will increase with time. Therefore, it can control these weights to store or delete relevant historical data information.

Definition 3 Synopsis data structure plays a crucial role in data stream clustering algorithm and it is employed to store the summary statistics information of micro clusters. The cluster C is simply denoted as $C(\delta, \mu, SS)$. δ is the number of data objects in the cluster. μ represents the

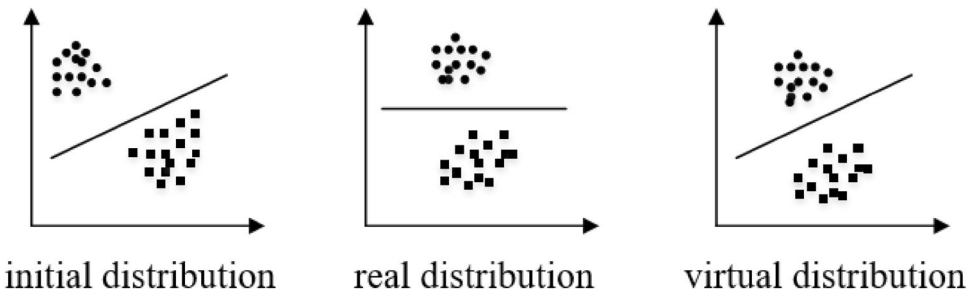

Fig. 3 Two types of concept drift

center of cluster and *SS* is the quadratic sum of data objects in the cluster.

In the paper [31], we summarize and compare the difference between static data set and dynamic data stream mining. Data stream have several characteristics compared with data set:

(1) Ordering. The data objects in data stream are generated in chronological order and the ordinal numbers are implicit at the arrival time or are recorded directly with the time stamp.
(2) Non-reproducibility. Once the data object in the data stream is scanned by processing node, it will not appear again unless extra storage is utilized to save the information.
(3) High speed. The data objects are generated at high speed and it requires fast processing speed of mining algorithm.
(4) Infinity. The data objects in data stream are continuously generated and data stream can be treated as infinite process.
(5) High dimension. The data objects in data stream have a large number of features and some of that can be redundant. Efficient dimension reduction method will be helpful and traditional methods for data set also can be transplanted into data stream, such as the localization of linear discrimination analysis (LDA) in the paper [40].
(6) Dynamic. The probability distribution of data objects in data stream changes over time and the change is out of control. It requires analysis model should be updated to fit the distribution changes.
(7) Costly labeling resource. The data labeling can be costly and it may be not immediate. Sometimes, it is not possible to determine the label. Therefore, clustering method can be applied as the preprocessing of other mining methods to analyze the distribution of data stream.

The characteristics of data stream make its analysis model different compared with data set and the analysis model should satisfy several requirements:

(1) The design of mining algorithm is necessary and it should have lower time and space complexity. Considering the limited storage capacity of devices and the infinity of data stream, it is impossible to store whole data objects. The memory requirement of mining algorithm should be independent of the number of data objects and it is better to be a fixed memory size. The ordering of data objects requires that the processing of each data object should be in time, and each data point can only be accessed by one-time scanning.
(2) The high dimension of data stream means dimension reduction methods can be employed to reduce time and memory consumption, such as feature selection and extraction.
(3) The change of data stream should be monitored and the occurrence of concept drift should be detected. The analysis model need to be adapted for the change of data stream and correctly describes the distribution of data objects when concept drift occurs. Therefore, the detection of concept drift is closely related with the updating of analysis model.

3.2 Problem and strategy

In this section, some problems existing in data stream mining methods are discussed and the corresponding strategies are proposed.

(1) The problem of data collection. The data objects in data stream is continuously generated and data collection system should be robust, because the downtime of data collection system means the loss of data objects. Flume is a log collection, aggregation and transmission system with high availability and reliability provided by Cloudera. Flume supports the data structure customization for senders, provides the simple treatment of data, and writes to diverse receivers. Kafka is a distributed publish subscribe message system with high throughput capacity. The purpose of Kafka is to unify online and offline message processing through the parallel loading mechanism of Hadoop, and provide real-time consumption through the cluster.

The sources of data stream are diverse and they do not always have strict data structure. The network stream is based on TCP/IP network and data collection is established in TCP/IP network with HTTP protocol. Since web sites are the first use case of large scale data collection, the log format used by web servers has become popular. JavaScript Object Notation (JSON) is one of the most popular log format which is easy to be parsed and extended.

(2) The problem of concept drift detection. Concept drift detection is closely related with data window and the appropriate window size is significant for fast and effectively detecting concept drift. Large window size can delay the detection and may cause missed detection. Conversely, small window size is sensitive to noises and may cause false alarm. The window size can be determined according to Hoeffding bound [10]. The occurrence of concept drift means the change of data distribution, and concept drift can be detected according to data distribution. The

change of data distribution between two adjacent data windows should be within a limited range. If the variation degree is large, it thinks concept drift occurs.

(3) The problem of anomaly detection based on data stream. Generally, data stream clustering and data stream classification are applied for anomaly detection. Similarity measurement is an important issue for data stream clustering and classification. Euclidean distance is a common choice and cosine distance also performs well in text processing. The concerned problem of data stream clustering using in anomaly detection is how to detect anomaly point. Clustering method belongs to unsupervised learning method and does not need additional label resources. It can divide data points into nearest cluster according to similarity measurement, but it cannot decide the label of data points. The application of clustering method in anomaly detection is based on two conditions: the numbers of normal behaviors are much more than that of anomaly behaviors; there are obvious differences between normal and anomaly behaviors. Consequently, the label of clusters can be inferred from its number.

The concerned problem of data stream classification using in anomaly detection is adaptive updating of classifiers. The additional label resources are required to update classifiers that make updating process more complex. Clustering method can be integrated with the updating of classifiers and the preliminary label results can be utilized as label resources of classifiers.

4 Improved anomaly detection model

In the paper [30, 31], we propose an anomaly detection model based on data stream clustering and employ two synopsis data structures to store summary statistics information of clusters as shown in Formula 1. *n-cluster* represents normal cluster and *s-cluster* denotes the cluster which is suspected as anomaly cluster.

$$\begin{cases} n-cluster : (\delta, \mu, SS, flag) \\ s-cluster : (\delta, \mu, SS, flag, list) \end{cases} \quad (1)$$

In Formula 1, δ is the number of data points in the cluster; μ is the cluster center; SS is the quadratic sum of data points in the cluster. The property *flag* is added to identify the type of clusters and *list* stores the index of data objects in *s-cluster*. When new data object arrives, these properties can be updated as Formula 2.

$$(\delta, \mu, SS) \rightarrow \left(\delta + 1, \frac{\mu \times \delta + o}{\delta + 1}, SS + o^2\right) \quad (2)$$

The improved anomaly detection model is shown in Fig. 4, and we add classifier and concept drift detecting module into anomaly detection model. Clustering module will update clusters according to new arrived data point. Classifier module is trained by initial training data set and label new arrived data point. Meanwhile, classifier module will be updated according to clusters information from clustering module when concept drifting module detects the occurrence of concept drift. The main ideas of improved anomaly detection model are as follows and detailed process can be obtained in the paper [31]:

(1) Clustering module waits for receiving data objects and clustering algorithm will generates initial clusters. These clusters are labelled as *n-cluster* or *s-cluster* according to the number of data objects. Classifier module are trained by initial training data sets;

(2) When data window receives enough data objects, they will be sent to concept drift detecting module. If it detects the occurrence of concept drift through the method mentioned in Sect. 2.3, the summary statistics of clusters in clustering module are utilized to update classifier. Otherwise, they are labelled directly by classifier and update the clusters of clustering module.

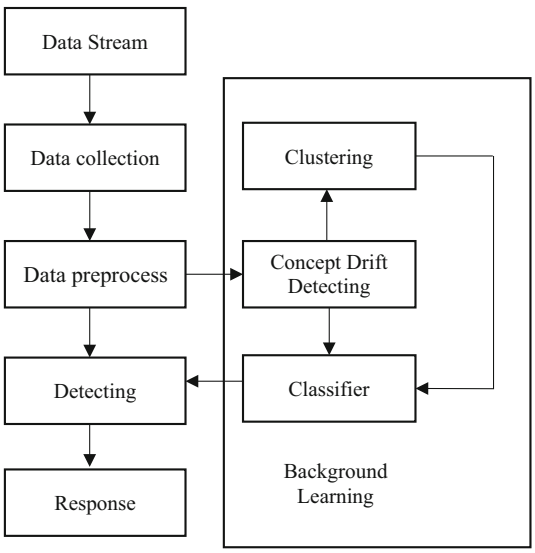

Fig. 4 Improved anomaly detection model

5 Conclusion

Traditional data mining methods are based on static offline data sets and these methods can perform better on data sets, but they may be not effective for data streams. The distribution of data stream may change over time and the change is out of control. The occurrence of concept drift causes original analysis model not work and it requires the retraining or updating of analysis model.

Focus on these problems, we discuss and propose some strategies for data stream mining. The improved anomaly detection model based on clustering method is also designed. The generated clusters can be updated immediately when new data object comes and the summary statistics of clusters are utilized to retrain or update classifiers when it detects the occurrence of concept drift.

Acknowledgements This work was funded by the National Natural Science Foundation of China (61772282, 61772454, 61373134, 61402234). It was also supported by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX17_0901) and Jiangsu Collaborative Innovation Center on Atmospheric Environment and Equipment Technology (CICAEET). It was also funded by the open research fund of Key Lab of Broadband Wireless Communication and Sensor Network Technology (Nanjing University of Posts and Telecommunications), Ministry of Education. Professor Seungwook Min is the corresponding author.

Compliance with ethical standards

Conflict of interest Ruxia Sun declares that she has no conflict of interest. Sun Zhang declares that he has no conflict of interest. Chunyong Yin declares that he has no conflict of interest. Jin Wang declares that he has no conflict of interest. Seungwook Min declares that he has no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Lee, W., Stolfo, S., Mok, K.: Mining audit data to build intrusion detection models. In: International conference on knowledge discovery & data mining, pp. 66–72 (1998)
2. Keegan, N., Ji, S.Y., Chaudhary, A., Concolato, C., Yu, B., Jeong, D.H.: A survey of cloud-based network intrusion detection analysis. Hum. Centric Comput. Inf. Sci. **6**(1), 19–35 (2016)
3. Yin, C., Zhang, S., Xi, J., Wang, J.: An improved anonymity model for big data security based on clustering algorithm. Concurr. Comput. **29**(7), 1–13 (2017)
4. Yin, C., Zhang, S.: Parallel implementing improved k-means applied for image retrieval and anomaly detection. Multimed Tools Appl. **76**, 1–17 (2017)
5. Wang, G., Hao, J., Ma, J., Huang, L.: A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering. Expert Syst. Appl. **37**(9), 6225–6232 (2010)
6. Li, L., Ye, J., Deng, F., Xiong, S., Zhong, L.: A comparison study of clustering algorithms for microblog posts. Clust. Comput. **19**(3), 1333–1345 (2016)
7. Li, W., Li, X., Yao, M., Jiang, J., Jin, Q.: Personalized fitting recommendation based on support vector regression. Hum. Centric Comput. Inf. Sci. **5**(1), 21–32 (2015)
8. Gu, B., Sun, X., Sheng, V.S.: Structural minimax probability machine. IEEE Trans. Neural Netw. Learn. Syst. **28**(7), 1646–1656 (2017)
9. Gu, B., Victor, S.S.: A robust regularization path algorithm for ν-support vector classification. IEEE Trans. Neural Netw. Learn. Syst. **28**(5), 1241–1248 (2017)
10. Gu, B., Sheng, V.S., Tay, K.Y., Romano, W., Li, S.: Incremental support vector learning for ordinal regression. IEEE Trans. Neural Netw. Learn. Syst. **26**(7), 1403–1416 (2015)
11. De la Hoz, E., de la Hoz, E., Ortiz, A., Ortega, J., Martínez-Álvarez, A.: Feature selection by multi-objective optimisation: application to network anomaly detection by hierarchical self-organising maps. Knowl Based Syst. **71**, 322–338 (2014)
12. Yin, C., Zhang, S., Kim, K.J.: Mobile anomaly detection based on improved self-organizing maps. Mob Inf Syst. **2017**, 1–9 (2017)
13. Ma, T., Zhang, Y., Cao, J., Shen, J., Tang, M., Tian, Y., Al-Dhelaan, A., Al-Rodhaan, M.: KDVEM: a k-degree anonymity with vertex and edge modification algorithm. Computing **97**(12), 1165–1184 (2015)
14. Fu, Z., Ren, K., Shu, J., Sun, X., Huang, F.: Enabling personalized search over encrypted outsourced data with efficiency improvement. IEEE Trans. Parallel Distr. **27**(9), 2546–2559 (2016)
15. Wang, J., Zhang, Z., Li, B., Lee, S., Sherratt, R.: An enhanced fall detection system for elderly person monitoring using consumer home networks. IEEE Trans. Consum. Electr. **60**(1), 23–29 (2014)
16. Younghee, K., Wonyoung, K., Ungmo, K.: Mining frequent itemsets with normalized weight in continuous data streams. J. Inform. Process. Syst. **6**(1), 79–90 (2010)
17. Fong, S., Hang, Y., Mohammed, S., Fiaidhi, J.: Stream-based biomedical classification algorithms for analyzing biosignals. J. Inform. Process. Syst. **7**(4), 717 (2011)
18. El-Semary, A.M., Mostafa, G.H.M.: Distributed and scalable intrusion detection system based on agents and intelligent techniques. J. Inform. Process. Syst. **6**(4), 481–500 (2010)
19. Krawczyk, B., Minku, L.L., Gama, J., Stefanowski, J., Woźniak, M.: Ensemble learning for data stream analysis: a survey. Inform. Fusion. **37**, 132–156 (2017)
20. Domingos, P., Hulten, G.: Mining high-speed data streams. In: ACM SIGKDD international conference on knowledge discovery and data mining, pp. 71–80 (2000)
21. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. ACM Sigmod. Rec. **29**(2), 1–12 (2000)
22. Czarnowski, I., Jędrzejowicz, P.: Ensemble online classifier based on the one-class base classifiers for mining data streams. Cybern. Syst. **46**(1–2), 51–68 (2015)
23. Gaur, M.S., Pant, B.: Trusted and secure clustering in mobile pervasive environment. Hum. Centric Comput. Inf. Sci. **5**(1), 1–17 (2015)
24. Guha, S., Meyerson, A., Mishra, N., Motwani, R.: Clustering data streams: theory and practice. IEEE Trans. Knowl. Data Eng. **15**(3), 515–528 (2003)
25. Aggarwal, C., Yu, P., Han, J., Wang, J.: A framework for clustering evolving data streams. In: International conference on very large data bases, pp. 81–92 (2003)
26. Chen, Y., Tu, L.: Density-based clustering for real-time stream data. In: ACM SigkDD international conference on knowledge discovery & data mining, pp. 133–142 (2007)

27. Ramírez-Gallego, S., Krawczyk, B., García, S., Woźniak, M., Herrera, F.: A survey on data preprocessing for data stream mining: current status and future directions. Neurocomputing **239**(C), 39–57 (2017)
28. Oh, S., Kang, S., Byun, Y., Jeong, T., Lee, W.: Anomaly intrusion detection based on clustering a data stream. In: ACIS international conference on software engineering research, management and applications, pp. 220–227 (2005)
29. Guerrieri, A., Montresor, A.: DS-means: distributed data stream clustering. In: International conference on parallel processing, pp. 260–271 (2012)
30. Yin, C., Zhang, S., Yin, Z., Wang, J.: Anomaly detection model based on data stream clustering. Clust. Comput. **2017**, 1–10 (2017)
31. Yin, C., Zhang, S., Wang, J.: Improved data stream clustering algorithm for anomaly detection. Adv. Multimed. Ubiquitous Eng. **448**, 620–625 (2017)
32. Hulten, G., Spencer, L., Domingos, P.: Mining time-changing data streams. In: ACM SigkDD international conference on knowledge discovery & data mining, pp. 97–106 (2001)
33. Rocha, R., Medas, P.: Accurate decision trees for mining high-speed data streams. In: ACM SigkDD international conference on knowledge discovery & data mining, pp. 523–528 (2003)
34. Gomes, H.M., Bifet, A., Read, J., Barddal, J.P., Enembreck, F., Pfharinger, B., Holmes, G., Abdessalem, T.: Adaptive random forests for evolving data stream classification. Mach Learn. **106**(9–10), 1469–1495 (2017)
35. Pietruczuk, L., Rutkowski, L., Jaworski, M., Duda, P.: How to adjust an ensemble size in stream data mining? Inform. Sci. **381**, 46–54 (2017)
36. Silva, J., Faria, E., Barros, R., Hruschka, E.: Data stream clustering: a survey. ACM Comput. Surv. **46**(1), 125–134 (2013)
37. Cao, F., Ester, M., Qian, W., Zhou, A.: Density-based clustering over an evolving data stream with noise. In: SIAM international conference on data mining, pp. 328–339 (2006)
38. Udommanetanakit, K., Rakthanmanon, T., Waiyamai, K.: E-stream: evolution-based technique for stream clustering. In: International conference on advanced data mining and applications, pp. 605–615 (2007)
39. Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., Bouchachia, A.: A survey on concept drift adaptation. ACM Comput. Surv. **46**(4), 44 (2014)
40. Laohakiat, S., Phimoltares, S., Lursinsap, C.: A clustering algorithm for stream data with LDA-based unsupervised localized dimension reduction. Inform. Sci. **381**, 104–123 (2017)

Ruxia Sun received the B.E. degree in Electrical Engineering from the Shandong University of Technology, China, in 1997. She is an associate Professor in the Nanjing University of Information Science & Technology, China. Her current research interests include machine learning, network security and intrusion detection.

Sun Zhang received his bachelor degree in Computer Science and Technology from Nanjing University of Information Science and Technology, China. Now he is studying for his master's degree in there. His current research interests are in machine learning, network security and intrusion detection.

Chunyong Yin received the B.S. degree from Shandong University of Technology, China in 1998. He received the M.S. and Ph.D. degrees in Computer Science from Guizhou University, China, in 2005 and 2008, respectively. He was a post-doctoral research associate at the University of New Brunswick, Canada, in 2011 and 2012. He is a Professor and Dean with the Nanjing University of Information Science & Technology, China. His current research interests include privacy preserving and sensor networking, machine learning and network security.

Jin Wang received the B.S. and M.S. degree from Nanjing University of Posts and Telecommunications, China in 2002 and 2005, respectively. He received Ph.D. degree from Kyung Hee University Korea in 2010. Now, he is a professor in the College of Information Engineering, Yangzhou University. His research interests mainly include routing algorithm design, performance evaluation and optimization for wireless ad hoc and sensor networks. He is a Member of IEEE and ACM.

Seungwook Min received the B.S. degree in Control and Instrumentation Engineering from Seoul National University in 1987, and the M.S. in Electrical Engineering from Korea Advanced Institute of Science and Technology, Seoul, in 1990. He received Ph.D. degree in Electrical Engineering from Polytechnic University (now NYU), U.S.A in 1999. From 1999 to 2002, he was a principal engineer at Samsung Electronics. Now he is a professor in the Department of Computer Science, Sang Myung University. His research interests include WLAN networking, mobile communications and UWB systems.