Research Article

Application of High-Dimensional Outlier Mining Based on the Maximum Frequent Pattern Factor in Intrusion Detection

Limin Shen,1 Zhongkui Sun,1,2 Lei Chen,3 and Jiayin Feng1

1School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China
2Qinggong College, North China University of Science and Technology, Tangshan 063000, China
3Graduate School, North China University of Science and Technology, Tangshan 063000, China

Correspondence should be addressed to Zhongkui Sun; sunzhk7965@stumail.ysu.edu.cn

Received 12 August 2020; Revised 1 May 2021; Accepted 14 June 2021; Published 22 June 2021

Abstract

As the Internet applications are growing rapidly, the intrusion detection system is widely used to detect network intrusion effectively. Aiming at the high-dimensional characteristics of data in the intrusion detection system, but the traditional frequent-pattern-based outlier mining algorithm has the problems of difficulty in obtaining complete frequent patterns and high time complexity, the outlier set is further analysed to get the attack pattern of intrusion detection. The NSL-KDD dataset and UNSW-NB15 dataset are used for evaluating the proposed approach by conducting some experiments. The experiment results show that the method has good performance in detection rate, false alarm rate, and recall rate and effectively reduces the time complexity.

1. Introduction

1.1. Intrusion Detection System. With the rapid development of modern information technology, network security has become the focus of attention. How to effectively detect the types of intrusion attacks, as well as the security of the early warning and protection system, has become one of the research directions of network security. Intrusion detection systems (IDSs) are most widely used in the world for identifying and detecting the intruders in computer networks, Internet, and cloud networks. The intrusion detection system analyses the network data collected by the computer system and the key points in the network, so as to find out the behaviour of violating the security policy and the traces of attacks and monitor and detect the network intruders. The IDS can be used to detect different types of attacks on the network, but the traditional firewall cannot perform these attacks well.

Generally, the intrusion detection system can be roughly divided into two categories according to its detection methods, namely, an anomaly detection system and detection system. Anomaly detection is also known as behaviour-based system detection, which detects the abnormal behaviour of the system to discover intrusion behaviour. Misuse detection is a knowledge-based detection or feature-based detection technology, whose premise is that intrusion behaviour and normal network access have different data characteristics. The intrusion detection system is divided into two stages, namely, the preprocessing stage and intrusion detection stage. By developing the intrusion detection system, the intrusion behaviour can be identified effectively.

1.2. Outlier Detection. Outlier mining is an important research direction in the field of data mining. Outlier data do not conform to the general rules of data and are not consistent with other parts of the data. It is those small-scale objects that are far away from other objects in the dataset. Although outlier data are "abnormal data" which are inconsistent with normal data, outlier detection can provide important information in some applications.

There are many reasons for outliers. Generally speaking, they can be divided into two situations: first, they are indeed caused by human or detection equipment errors; second, they are caused by the nature of things themselves, and they are the data reflection of the real nature of things. The outlier
analysed in this paper belongs to the second case. The outlier
data generated by human operation are significantly dif-
ferent from the normal network behaviour, in order to find
the real potential valuable knowledge through outlier
mining.

In the real network activities, most of the network be-
haviours are normal, the intrusion behaviour can be
regarded as the abnormal phenomenon of the amount of
data far less than the normal behaviour, and the data cor-
responding to the normal behaviour and the intrusion be-
haviour have different data characteristics. Based on the
characteristics of intrusion behaviour data, intrusion be-
haviour can be regarded as “outlier” data [1].

1.3. Association Rule Mining. Association rule mining, as an
important part of data mining, has been a hot research topic.
Association rules are a collection of items in the database
that exceed the specified minimum support and minimum
confidence. Association rules are usually expressed as X \rightarrow Y,
support = s, and confidence = c, in which X is the precon-
dition of the rule, Y is the conclusion of the rule, the support
s represents the frequency of the rule, and the confidence c
represents the strength of the rule.

The goal of association rule mining is to find out all the
strong association rules. The mining process is divided into
two steps:

Step 1: all rules that are not less than the minimum
support threshold s are found, i.e., all frequent patterns
Step 2: by setting the confidence threshold c, the
conversion rule is used to filter out the set of items less
than the minimum confidence c, and the corresponding
association rules are obtained

In this paper, it is only needed to get the maximum
frequent patterns based on frequent pattern, so it is only
needed to complete Step 1 to get the frequent pattern.

1.4. Maximum Frequent Pattern. If the maximum frequent
pattern needs to be explained, the concept of supersets must
be introduced first, which is defined as follows: if every
element in set S2 is in set S1 and set S1 may contain elements
that are not in S2, then set S1 is a superset of set S2. If set S1 is a
superset of set S2, then set S2 is a true subset of set S1, and vice
versa.

With the superset, the maximum frequent pattern is
defined as follows: if all supersets of frequent pattern X are
nonfrequent patterns, then X is called as a maximum fre-
quent pattern.

With the increasing number and dimension of collected
data in the intrusion detection system, researchers have
proposed a variety of typical high-dimensional outlier
mining algorithms for the complexity, sparsity, and diversity
of high-dimensional data. Among them, outlier mining
based on frequent pattern is widely used in intrusion de-
tection because of its easy-to-understand nature and low
time complexity. On the basis of frequent-pattern-based
outlier mining algorithm, using the concept of maximum
frequent pattern in association rules, an improved high-
dimensional outlier mining algorithm based on the maxi-
mum frequent pattern is proposed in this paper. The al-
gorithm transforms frequent pattern mining into maximum
frequent pattern mining. On the premise of good detection
performance, the time complexity is reduced.

2. Literature Survey

In the real network, the data are high dimensional in the
intrusion detection system. Some researchers proposed the
means to reduce the dimension of high-dimensional data
with the way of feature extraction or feature selection and
then analysed the processed data with the traditional data
mining methods.

Ganapathy [2] proposed an intelligent algorithm for
feature selection and classification to design an effective
intrusion detection system, which can be used to provide
security to networks effectively.

Tian et al. [3] proposed a hierarchical outlier detection
model based on PCA, an anomaly data model based on PCA
was established based on normal data to filter data firstly,
and then, the abnormal data types were analysed to detect
both anomaly and misuse attack.

Zyad et al. [4] proposed a way to use the trimmed av-
erage vector to estimate the average vector on the basis of
PCA, so as to make the trimmed PCA have better
robustness.

To solve the problem of high-dimensional data in IDS,
Riyaz and Ganapathy [5] proposed a new fuzzy rule and
information gain ratio-based feature selection algorithm
(FRFSA), and the existing classifiers called SVM and LSSVM
were used for effective classification. The experimental result
shows that the proposed work exceeds the performance
measure when compared to the existing algorithms on
classification for feature selection.

Nancy et al. [6] proposed a dynamic recursive feature
selection algorithm for feature selection and then used an
intelligent fuzzy temporal decision tree algorithm to effec-
tively detect intruders, which can effectively reduce the false
positive rate, energy consumption, and delay of the system.

The method of dimension reduction can eliminate some
features and reduce the time complexity, but each feature
represents a different outlier value. If the features are selected
incorrectly, it will get the wrong outlier value, which will
produce an approximate result that is not suitable for future
calculation [7]. The complexity, sparsity, and diversity of
high-dimensional data restrict the traditional mining al-
gorithm. When dealing with high-dimensional data, data
mining algorithms suitable for low-dimensional data usually
encounter the problems of algorithm efficiency reduction
and the traditional definition based on distance and density
is invalid, which reduces the accuracy of intrusion detection
[8].

Researchers have proposed intrusion detection methods
for high-dimensional data. Zhang et al. [9] proposed SPOT
technology for anomaly detection in a high-dimensional
data network data stream, which has good detection effect.
Prajapati and Bhartiya [10] proposed a nearest neighbour search algorithm based on the advantages of K-mean algorithm and fuzzy C-mean (PCM) algorithm to solve the problem of uneven data and rigid clustering in high-dimensional data, which can realize nearest neighbour search in a shorter time.

In general, the “attack” data in intrusion behaviour are regarded as abnormal data, and outlier mining is to mine those abnormal data which deviate from normal behaviour in large-scale data, so outlier mining is very important for analysing intrusion behaviour. For high-dimensional outlier mining, researchers have proposed several typical mining algorithms: outlier mining algorithm based on spatial projection [11, 12], outlier mining algorithm based on a hypergraph model [13, 14], and outlier mining algorithm based on frequent patterns. The outlier mining algorithm based on frequent patterns is simple, easy to understand, and has lower time complexity than the previous two algorithms, so researchers have conducted extensive research.

In the early stage, He et al. [15] proposed an outlier mining algorithm based on frequent patterns (FindFPOF) and proposed a measurement factor of frequent pattern outlier factor (FPOF). It is believed that the less frequent the patterns contained in a data record, the more likely they would be an outlier, so outliers could be found by calculating the frequent pattern factor of each data.

Zhou [16] proposed a new metric called weighted frequent pattern outlier factor for categorical data streams based on FindFPOF and proposed a fast outlier detection method for high-dimensional categorical data streams based on frequent pattern (FODFP-Stream), which has good applicability and validity.

Wang and Tang [17] proposed an algorithm based on frequent patterns-NFPOF, which further accurately locates abnormal properties of each outlier data through the related attributes of frequent patterns.

Yuan et al. [18] proposed a weighted frequent-pattern-based outlier (WFP-Outlier) to solve the problem whose weights seriously affect outlier detection results, which can find implicit outliers from weighted data streams.

To solve the problem of being incapable of detecting new type of attacks, Jaisankar [19] proposed a new intelligent-agent-based IDS using Fuzzy rough-set-based outlier detection and Fuzzy rough-set-based SVM. The system adopted Fuzzy rough-based SVM in our system to classify and detect anomalies efficiently. The experimental result shows that the proposed intelligent-agent-based model improves the overall accuracy and reduces the false alarm rate.

In order to solve the problem of high false positives, Ganapathy [20] proposed a new intrusion detection model using a new Weighted-Distance-Based Outlier Detection (WDBOD) algorithm and an Enhanced Multiclass Support Vector Machine algorithm, which has low false alarm rate and high accuracy.

Combined with attribute selection, outlier detection, and the enhanced multiclass support vector machine classification method, Ganapathy et al. [21] proposed a new intelligent-agent-based intrusion detection model for mobile ad hoc networks. Using the proposed Intelligent Agent Weighted Distance Outlier Detection algorithm and Intelligent-Agent-based Enhanced Multiclass Support Vector Machine algorithm, the proposed model can detect anomalies with low false alarm rate and high accuracy.

To sum up, high-dimensional outlier mining based on frequent patterns plays an important role in intrusion detection, but there are two problems in the algorithms based on frequent patterns. First, it needs to mine the complete frequent patterns in the dataset, but it is very difficult to find the complete set of frequent patterns in high-dimensional data. Second, the time complexity of mining algorithm for frequent patterns is exponentially related to the dimension of data, the higher the dimension, the greater the time complexity. High-dimensional outlier mining algorithm based on frequent patterns has the problems of difficulty in obtaining complete frequent patterns and high time complexity. So, a high-dimensional outlier mining algorithm based on the maximum frequent pattern factor is proposed in this paper using the concept of maximum frequent pattern factor in association rules. Also, the algorithm is applied in intrusion detection, which reduces the time complexity on the premise of ensuring good detection performance.

3. Proposed Work

3.1. Relevant Theories. We let \( D = \{t_0, t_2, \ldots, t_m\} \) be a dataset containing \( n \) network behaviour records \( t \), and \( t_k \) is called a transaction. Also, \( I = \{i_1, i_2, \ldots, i_p\} \) is the collection of all attributes in the network behaviour record, and \( i_m \) is called an item.

**Definition 1.** Itemset: any subset \( X \) of \( I \) is called the itemset of \( D \). We let \( t_k \) be a transaction of \( D \), and \( X \) is a itemset of \( D \); if \( X \subseteq t_k \), then the itemset \( D \) is contained in the transaction \( t_k \).

**Definition 2.** Support: the support number of itemset \( X \) in dataset \( D \) and is recorded as \( \text{support}(X) \).

**Theorem 1.** \( X, Y \) are set as itemsets in dataset \( D \); then,
(1) If \( X \subseteq Y \), then \( \text{support}(X) \geq \text{support}(Y) \)
(2) If \( X \subseteq Y \) and \( X \) is not a frequent pattern, then \( Y \) is not a frequent pattern
(3) If \( X \supseteq Y \) and \( Y \) is a frequent pattern, then \( X \) is a frequent pattern

where \( D \) is the total number of transactions in dataset \( D \).

**Definition 3.** Frequent pattern: if the support \( X \) is not less than the minimum support (MinSP) which is specified by the user, then \( X \) is a frequent pattern; otherwise, it is an infrequent pattern.
and if \( K \) value is increased continuously, the change of SSE will tend to be gentle, that is to say, the relationship graph between SSE and \( K \) is the shape of an elbow, and the corresponding \( K \) value of this elbow is the optimal number of clusters.

The square sum of error (SSE) of the core index of the elbow method is defined as

\[
SSE = \sum_{i=1}^{K} \sum_{p \in C_i} |p - m_i|^2,
\]

where \( C_i \): the \( i_{th} \) clustering, \( p \): sample points in \( C_i \), \( m_i \): the centroid of \( C_i \) (mean value of all samples in \( C_i \)), and SSE: clustering error of all samples, representing the quality of the clustering effect.

For the sensitive problem of the selection of an initial cluster centre, the maximum distance method is used to select \( K \) samples as the initial centre points based on the fact that the farthest sample points are most unlikely to be divided into the same cluster.

3.3. The Proposed Algorithm. The concept of maximum frequent pattern factor (MFPOF) is proposed based on the frequent pattern factor (FPOF) in FindFPOF algorithm.

Definition 6. Maximum frequent pattern factor (MFPOF): \( \text{MFPS} (D, \text{MinSP}) \) is the maximum frequent pattern sets in dataset \( D \) that meets a given minimum support threshold. The MFPOF of each network behaviour record \( t \) is defined as

\[
\text{MFPOF} (t) = \frac{\sum_{X \in t, X \in \text{MFPS} (D, \text{MinSP})} \text{support} (X)}{\| \text{MFPs} (D, \text{MinSP}) \|}
\]

where \( \| \text{MFPs} (D, \text{MinSP}) \| \) is the number of the maximum frequent patterns in frequent patterns and the support\((X)\) is the support of a maximum frequent pattern \( X \).

The description of the high-dimensional outlier mining algorithm based on maximum frequent patterns (MFPOF-OM) is shown as Algorithm 1.

3.4. Automatically Constructing Intrusion Detection Patterns Based on Association. Association analysis can automatically discover the data characteristics of network behaviour. The maximum frequent patterns generated by association analysis can reflect the maximum common characteristics of network behaviour data, which are expressed by the attribute values of network behaviour data. So, these attribute values can be used to build intrusion detection patterns with strong classification ability [22].

Taking the outlier dataset obtained by MFPOF-OM algorithm as input and setting a minimum support threshold, the maximum frequent patterns of the outlier dataset can be obtained referring to Step 1–3 of Algorithm 1, which are the intrusion detection patterns of network attack.

3.5. System Architecture. According to the abovementioned analysis, the architecture of the system proposed in this work consists of six major modules such as data preprocessing, an
4. Results and Discussion

4.1. Dataset and Experimental Environment. The specifications of the hosts adopted in the experiments are Core Intel Core i5-6300HQ, 2.3 GHz CPU, 16 GB RAM, and Windows 7. The proposed method is verified in MATLAB 2012. The NSL-KDD dataset [23] and UNSW-NB15 dataset [24] are used as the experimental datasets to verify the proposed method in this paper.

First, the experimental results of the proposed algorithm are analysed in the NSL-KDD dataset, and then, the proposed algorithm is compared with other researchers’ algorithms to verify the effectiveness; lastly, the experimental results in the NSL-KDD dataset and UNSW-NB15 dataset are compared to verify the applicability of the proposed algorithm.

The NSL-KDD dataset is an effective benchmark dataset to help researchers compare different intrusion detection methods. There are 125,973 connection records in the NSL-KDD dataset. Each connection record is described by 41 attributes about the network packet, network traffic, host traffic, and content information. The 22 categories of attacks are from the following four classes: DoS, R2L, U2R, and Probing. Also, the 20th attribute (num_outbound_files) can be deleted because its attribute value is all 0, so its information entropy is 0 according to information theory.

The raw network packets of the UNSW-NB15 dataset are created for generating a hybrid of real modern normal activities and synthetic contemporary attack behaviours. It is suitable for researchers to study the intrusion detection system. There are 175,341 records in the training set and 82,332 records in the testing set. This dataset has totally 49 features with the class label and 9 families of attacks, namely, Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, and Worms.

The NSL-KDD dataset is a factual benchmark in the field of network intrusion detection, which lays a foundation for the research of network intrusion detection based on computational intelligence. First, the NSL-KDD dataset eliminates duplicate records and classifiers that prefer more duplicate records. Second, it eliminates the imbalance between the number of records and reduces the false positive rate. Therefore, although the NSL-KDD dataset is older, it is widely used to evaluate the performance of the IDS. The UNSW_NB15 dataset is a comprehensive network attack traffic dataset, which combines the real normal network traffic attack activities and modern network traffic comprehensive attack activities and can better reflect the real environment of the network, so it is widely used in abnormal intrusion detection [25, 26].

The proposed algorithm needs to mine the maximum frequent pattern, which requires that the data type must be discrete. Taking the NSL-KDD dataset as an example, the dataset values’ processing is introduced, which is suitable for the proposed algorithm. According to the analysis of the

outlier mining module, constructing intrusion detection patterns, attack patterns base, pattern match, and an alarm system, as shown in Figure 1. The data preprocessing module is for performing preprocessing activities, but its main function is to discretize the data and make it suitable for the proposed algorithm. The outlier mining module is used to obtain the outlier data by the proposed algorithm. On the basis of acquiring outlier data, an intrusion detection pattern module is used to obtain intrusion detection patterns, so as to construct the attack pattern library module. The pattern match module is used to match the testing data with the attack rule base. If the match is successful, it indicates that there is an intrusion attack and transfers to the alarm module to trigger the alarm.

Algorithm 1: MFPOF-OM algorithm.

```
Input: D // network behaviour dataset
MinSP // minimum support threshold
k // number of outliers threshold
Output: k network behaviour outlier data records
Begin
// Step 1–3: mining the maximum frequent item sets based on PF-Tree Algorithm
Step 1: To D, the HeaderTable (D) is generated to satisfy the MinSP; // Calculating the header table of PF-tree
Step 2: To D, the frequent item set tree is generated to satisfy the given MinSP by using the PF-Tree Algorithm, and denoted as T_j; // Obtains frequent item set tree according to the PF-Tree algorithm
Step 3: Obtains maximum frequent item sets based on an improved PF-Tree, and obtains MFPs (D, MinSP) and support (X); // Obtains maximum frequent item sets
// Step 4–7: Mine k outliers data with minimum MFPOF value based on the obtained MFPs
Step 4: foreach t in D
According to formula (3), calculates the maximum frequent patterns factor of each record t: MFPOF(t);
end foreach // Calculating maximum frequent factor of each transaction t
Step 5: Obtains a MFPOF value of each network behaviour records t;
Step 6: For all t, they are sorted in ascending order according to MFPOF(t);
Step 7: Return the first k network behaviour record with the minimum MFPOF value, and they are k outlier data in the network behaviour data.
End
```

The specifica-
NSL-KDD dataset, the attribute data type of the dataset can be divided into the text type and numerical type, and the numerical type can be divided into the discrete type and continuous type. The types of data are shown in Table 1 for the text-type and numerical discrete-type data which have met the data requirements. However, the continuous numerical data represented by columns 1, 5, and 6 are discretized using the discretization algorithm given in Section 3.2 and transformed into reliable and accurate data suitable for data mining.

4.2. Experiments in the NSL-KDD Dataset. Experiment A: the experimental results of the proposed algorithm in the NSL-KDD dataset are analysed in the experiment. The accuracy, false positive rate, and complexity analysis are used as the performance evaluation criteria to determine the results. Four groups of sample data were extracted from the dataset: Normal + DoS, Normal + Probing, Normal + R2L, and Normal + U2R.

4.2.1. Experiment Results of Four Network Attack Patterns. By comparing the detection rate and false positive rate under different MinSP thresholds of four groups of sample data, Normal + DoS, Normal + Probing, Normal + R2L, and Normal + U2R, the detection effect of the proposed algorithm is illustrated, and then, the feasibility of the proposed algorithm is verified. The experimental results of DoS, Probing, R2L, and U2R intrusion detection patterns obtained from the analysis of four groups of sample data are shown in Figure 2.

Probing attack detection patterns are taken as an example for data analysis. The Normal + Probing sample set contains 62000 pieces of data, the threshold value of MinSP is different, and the detection patterns are also different in the experiment. The experimental results are shown in Figure 2(b), which shows the detection patterns acquired under the MinSP thresholds of 58500, 59000, and 60000 and uses the acquired Probing detection patterns to detect five data types (DoS, Probing, R2L, U2R attack data, and Normal data), respectively. It is found that when the threshold value is 59000, the accuracy of Probing detection patterns to Probing data is 88%, and the false alarm rate is 2% to Normal data, 4% to DoS, 1% to R2L, and 10% to U2R data. When the threshold values are 58000 and 60000, the results are as shown in Figure 2(b) and will not be described one by one.

By comparing the four intrusion detection attack modes in Figure 2, it is found that the accuracy will be better when the minimum support threshold is larger, and the detection error of other data is basically the same, although the size varies. It is determined by the characteristics of outlier mining. The larger the threshold is, the fewer the number of outliers is, which can better reflect the characteristics of attack-type data. Of course, the threshold should not be too large, and the accuracy will be reduced if the threshold is too large. Through the comprehensive analysis of detection rate and false detection rate under multiple thresholds, the intrusion detection mode with the best comprehensive detection result is selected as the acquired intrusion detection mode, and the threshold value at this time is taken as the acquired intrusion detection pattern threshold: the threshold of DoS attack is 59100, the threshold of Probing attack is 59000, the threshold of R2L attack is 59600, and the threshold of U2R attack is 59500. The evaluation parameters are shown in Table 2.

Comparing the four subgraphs in Figure 2, it is found that U2R-type data have the highest detection errors in DoS, Probing, and R2L attack intrusion detection patterns, which are 4%, 10%, and 33%, respectively, and compared with the other three attack intrusion detection patterns, the accuracy of U2R attack intrusion detection mode is relatively low, only 87%, which is determined by the number of U2R, only 52 pieces of U2R data in the NSL-KDD dataset, so data mining cannot fully discover its data characteristics, resulting in incomplete detection performance.

Comparing Figure 2(c) with Figure 2(d), it is found that there are higher errors in the detection of U2R data by using R2L attack intrusion detection patterns and R2L data by using U2R attack intrusion detection patterns, which shows that R2L-type data and U2R-type data have higher data similarity compared with other three types of data, which is consistent with the characteristics of two kinds of network attacks in reality.

4.2.2. Complexity Analysis. In this section, the complexity of 4 groups of sample data, Normal + DoS, Normal + Probing, Normal + R2L, and Normal + U2R, will be analysed. The FindFPOF algorithm based on frequent patterns and other outlier mining algorithms based on weighted frequent patterns need to mine frequent patterns first, and the time complexity is similar. Here, FindFPOF algorithm is taken as an example to illustrate.

<table>
<thead>
<tr>
<th>Table 1: NSL-KDD dataset attribute data types.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute types</td>
</tr>
<tr>
<td>-------------------------------------</td>
</tr>
<tr>
<td>Text type</td>
</tr>
<tr>
<td>Numerical discrete type</td>
</tr>
<tr>
<td>Continuous numerical data</td>
</tr>
</tbody>
</table>
The total time complexity of FindFPOF algorithm is \(O(m^2 + m*n + m*logm)\), where \(m\) is the amount of data and \(n\) is the amount of frequent patterns.

The MFPF-OM algorithm has three steps: (1) mining maximum frequent patterns from the dataset, the time complexity is \(O(m^2)\); (2) calculating the MFPOF(t) of each network behaviour record, the time complexity is \(O(m*l)\); and (3) discovering \(K\) network behaviour outliers, the time complexity is \(O(m*logm)\). Therefore, the time complexity from the abovementioned three steps is proved as follows: 

\[ T(\text{MFPOF-OM}) = O(m^2 + m*l + m*logm) \]

where \(m\) is the number of data and \(l\) is the number of maximum frequent patterns.

The number of frequent patterns \((n)\) in FindFPOF algorithm and the number of maximum frequent patterns \((l)\) in MFPF-OM algorithm for 4 groups of sample are shown in Table 3.

For massive data, the value of \(m\) is large enough, and in theory, the time complexity of the two algorithms can be simplified to \(O(m^2)\). But in practice, when the value of \(m\) is not large enough, the proposed algorithm only needs to mine the maximum frequent patterns in Step 3, and \(l \ll n\), as shown in Table 3, so MFPOF-OM algorithm has a better time complexity than the FindFPOF algorithm when calculating MFPOF \((t)\) in Step 4 of the algorithm.

### 4.3. Comparative Experiments between the Proposed Algorithm and Other Algorithms

Experiment B: in order to verify the accuracy of the proposed method, it is compared with the SVM method, Intelligent DT method [6], LSSVM + FRFSA method [5], and Outlier Detection + EMSVW method [20]. The accuracy is used as the performance evaluation criteria to determine the results. The evaluation parameters are shown in Table 4.

The results are shown in Figure 3, in which M1 represents the SVM method, M2 represents the Intelligent DT method, M3 represents the LSSVM + FRFSA method, M4 represents the Outlier Detection + EMSVW method, and M5 represents the proposed method in this paper. The results show that the MFPOF-OM method is very close to the other methods in accuracy of Probing and DoS, but slightly inferior. However, it has a great advantage in the accuracy of R2L and U2R, which shows that the improved dimensional outlier mining method has good characteristics in dealing with outlier data because of the small amount of R2L and U2R attack data in the NSL-KDD dataset. The accuracy data of R2L and U2R are empty in Figure 3 because there are no relevant data in [20]. The overall performance analysis shows that the performance of the proposed method is reliable, can effectively detect the intrusion behaviour in network data, and can meet the actual operation requirements.
4.4. Comparative Experiments between the NSL-KDD Dataset and UNSW-NB15 Dataset. Experiment C: in this experiment, the proposed method is tested and compared in the NSL-KDD dataset and UNSW-NB15 dataset, and the performance of the proposed algorithm is estimated by using the performance metrics, namely, precision, recall, and F1-measure and ROC. The two datasets have different attack patterns and data characteristics, so it is impossible to compare each pattern separately, and only the overall performance index is analysed in two datasets in this paper. The overall performances of precision, recall, and F1-measure in the two databases are shown in Table 5. Figure 4 shows the comparison results of precision, recall, and F1-measure in two different databases. It is found that although the detection results of the UNSW-NB15 dataset are better than those of the NSL-KDD dataset in some values, the detection results of the NSL-KDD dataset are generally better than those of the UNSW-NB15 dataset from the whole ROC curve.

Figure 5 shows the ROC curves in two different databases. It is found that although the detection results of the UNSW-NB15 dataset are better than those of the NSL-KDD dataset in some values, the detection results of the NSL-KDD dataset are generally better than those of the UNSW-NB15 dataset from the whole ROC curve.

By comprehensively comparing the performance indexes in Figures 4 and 5, it is found that the proposed method’s technique achieves better performances for the NSL-KDD dataset. The reason is that some malicious records in the UNSW-NB15 one are not high because of the lower variances between them and normal records, and the data are optimized in the NSL-KDD database, which is more suitable for the detection of malicious records. But on the whole, it shows very good performance in the NSL-KDD dataset and UNSW-NB15 dataset, which proves the effectiveness of the proposed method in high-dimensional anomaly detection.

### Table 2: The result of two mining algorithms.

<table>
<thead>
<tr>
<th>Sample set (sample size)</th>
<th>Threshold value</th>
<th>Accuracy (%)</th>
<th>False positive rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal + DoS (63000)</td>
<td>58500</td>
<td>88</td>
<td>1</td>
</tr>
<tr>
<td>Normal + Probing (62000)</td>
<td>59000</td>
<td>95</td>
<td>2</td>
</tr>
<tr>
<td>Normal + R2L (60900)</td>
<td>59600</td>
<td>95</td>
<td>3</td>
</tr>
<tr>
<td>Normal + U2R (60052)</td>
<td>59500</td>
<td>87</td>
<td>3</td>
</tr>
</tbody>
</table>

### Table 3: The result of two mining algorithms.

<table>
<thead>
<tr>
<th>Sample dataset</th>
<th>Number of samples (m)</th>
<th>Number of FP (n)</th>
<th>Number of MFP(l)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal + DoS</td>
<td>63000</td>
<td>23</td>
<td>4</td>
</tr>
<tr>
<td>Normal + Probing</td>
<td>62000</td>
<td>19</td>
<td>1</td>
</tr>
<tr>
<td>Normal + R2L</td>
<td>60900</td>
<td>21</td>
<td>3</td>
</tr>
<tr>
<td>Normal + U2R</td>
<td>60052</td>
<td>23</td>
<td>2</td>
</tr>
</tbody>
</table>

### Table 4: Comparison of detection rates of different algorithms.

<table>
<thead>
<tr>
<th></th>
<th>SVM</th>
<th>Intelligent DT</th>
<th>LSSVM + FRFS</th>
<th>Detection + EMSMVW</th>
<th>Proposed method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probing</td>
<td>95.42</td>
<td>99.59</td>
<td>92</td>
<td>99.1</td>
<td>95</td>
</tr>
<tr>
<td>DoS</td>
<td>94.29</td>
<td>99.2</td>
<td>95</td>
<td>99.2</td>
<td>92</td>
</tr>
<tr>
<td>R2L</td>
<td>45.34</td>
<td>50.88</td>
<td>38</td>
<td>Null</td>
<td>95</td>
</tr>
<tr>
<td>U2R</td>
<td>31.34</td>
<td>35.88</td>
<td>38</td>
<td>Null</td>
<td>87</td>
</tr>
</tbody>
</table>

4.4. Comparative Experiments between the NSL-KDD Dataset and UNSW-NB15 Dataset. Experiment C: in this experiment, the proposed method is tested and compared in the NSL-KDD dataset and UNSW-NB15 dataset, and the performance of the proposed algorithm is estimated by using the performance metrics, namely, precision, recall, and F1-measure and ROC. The two datasets have different attack patterns and data characteristics, so it is impossible to compare each pattern separately, and only the overall performance index is analysed in two datasets in this paper. The overall performances of precision, recall, and F1-measure in the two databases are shown in Table 5. Figure 4 shows the comparison results of precision, recall, and F1-measure in two different databases. It is found that although the detection results of the UNSW-NB15 dataset are better than those of the NSL-KDD dataset in some values, the detection results of the NSL-KDD dataset are generally better than those of the UNSW-NB15 dataset from the whole ROC curve.

By comprehensively comparing the performance indexes in Figures 4 and 5, it is found that the proposed method’s technique achieves better performances for the NSL-KDD dataset. The reason is that some malicious records in the UNSW-NB15 one are not high because of the lower variances between them and normal records, and the data are optimized in the NSL-KDD database, which is more suitable for the detection of malicious records. But on the whole, it shows very good performance in the NSL-KDD dataset and UNSW-NB15 dataset, which proves the effectiveness of the proposed method in high-dimensional anomaly detection.
5. Conclusions

In this paper, a high-dimensional outlier mining algorithm based on the maximum frequent pattern factor (MFPOF-OM) has been proposed by using the related technology of high-dimensional outlier mining based on frequent patterns. This work has two advantages: first, the MFPOF-OM algorithm only needs to mine the maximum frequent pattern set, which solves the problem of mining completely frequent patterns in frequent pattern outlier algorithm; second, it can greatly reduce the number of maximum frequent patterns, thus reducing the time complexity of the algorithm. Experimental results show that the proposed method is feasible, which can further reduce the time complexity while ensuring the excellent detection performance compared with the contrast algorithms.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61772450) and Hebei Province Natural Science Foundation of China (F2017203307).

References


