Behavior anomaly detection based on big data analysis of Internet of Things

Jinliang Yang1,*, Xuan Lan2, Liansheng Huang2 and Jigang Zeng2

1 School of Natural and Computer, Ganzhou Teachers College, Ganzhou, 341000, China
2 Information Center, Jiangxi College of Construction, Nanchang, 330000, China
*Corresponding author’s Email: 4615120@qq.com
*Corresponding author’s ORCID: https://orcid.org/0000-0003-4481-6821

Abstract. The technical requirements of behavior anomaly detection are higher and higher. Using the Internet of things technology combined with a variety of big data analysis algorithms, we can achieve accurate behavior anomaly detection by classifying behavior data sets to a large extent. In this paper, PLA - PRF (parallel random forest) algorithm is used to realize the behavior anomaly detection model of Internet of things integrating big data analysis. In behavior detection, the PRF algorithm and DFS algorithm are compared in the case of a different number of decision trees. The results show that, compared with DRF algorithm, PLA-PRF, SPARK MLRF(Spark Machine Learning Random Forests) and PRF algorithm perform better on the four datasets, with kappa values increased by about 3.13%, 2.56% and 1.98% respectively. In contrast, PLA-PRF algorithm has higher accuracy in the case of a small sample size. With the increase of sample size, the accuracy of behavior anomaly detection gradually decreases; because the algorithm is in subspace in the process of construction, some high pheromone features are abandoned, which makes the new spatial information of features insufficient, resulting in the decision tree training process does not learn the inherent laws of abandoned data. Compared with spark MLRF and DRF, PLA-PRF has a faster execution speed in large data sets, and with the increase of data volume, the advantage is more prominent. This is because PLA-PRF uses data reuse strategy "DRS" in the process of parallelization, which reduces the data communication overhead in a distributed environment and improves the parallelization efficiency of the algorithm.

1. Introduction

With the continuous in-depth research in big data, anomaly detection has gradually become a hot issue. Anomaly detection is essentially a classification problem of data imbalance, which is widely used in credit card anti-fraud, network intrusion detection, industrial equipment fault detection, and other production and life fields. Compared with a large number of normal data, the amount of abnormal data is less, and the expression of abnormal data is quite different from that of normal data. Due to the extreme imbalance of data, and the significant difference between the characteristics of normal data and abnormal data, the common "method" training is difficult, and it is often unable to detect anomalous data effectively.

Wang C. first proposed a scalable random forest anomaly detection algorithm based on spark in big data, in order to solve the problems of frequent iterations and low performance of random forest
algorithm when processing large-scale data [1]. The algorithm uses spark framework to implement computing tasks’ cache execution and eliminates the iterative dependence. However, the algorithm does not consider the problem of too large covariance matrix caused by too many redundant features in the process of feature transformation. Garg's is analyzed in the time domain and frequency domain of anomaly detection, and the recognition accuracy is low in use. It is challenging to identify disturbances with high similarity [2]. In 2018, park J used characteristic parameters such as average fragment interval, fragment length, peak to average ratio and frequency domain energy to recognize patterns such as foot stepping, bicycle rolling and slapping, with high recognition accuracy, but insufficient recognition ability for similar patterns [3]. Zheng W. proposes a parallel random forest algorithm WECRF based on the weighted selection of anomaly detection subspace. In the process of subspace selection, low pheromone features are given lower selection weight to reduce the probability of being selected [4]. Dymora P proposes a parallel random forest algorithm DRF based on hierarchical subspace selection in spark environment. A statistical criterion is applied to divide the feature into three parts, and then the feature subspace is constructed by hierarchical sampling [5]. These algorithms do not consider the problem that the node communication overhead is too large due to the low utilization of feature information in the process of parallel training decision tree. Therefore, in the spark environment, Parada proposed a parallel random forest algorithm SPARK MLRF for anomaly detection, designed an RDD data partition strategy, and combined it with partition sampling to reduce the data transmission operation. Still, because it uses the horizontal partition method, although the local communication frequency is reduced, the global communication overhead is not reduced, and the algorithm increases with the forest environment. As the scale K continues to increase, the amount of data communication operation during the training of decision tree will increase linearly, which does not fundamentally solve the problem of large communication overhead during the training of decision tree [6]. Although these algorithms remove some redundant features and compress the size of feature set, they do not completely solve the problem of large covariance matrix in abnormal behavior detection box. In addition, when these algorithms construct feature subspaces on computing nodes, they all select features in a uniform and random way, without considering the problem of insufficient coverage of feature information in subspace of distributed environment. Solving the problem of insufficient coverage of subspace feature information is always an important research content of parallel random forest algorithm. In order to deal with this problem, although these algorithms have improved the way of evenly selecting feature subspace, they have not completely solved the problem of insufficient feature information coverage in subspace. In addition, aiming at the above three problems, this paper proposes a parallel random forest algorithm PLA - PRF based on PLA and stratified sampling. In this paper, PLA - PRF (parallel random forest) algorithm is used to realize the behavior anomaly detection model of Internet of things integrating big data analysis. In behavior detection, the PRF algorithm and DFS algorithm are compared in the case of a different number of decision trees.

2. Internet of Things Big Data and Abnormal Behavior Detection

2.1. Internet of Things Applications

With the progress and broad application of the Internet of Things and the emergence of wireless communication and mobile technology, IoT and cloud computing have become important concepts [7]. Big data Internet of things is the specific performance of Internet of things, cloud computing and other related technologies in the Internet of things system industry. There are more and more big data Internet of things perception layer terminal access, which brings a lot of technology improvement and convenience. Simultaneously, each link of its operation terminal equipment has produced a huge amount of data, so controlling these data has become an urgent problem [8]. Data security has always been a topic of concern in the information age. Many solutions have been proposed in the face of the Internet of Things’ security problems in the cloud computing environment. Also, there is a big data Internet of things security service framework, which can be used to detect behavior anomaly detection
in the IOT cloud environment [9]. Secure data storage can be realized in IOT system integrated with the cloud. The scheme uses key and public key to encrypt. In the proposed scheme, all security operations are unloaded to nearby servers, so the processing burden is reduced [10]. It also presents a secure search scheme using authorized users to encrypt, store and share data in the cloud. When the search process is complete, validation starts, in which case the shared data is retrieved. The scheme ensures the integrity of shared data and search result data [11]. This paper proposes a security service framework of Internet of things based on multi-layer cloud architecture model. This mode can achieve effective and seamless interaction on devices provided by IOT services [12]. The model includes private cloud and public cloud to give high security to user data in cloud storage. Highly secure data is stored in the private cloud, while ordinary data is stored in the public cloud, so cloud users can easily access the public cloud [13-14]. The above scheme effectively improves the security of the Internet of things, but data encryption is a single algorithm or less identity authentication [15].

2.2. PLA - PRF Algorithm
PLA - PRF algorithm mainly includes three stages: feature transformation, subspace selection and parallel training decision tree. In the stage of feature transformation, "MFS" strategy is proposed, in which the characteristic matrix is used to decompose the equation \( n \) to reduce the calculation cost \( K \) of covariance matrix, and the transformed principal component feature \( t \) is obtained

\[
\frac{dN}{dt} = r n \left( \frac{k - N}{k} \right) 
\]

\[
N = \frac{k}{1 + e^{a - rt}}
\]

In the stage of subspace selection, "ehsca" strategy is proposed, and the error constraint formula is used to select the principal component feature \( P \) hierarchically, so as to improve the subspace feature information coverage \( (y) \)

\[
N = k + \frac{N_0 - k}{1 + \left( \frac{x}{T} \right)^p}
\]

\[
y = d + \frac{e - d}{1 + \left( \frac{x}{c} \right)^b}
\]

Parallel training decision tree stage: in spark environment, an RDD data reuse strategy "DRS" is designed, which divides RDD data vertically and stores its index in DSI (data search index) table to realize the reuse of feature information and reduce the communication overhead of nodes

\[
\ln \left( \frac{k - n}{N} \right) = a - rt
\]

3. Experimental Design

3.1. Research Objects
This paper uses PLA - PRF (parallel random forest) algorithm to realize the behavior anomaly detection model of Internet of things integrating big data analysis. In behavior detection, the PRF algorithm and DFS algorithm are compared in the case of a different number of decision trees.

3.2. Algorithm Steps
Firstly, input the original data and divide it into RDD data blocks of the same size; for the feature set of each RDD data block, call "MFS" strategy to carry out feature transformation, eliminate redundant features, obtain principal component features, and store them in each task node. According to the "ehsca" algorithm, the feature subspace is generated by hierarchical subspace selection. According to
the feature subspace, the "DRS" strategy is used to train the decision tree in parallel, and the corresponding DAG tasks are generated. Finally, the tasks in DAG are submitted to spark's task scheduler to complete all model training.

4. Behavior Anomaly Detection and Analysis of Big Data Analysis Integrated with the Internet of Things

As shown in Figure 1, when the number of decision trees is equal to 10, the average behavior anomaly detection accuracy of all comparison algorithms is low. With the increase of the number of decision trees, the average behavior anomaly detection accuracy of these algorithms gradually increases in a fluctuating manner. When the number of decision trees increases to 1300, compared with PRF and DRF algorithms, the average behavior anomaly detection accuracy of PLA-PRF algorithm is about 6.1% and 1.2% higher. When the number of decision trees reaches 1500, the accuracy of PLA-PRF algorithm is about 4.6% higher than that of SPARK MLRF. It can be seen that with the increase of decision tree size, the accuracy gain of PLA-PRF algorithm is significantly higher than that of the other three algorithms. This is because PLA-PRF algorithm uses "ehsca" strategy to improve the subspace feature information coverage, thus improving the accuracy of behavior anomaly detection. Therefore, it can be concluded that the accuracy of PLA-PRF algorithm is the highest when the forest scale is large.

To verify the accuracy of PLA-PRF algorithm, this paper uses URL data set to carry out comparative experiments. The average detection accuracy of PLA-PRF algorithm in different decision tree sizes makes a comprehensive comparison with PRF, DRF, and SPARK MLRF algorithms. The experimental results are shown in Table 1.

Table 1. Average classification accuracy under different decision tree scales

| Item   | Accuracy | Precision | Decision tree | Scale | Kappa | EHSCA |
|--------|----------|-----------|---------------|-------|-------|-------|
| PRF    | 2.48     | 3.7       | 1.95          | 2.3   | 2.25  | 1.49  |
| DRF    | 4.58     | 5.76      | 2.67          | 5.77  | 3.89  | 5.78  |
| Spark  | 1.06     | 2.32      | 2.28          | 4.84  | 5.67  | 2.78  |
| MLRF   | 1.48     | 1.12      | 1.15          | 3.89  | 2.6   | 2.09  |

To verify the accuracy of PLA-PRF algorithm in behavior anomaly detection under different data sets, this paper conducts experiments on four data sets respectively, and comprehensively compares them...
with PRF, DRF algorithm and SPARK MLRF algorithm according to kappa value. The comparison results are shown in Figure 2.

**Figure 2.** Comparison of algorithms on different data sets

Compared with DRF algorithm, PLA-PRF, SPARK MLRF and PRF algorithm perform better on the four datasets, with kappa values increased by about 3.13%, 2.56% and 1.98%, respectively. However, PLA-PRF algorithm has higher accuracy in the case of small sample size, and with the increase of sample size, the accuracy of behavior anomaly detection gradually decreases, because the algorithm is in the process of subspace construction some high pheromone features are abandoned, which makes the new spatial information of features insufficient, resulting in the decision tree training process does not learn the inherent laws of abandoned data. From URL, games and words, it can be seen that the performance of SPARK MLRF algorithm is slightly better than that of PLA-PRF and PRF algorithm on the data set with fewer features and lower complexity. On the data set with more features, the PLA-PRF algorithm's average accuracy is 3.1% and 5.9% higher than that of SPARK MLRF and PRF, respectively. When the sample size reaches 65000, the accuracy of PLA-PRF is 8.9% higher and 15.6%.

**Table 2.** PRF algorithm subspace construction

| Item          | Kappa value | PRF  | DRF  | Spark-MLRF |
|---------------|-------------|------|------|-------------|
| URL dataset   | 0.51        | 0.82 | 0    | 0.09        |
| Games         | 3.82        | 3.67 | 1.44 | 3.64        |
| Words         | 4.61        | 4.83 | 2.56 | 4.44        |
| Gas           | 4.61        | 3.57 | 1.36 | 3.08        |
| Coefficient graph | 4.69    | 4.52 | 3.87 | 4.72        |

As shown in Table 2, the dimension reduction algorithm is used in the subspace construction process of PRF algorithm. Although the main component features are retained, the hierarchical extraction is not carried out in the subsequent subspace selection, resulting in uneven feature subspace information. In the face of large data sets with more features, with the increase of 10 times of data partition, the lack of information in the training subset is more and more serious, and the PRF algorithm is more effective, and the accuracy of abnormal behavior detection is reduced. However, spark PRF algorithm
solves the problem of missing subspace information of PRF to a certain extent through hierarchical extraction, because it samples each partition separately, with the increase of the size of the data set, the proportion of random selection of the data set increases, and the accuracy of behavior anomaly detection decreases; while PLA-PRF algorithm uses "ehsca" strategy for subspace selection, which fully ensures the accuracy of behavior anomaly detection. The information coverage of feature subspace is improved, and the accuracy of behavior anomaly detection is improved. In conclusion, PLA-PRF algorithm is suitable for large-scale and complex behavior data sets.

5. Conclusions
In order to solve the problem of low accuracy and low parallelization efficiency of traditional random forest algorithm in big data environment, this paper proposes a parallel random forest algorithm based on PLA and hierarchical subspace sampling. To solve the problem that the covariance matrix of parallel random forest algorithm is too large in the process of feature transformation, a feature transformation method based on PLA is proposed. Then, for the obtained principal component features, an error constrained hierarchical subspace construction algorithm is proposed to optimize the feature subspace and solve the problem of insufficient feature information coverage in the subspace. Finally, in the parallelizing training decision tree in spark environment, a data reuse technology is designed, which effectively reduces the communication overhead of nodes in a distributed environment and improves the parallel efficiency. Although PLA-PRF algorithm has made some progress in the accuracy and parallel efficiency of behavior anomaly detection, there is still room for improvement. Whether the swarm intelligence algorithm can be appropriately introduced to optimize the original feature set and improve the feature set's optimization accuracy is the next problem to be solved.

References
[1] Wang C. IoT anomaly detection method in intelligent manufacturing industry based on trusted evaluation[J]. The International Journal of Advanced Manufacturing Technology, 2020, 107(3): 993-1005.
[2] Garg S, Kaur K, Kumar N, et al. A Hybrid Deep Learning-Based Model for Anomaly Detection in Cloud Datacenter Networks[J]. Network and Service Management, IEEE Transactions on, 2019, 16(3):924-935.
[3] Park J, Cho W, Kim K. Anomaly Detection Analysis using Repository based on Inverted Index[J]. Journal of KIISE, 2018, 45(3):294-302.
[4] Zheng W, Wu Q, Zhang M, et al. Disruption predictor based on neural network and anomaly detection on J-TEXT[J]. Plasma Physics and Controlled Fusion, 2020, 62(4):12-13.
[5] Dymora P, Mazurek M. Anomaly Detection in IoT Communication Network Based on Spectral Analysis and Hurst Exponent[J]. Applied Sciences, 2019, 9(24):5319-5320.
[6] Parada, Raúl, Melià-Seguí, Joan, Pous R. Anomaly Detection Using RFID-Based Information Management in an IoT Context[J]. Journal of Organizational & End User Computing, 2018, 30(3):1-23.
[7] Al-Hawawreh M, Moustafa N, Sitnikova E. Identification of malicious activities in industrial Internet of things based on deep learning models[J]. Journal of Information Security & Applications, 2018, 41(8):1-11.
[8] Elmenshawy D, Helmy W, El-Tazi N. A Clustering based Approach for Contextual Anomaly Detection in the Internet of Things[J]. Journal of computer sciences, 2019, 15(8):1195-1202.
[9] Deng X. Research on the Anomaly Detection Method in Intelligent Patrol Based on Big Data Analysis[J]. Journal of Computer and Communications, 2019, 07(8):1-7.
[10] Zhang H, Luo Y, Yu Q, et al. A Framework of Abnormal Behavior Detection and Classification Based on Big Trajectory Data for Mobile Networks[J]. Security and Communication Networks, 2020, 2020(4):1-15.
[11] Zhang, Yu, Zhu, et al. Anomaly Detection Based on Mining Six Local Data Features and BP Neural Network[J]. Symmetry, 2019, 11(4):571-571.
[12] Maya S, Ueno K, Nishikawa T. dLSTM: a new approach for anomaly detection using deep learning with delayed prediction[J]. International Journal of Data Science and Analytics, 2019, 8(2):137-164.

[13] Shen M, Jiang X, Sun T. Anomaly Detection based on Nearest Neighbor Search with Locality-Sensitive B-tree[J]. Neurocomputing, 2018, 289(10):55-67.

[14] Sarmadi H, Karamodin A. A novel anomaly detection method based on adaptive Mahalanobis-squared distance and one-class kNN rule for structural health monitoring under environmental effects[J]. Mechanical systems and signal processing, 2020, 140(6):1-24.

[15] Tang Z, Chen Z, Bao Y, et al. Convolutional neural network-based data anomaly detection method using multiple information for structural health monitoring[J]. Structural Control & Health Monitoring, 2019, 26(1):1-22.