Research on Attributes Reduction Method of Intrusion Detection Data Based on Rough Set Theory

Xueliang Fu*, Yaodan Zhang, Honghui Li and Yumeng Hu
College of Computer and Information Engineering, Inner Mongolia Agriculture University, China

*Corresponding author email: fuxl@imau.edu.cn

Abstract. With the huge amount of network connection data, high data dimensions and complex attack types, it is difficult for traditional data analysis techniques to obtain satisfactory results in the Network Security Intrusion Detection System. This paper proposed a RSDB method based on the Rough Set Theory and the Data Deblock Thought, which can help us to reduce the data conditional attributes and remix the data. The method considers two factors: the relationship between conditional attributes and decision attribute, and the data integrity. Through the experiment, this method can solve these difficult above problems effetely, while ensuring the original data is not distorted. Through many experiments and their experimental results, RSDB can not only make the classification accuracy over than 90%, but also its average detection time is in the millisecond level.

Keywords: Big data; Data mining; Rough set; NSL KDD; Intrusion detection.

1. Introduction
Along with the development of Internet of Things, Cloud Computing, Big Data, Artificial Intelligence and 5G communication technology, the network information security has become a hot issue, and the network intrusion detection technology has also become the focus of experts and scholars. From the data level, the security of big data itself involves the collection, transmission, storage, processing, exchange, destruction and other links [1]. Each of these links faces different degrees and types of threats, so it is necessary to take corresponding and active security measures. From the system level, in order to make sure the big data system works stably, safely and reliably, and prevent the big data system from being destroyed, infiltrated or used illegally, it is a basic requirement to establish a solid, meticulous, and robust protection system. From the service level, standardizing the content of big data security service, improving the risk identification ability of big data security, establishing a sound big data security guarantee system, reducing the hidden danger of big data security and the frequency of security events are the premise and guarantee in the field of big data security. And the most effective part of big data technology in the safe life cycle is the early warning and detection, and they can provide positive and reliable guidance for protection and response work [1]. With the arrival of big data era, it provides us massive amounts of data information, and the huge data information has also made network security intrusion detection technology face unprecedented challenges. At present, network attack presents the characteristics of large-scale, distributed and diversified [2], therefore, it also proposes higher requirements and expectations for network security intrusion detection algorithm. How to improve the detection rate, reduce the false alarm rate and improve the generalization ability of intrusion detection algorithm is the importance in the Network Security area.
At present, most intrusion detection algorithms focus on some kinds of machine learning or deep...
learning algorithms depth improvement. Among them, Hui XU et al. proposed a moth fire control optimization (MFO) algorithm. Through the experimental results, they found that MFO could improve the global convergence of the algorithm and avoid falling into local optimum [3]. Hui-you HUA et al. proposed an algorithm which was named after Cluster-KNN and it was compared with KNN algorithm which is traditional. Cluster-KNN has a high time efficiency in the classification stage, and it invades other similar fields in terms of accuracy, FP and FN. The detection method also has considerable advantages, but the algorithm does not identify the specific attack type in abnormal situations [4]. Zhao-jun GU et al. proposed a KNN algorithm which is based on extreme learning machine (ELM) feature map. The intrusion detection model based on ELM-KNN algorithm improves the accuracy of intrusion detection [5]. Fa-mei HE et al. proposed a clustering algorithm which is mainly according to the feature grouping under the k-means algorithm. They achieved that the intrusion detection effect of the new algorithm is better than before, and its disadvantages have not greatly improved the detection rate of U2R and R2L [6].

This paper proposed a RSDB method based on Rough Set attributes reduction, which can be used in Network Security Intrusion Detection System. First, we apply the Rough Set knowledge attributes reduction theory to reduce association and redundancy attributes, and delete redundant attributes. And then, according to the Data Deblock Thought, the original data set can be re-divided and remixed. Finally, the classic machine learning algorithm was used to classify the data subsets, which verified the effectiveness of the RSDB method.

2. Data Sets Introduction

2.1. KDD CUP 99
The KDD CUP 99 data set comes from the intrusion detection project conducted by MIT Lincoln Lab and DARPA in 1998 to simulate a real network environment in a military network [7]. The simulation attack of this experiment is divided into four attack types, namely port scan attack (Probe), denial of service attack (DOS), unauthorized local superuser privilege access (U2R), unauthorized access from remote host(R2L). And there are 41 conditional attributes and 1 decision attribute in this data set. Ge-lin ZHANG et al. used the data set to combine the NMF algorithm with the principal component analysis algorithm to optimize the initialization problem of the NMF algorithm and apply it to intrusion detection [8].

2.2. NSL KDD
Data sets used in this paper is NSL KDD [9]. It is a subset of the data obtained by the researchers through the improved KDD CUP 99, which does not contain redundant records, so the classifier does not favor more frequent records. The training set of this data set contains 125973 pieces of data (refer with: Table 1).

| Attacks | Count |
|---------|-------|
| Normal  | 67343 |
| Dos     | 45927 |
| Probe   | 11656 |

Table 1. Data distribution.

For fine-grained classification of samples from the original dataset, it contains 22 attack types and 1 normal type. For coarse-grained classification, it contains 4 attack types and 1 normal type. So, we need to deblock the original data into level 1 classification first, and then deblock into its corresponding subclasses. Therefore, according to the definition of attack type, 21 attack types are classified as 4 attack types (refer with: Table 2).
Table 2. Classification mark.

| Attacks  | Specific classification marks                                    |
|----------|------------------------------------------------------------------|
| Normal   | Normal                                                          |
| DOS      | back, land, neptune, pod, smurf, teardrop                       |
| Probe    | ipsweep, nmap, portsweep, satan                                  |
| R2L      | ftp_write, guess_passwd, imap, multihop, phf, spy, warezclient, warezmaster |
| U2R      | buffer_overflow, loadmodule, perl, rootkit                       |

Based on this data set, Jing-ming XIA et al. proposed an improved network intrusion detection method for random forest classifier. The data were divided into different clusters by Gaussian mixed model clustering algorithm, and different random forest classifiers were trained for each cluster. These trained random forest classifiers were used for network intrusion detection [10].

3. The Rough Set Theory and Application

The development of knowledge reduction theory based on rough sets has provided many effective new methods for data mining [11]. The Rough Set Theory is a data analysis theory, which was proposed by Polish mathematician Z. Pawlak in 1982 [12]. Now it has made some research achievements in many fields, such as feature extraction, decision analysis, pattern recognition, expert system, fuzzy set, neural network and so on, and established a very complete theoretical system [13]. The Rough Set Theory is based on the classification mechanisms, Z. Pawlak understood the classification as the equivalence relationship in a specific space, and thus used this equivalence relationship to divide the space. The core idea of Rough Set Theory is the process of approximating the inaccurate or uncertain knowledge by using the existing known knowledge base. It does not need to provide any prior information other than data itself when dealing with uncertain and imprecise problems, which is one of the biggest differences from other traditional methods, so the theory is more objective and reliable when describing or dealing with the uncertainty the problems.

3.1. Attributes Reduction

For improving the classification efficiency, we ought to reduce the 41 conditional attributes in the original dataset first. Data analysis found that only 4 data are non-zero values in the urgent conditional attribute, and the values of num_outbound_cmds and is_host_login conditional attributes are both 0 value, and the value of land conditional attribute is 1 value when the attack type is land. At this point, there are four conditional attributes and one decision attribute can be deleted (refer with: Table 3).

Table 3. Attributes reduction.

| Reduction categories | Value type                          |
|----------------------|-------------------------------------|
| Conditional attributes | urgent, num_outbound_cmds, is_host_login, land |
| Attack               | land                                |

At this point, using the horizontal analysis observation method, the data set is left with 36 conditional attributes and 1 decision attribute; using the longitudinal analysis observation method, the data set is left with 21 attack types and 1 normal type.

3.2. RSDB

Because of the large amount of data in the original data set, a reasonable way is needed to segment the data effectively. RSDB method is put forward to protect that the intrusion detection data is not distorted in the process of processing. The character-type conditional attributes with the widest range of values in the data set is used as the basis for dividing the data subset. A total of three character-type conditional attributes are protocol_type, service, flag in the dataset (refer with: Table 4).
Table 4. Character conditional attributes value type.

| Character conditional attributes | Value type                                                                 | Count |
|----------------------------------|-----------------------------------------------------------------------------|-------|
| protocol_type                    | TCP, UDP, ICMP                                                              | 3     |
| Service                          | aol, auth, bgp, courie, csnet_ns, etf, daytime, discard, domain, domain_u,  | 70    |
|                                  | echo, eco_i, ecr_i, efs, exec, finger, ftp, ftp_data, gopher, harvest,     |       |
|                                  | hostnames, http, http_2784, http_443, http_8001, imap4, IRC, iso_tsap,    |       |
|                                  | klogin, kshell, ldap, link, login, mtp, name, netbios_dgm, netbios_ns,    |       |
|                                  | netbios_ssn, netstat, nnsp, nntp, ntp_u, other, pm_dump, pop_2, pop_3,    |       |
|                                  | printer, private, red_i, remote_job, rje, shell, smtp, sql_net, ssh,      |       |
|                                  | sunrpc, supdup, systat, telnet, tftp_u, tim_i, time, urh_i, urp_i, uucp,  |       |
|                                  | uucp_path, vmnet, whois, X11, Z39_50                                        |       |
| Flag                             | OTH, REJ, RSTO, RSTOS0, RSTR, S0, S1, S2, S3, SF, SH                      | 11    |

From the Table 4, the maximum number of values is the service conditional attribute. So, the original data set could be deblocked which can be divided into 70 data subsets. After analyzing each subset, there is only one attack type in the data subsets which are numbered 62-70. So, we can correspond attack type through the values of these 9 service attributes directly, and they do not need to participate in the next classification work.

Through the similarity analysis of each subset, the conditional attributes and attack types which are left behind are totally consistent. So, using the RSDB method again to fuse new data subsets. Take some service attributes as an example (refer with: Table 5).

Table 5. Examples of data subset fusion.

| Data set Number | Service Value | Attack Count | Gross Count | Dimension Attributes Number | Label |
|-----------------|---------------|--------------|-------------|-----------------------------|-------|
|                 |               | Normal | Dos | Probe | U2R | R2L |             |                   |                   |
| 13               | efs           | 0     | 478 | 7     | 0   | 0   | 485         | 19 2,6-17,26,27,32|                   |
| 14               | exec          | 0     | 465 | 9     | 0   | 0   | 474         | 19 2,6-17,26,27,32|                   |
| 25               | klogin        | 0     | 425 | 8     | 0   | 0   | 433         | 19 2,6-17,26,27,32|                   |
| 26               | kshell        | 0     | 292 | 7     | 0   | 0   | 299         | 19 2,6-17,26,27,32|                   |

By the RSDB method, the attack categories and conditional attributes in each data subset are further reduced, which provides practical and effective help for the following research (refer with: Figure 1).
Analyzing Network Connection Data Packets → Attribute Reduction → Extract the character-type conditional attribute with the widest range of values → Data Set A

R-Block → Data Set B → Classify

Figure 1. Methodological framework.

3.3. Results and Discussion
The validity of the RSDB method is tested using the SVM algorithm and the KNN algorithm. There are 32 data subsets, so only partial data subsets detection result will be shown (refer with: Table 6, Table 7).

Table 6. Classification accuracy.

| Algorithm | Original | 1  | 2  | 3  | 4  | 5  |
|-----------|----------|----|----|----|----|----|
| SVM       | 54.5%    | 87.4% | 97.4% | 94.3% | 95.9% | 95.9% |
| KNN       | 61.4%    | 84.1% | 98.3% | 92.8% | 93.6% | 94.4% |

Table 7. Average detection time.

| Algorithm | Original | 1     | 2     | 3     | 4     | 5     |
|-----------|----------|-------|-------|-------|-------|-------|
| SVM       | 0.0421s  | 0.0018s  | 0.0073s  | 0.0015s  | 0.0013s  | 0.0024s  |
| KNN       | 0.0738s  | 0.0054s  | 0.0093s  | 0.0044s  | 0.0039s  | 0.0042s  |

The test results show that the accuracy of 17 data subsets is over than 90%, the accuracy of 3 data subsets is between 80% -90%, and the accuracy of the remaining 12 data subsets is less than 80%. According to the RSDB method, the classification accuracy and the average detection time are obviously better than the original data sets which is classified directly. However, there are still some data subsets with higher relative dimension and smaller data volume, which have poor classification performance and slower average detection speed, and will be further studied in the future.

4. Summary
Data dimensionality reduction is the pre-preparation of data mining algorithms. Reasonable dimensionality reduction can effectively reduce the computational cost of the later algorithms while maintaining a relatively accurate classification effect. Therefore, data dimensionality reduction has always been one of the worthies of attention in Data Mining. This paper proposes RSDB method based on the idea of Rough Set Theory, in order to test effectiveness, we use SVM and KNN algorithm to detect the validity and implement ability. It is found that most data sets perform well, but some data sets cannot be classified accurately and efficiently. How to turn such unequal data into reliable data is one of the issues that we will study later.

Acknowledgement
This research was financially supported by the National key research and development plan: integrated water ecological management and water resource smart regulation technology demonstration of typical lakes in Inner Mongolia (2019YFC0409205); Chinese Natural Science Foundations
Natural Science Foundation of Inner Mongolia Autonomous Region of China (NO.2019MS06015).

References
[1] Special Working Group on Big Data Security Standards of the National Technical Committee for Standardization of Information Security. Big Data Security Standardization White Paper (2018) [M]. Beijing: Standards Press of China, 2018
[2] Wang Youjin. Application of data mining in Intrusion Detection System[D]. Jilin University, 2017.
[3] Xu Hui, Fang Ce, Liu Xiang and Ye Zhiwei. Applying an Improved Moth-Flame Optimization Algorithm to Network Intrusion Detection [J]. Journal of Computer Applications, 2018, 38(11): 3231-3235+3240.
[4] HUA Hui-you, CHEN Qi-mai, LIU Hai, ZHANG Yang, YUAN Pei-quan. Hybrid Kmeans with KNN for Network Intrusion Detection Algorithm [J]. Computer Science, 2016, 43(03): 158-162.
[5] GU Zhao-jun, LI Bing, LIU Tao. Network intrusion detection based on ELM-KNN algorithm [J]. Computer Engineering And Design, 2018, 39(08): 2412-2416+2421.
[6] HE Famei, MA Huizhen, WANG Xuren, FENG Anran2. Research on anomaly intrusion detection system based on feature grouping clustering [J/OL]. Computer Engineering: 1-7 [2020-03-10].
https://doi.org/10.19678/j.issn.1000-3428.0054476.
[7] Hettich S, Bay S. KDD cup 1999 data [EB/OL]. http://kdd.ics.uci.edu/data bases/kddcup99/kddcup99.html
[8] ZHANG Gelin, LI Yong. Non-negative Matrix Factorization Optimization and Its Application in Network Intrusion Detection [J]. Netinfo Security, 2018(08): 73-78.
[9] TAVALAAEE M, BAGHERIE, LU W. A detailed analysis of the KDD CUP99 data set [C] // Proceedings of the 2nd IEEE Symposium on Computational Intelligence for Security and Defense Applications. Washington D.C., USA: IEEE Press, 2009.
[10] XIA jing-ming, LI Chong, TAN Ling, ZHOU Gang. Improved random forest classifier network intrusion detection method [J]. Computer Engineering And Design, 2019, 40(08): 2146-2150.
[11] WANG Guo-Yin, YAO Yi-Yu, YU Hong. A Survey on The Rough Set Theory and Applications [J]. Chinese Journal Of Computers, 2009, 32(07): 1229-1246.
[12] PAWLAK Z. Rough sets [J]. Inter J of Computer and Information Sciences, 1982, 11 (2): 341-356.
[13] WANG Guo-Yin, The Rough Set Theory and Knowledge Acquisition [M], Xi'an Jiaotong University Press, 2001: 23-31, 117-146.