Research of Intrusion Detection Algorithm Based on Parallel SVM on Spark

Hongbing Wang, Youan Xiao and Yihong Long
School of Information Engineering
Key Laboratory of Fiber Optic Sensing Technology and Information Processing (Wuhan University of Technology), Ministry of Education
Wuhan University of Technology
Wuhan, Hubei Province, China
whbing@whut.edu.cn

Abstract—**As network intrusion data's scale gets larger and larger, designing parallel schemes for intrusion detection have been becoming research focus in the field of information security. In order to solve the problem that the intrusion detection algorithm is high time-consuming, the classification of large amounts of data occupies lots of memory and the efficiency of single detection is low, a parallel principal component analysis (PCA) combined support vector machine (SVM) algorithm based on spark platform is proposed (SP-PCA-SVM) in this paper. This method adopts the way of principal component analysis (PCA) for training and predicting data, and then introduces a fusion of Bagging integration strategy and SVM algorithm, and finally uses spark distributed framework to achieve. The results show that, for a large number of intrusion data, the new parallel training method, to a certain extent, reduces the training time and improves model learning efficiency.**

Keywords-intrusion detection; SVM; spark; Parallel Computing; PCA

I. INTRODUCTION

Intrusion detection works by means of the detector to separate the normal network data and abnormal data. The classifications of the detector algorithm emerge in endlessly, such as association rules, neural network and genetic algorithm, et al [1]. However, the efficiency and accuracy of detection of these methods are very related to the number of samples and fluctuating. The Support Vector Machine (SVM) is a classical classification algorithm that has a wide range of application [2]. As a kind of machine learning method with complete theoretical guidance and excellent experimental performance, SVM can solve the problem of small sample, nonlinear and high dimension through the principle of risk minimization, and has good ability of generalization, which is very suitable for network intrusion Detection [3].

SVM is powerful classification and regression tool. So far, there have been some SVM models, such as Sequential Minimal Optimization (SMO), libSVM, lightSVM, et al. However, using directly of them is not suitable for handling large-scale data sets. When the size of the training sample becomes larger, the memory and time occupied by the training of SVM algorithm increase dramatically [4]. Besides, single SVM algorithm can't effectively deal with large-scale data sets. To solve the problem of insufficient for SVM to handle large amounts of data, the current resolution strategy is roughly divided into parallel SVM algorithm [5],[6] or adopting the strategy of divide and conquer to narrow the data [7]. For example, Zhanquan Sun [8] from Key Laboratory for Computer Network of Shandong Province, and Geoffrey Fox of Indiana University proposed a parallel SVM scheme based on iterative MapReduce, which has a good effect on generalized data, but it is bloated for intrusion detection data classification. Zeshen Liu and Zhisong Pan [9] proposed the improved strategy of cascading support vector machine parallel algorithm. To a certain extent, it can reduce the training time, but it is still to be improved for high dimensional data; Jianpei Zhang et al. [10] proposed a two-layer parallel SVM algorithm, which avoids influences of the classifier on performance when the distribution of the initial samples is too different, but the training efficiency of the data is not high enough. Mingyu Qi et al. [3] proposed a SVM classifier based on PCA has a good effect on data preprocessing, but because the whole process is carried out on a single SVM, the speed of the SVM classifier is greatly reduced when the data volume increases. Through these studies, it is shown that current parallel SVM algorithm or other improved SVM algorithm still has room for improvement in data training speed.

Aiming at the high time-consuming problem of parallel training data, this paper proposes a combination of SVM parallel integration strategy and PCA dimensionality reduction based on Bagging, and implements it on spark parallel computing platform. It achieves the optimization purposes not only eliminating the bottleneck which cannot deal with large-scale data sets, but also improving the efficiency of high-dimensional intrusion data.

II. RELATED TECHNOLOGIES

A. Intrusion Detection and Support Vector Machine

Design of Intrusion detection analyzer is essentially to determine a discriminant function $f: R^n \rightarrow \{-1,+1\}$. It can divide input data set $D=\{x_i \mid i=1,2,\cdots l\}(x_i \in R^n)$ into two

This work is supported by National Science & Technology Pillar Program (No. 2014BAH26F03).

categories and results of the categories are recorded as $y_i \in \{+1,-1\}$,where l is the number of samples and n is the dimension. If y_i = 1, it indicates that the corresponding sample is a normal sample; if y_i = -1, it indicates that the corresponding sample is an abnormal sample. That is:

$$f(x_i)=\begin{cases}+1, & x_i \textit{ is a normal data}\\ -1, & x_i \textit{ is a abnormal data}\end{cases} \tag{1}$$

So, intrusion detection can be deemed to a standard classification problem. system.

Support Vector Machine (SVM) algorithm is a learning algorithm based on principle of VC dimension learning and structural risk minimization, which seeks the best compromise between model complexity and learning ability based on limited sample information [6]. SVM has become an important algorithm to research the intrusion detection system.

Given the set of training samples, in the case of linear separable, then:

$$\begin{cases}(\omega \bullet x_i)+b \geq 0 \; y_i=+1\\ (\omega \bullet x_i)+b \leq 0, y_i=-1\end{cases} \tag{2}$$

The problem of classification is transformed into the optimal hyperplane problem by quadratic programming, The following formula:

$$\begin{cases}\min(\frac{1}{2}\|\omega\|^2)\\ s.t.\; y_i((\omega \bullet x_i)+b)\geq 1, i=1,\cdots,l\end{cases} \tag{3}$$

The linearly separable optimal hyperplane is shown in Fig. 1. The classification interval is Δ ; H_0 is the “optimal class hyperplane”; H_1 , H_2 are the “interval hyperplanes”; H_3 , H_4 are the “classification hyperplane” [4].

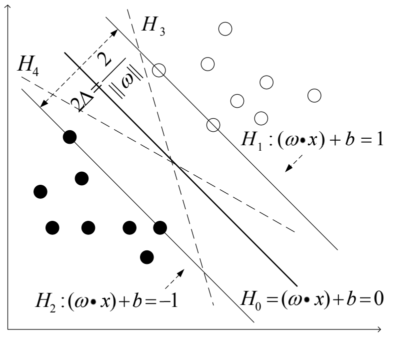
H_3
H_4
$2\Delta=\frac{2}{\|\omega\|}$
$H_1:(\omega\bullet x)+b=1$
$H_2:(\omega\bullet x)+b=-1$
$H_0=(\omega\bullet x)+b=0$

Figure 1. Linear separable optimal hyperplane

In the case of linear non-separable, by introducing nonnegative relaxation factor $\varepsilon \geq 0$ and castigate factor C to allow the existence of misclassified samples [11], the constraints become:

$$\begin{cases}\min(\frac{1}{2}\|\omega\|^2)+C\sum_{i=1}^{l}\xi_i\\ s.t.\; y_i((\omega \bullet x_i)+b)\geq 1-\xi_i, \xi_i \geq 0, i=1,\cdots,l\end{cases} \tag{4}$$

Considering the classification problem of complex data, SVM uses kernel representation to transform the original linear non-separable problem into linear separable problem of high dimensional feature space, which greatly reduces the computational cost of nonlinear transformation. By mapping $\phi: X \rightarrow F$, this means that nonlinear classification is divided into two steps:

- Transform the data into the feature space with the map.
- classify of feature space.

B. Bagging Integration Strategy

Integrated learning is a kind of learning methods that refers to using a set of learning machines to improve the performance of the entire system. Bagging is one of the classic integration algorithms[12], and a parallel integrated learning method. In this method, the classifier is composed of several sub-classifiers. The training set of each sub-classifier is randomly selected from the original training set, and the training samples are allowed to be selected repeatedly.

III. Design and Implementation of Parallel SVM Algorithm on spark

In the intrusion detection environment, the single detector is liable to malfunction because of the attack (the single point of failure problem), and for the data of high dimension and high feature, learning algorithm is inefficient. In view of the above problems, this paper proposes a multi-detection distributed parallel detection scheme based on the method of integrated learning, named SP-PCA-SVM.

A. Improved Bagging-based SVM Integration Strategy

For classification problems, the performance of the classifier is very dependent on the sample of the study, and the sample of useless information, such as redundancy, noise or unreliable information, will weaken the ability of the classifier. In intrusion detection, intrusion behavior information is often concentrated only in some features, such as Dos and Probe type of intrusion is mainly related to the traffic attributes, U2R and R2L mainly related to the content attributes. Redundant features for learning algorithms also have an impact. With the increase of the uncorrelated feature, the learning problem is not easy to describe, the classification accuracy will be greatly reduced, the speed of the learning algorithm will be affected. In order to reduce the interference of a large number of redundant information, and to ensure efficient and parallel way, the design of classifier in this paper is under the premise of Bagging, and dimensionality processing part is added.

For the given training sample X , training and predicting are as the following steps. For the training phase, first, K samples are obtained from the training sample set, denoted as TR_i ($i=1,2,...,k$); For each sample set, it will be analyzed by the Principal Component Analysis (PCA) to feature extract. That is to say, the training feature is reduced dimension, and the input feature space is transformed into a new feature space, where the obtained feature number is less than that in the original space and can reflect the original characteristics of the most useful information; Then, the data processed by PCA are trained as a training set to obtain the classifier $PCA\,SVM.i$ ($i=1,2,...,k$). For the predicting phase, first, the input predicting samples are reduced dimension by PCA, and then, the resulting data are classified by classifier $PCA\ SVM.i$ ($i=1,2,...,k$); Finally, the majority voting method is used to obtain more reliable result of the classification. The improved integration strategy is shown in Fig. 2, and its algorithm pseudocode is shown in Algorithm 1.

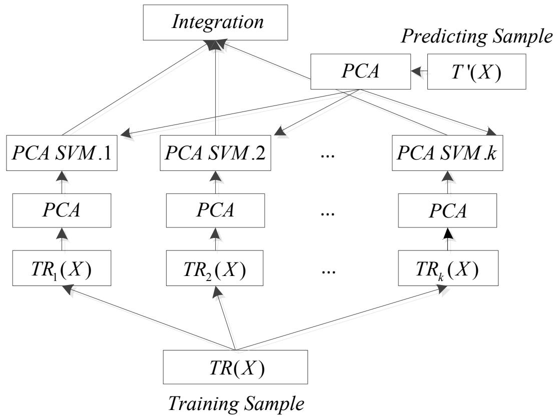

Figure 2. Improved Bagging-based SVM integration

```
Algorithm 1: Improved Bagging-based SVM
Input: training set $TR=\{(x_i,y_i)\,|\,i=1,2,\cdots,l\}$, integrated scale $K$.
Training phase:
1 Initialize parameters $SVM=\varnothing$;
2 $For\ i=1,\cdots K$
      obtain samples $TR_i$ from the training set $(x_p,y_p)$;
3 Calculate the mean $\mu=\frac{1}{l}\sum_{i=1}^{l}x_i$, covariance $\Sigma=\frac{1}{l}\sum_{i=1}^{l}(x_i-\mu)(x_i-\mu)^{\mathrm{T}}$,
  the unit eigenvector of the $\Sigma$ eigenvalue are $e_1,e_2,\cdots,e_p$, (p is the
  dimension), then, use the above calculation results to map the sample set to the
  main component space $\tilde{R}^p$, that is $\tilde{x}_i=e_i^{\mathrm{T}}x$, to get $PCA$ sampling set
  $PCA-TR_i$;
4 Use $PCA-TR_i$ as the training set to get the classifier
  $PCA\ SVM.i\ (i=1,2,\cdots K)$;
predicting phase:
1 For predicting sample $T'$, referring to step 3 of the training phase, to get
  $PCA$ sampling set $PCA-T'$;
2 Classify the $PCA-T'$ with classifier $PCA\ SVM.i$;
3 Obtain the predicting result by adopting majority voting method.
Output: $(\tilde{x}_i,\tilde{y}_i)$, $accuracy={}^{count(\tilde{y}_i=y_i)}\!/_{count(y_i)}$, time= endTime-
startTime
```

B. Parallel SVM Implementation on spark

According to the description of the Improved Bagging-based SVM algorithm and combine the parallel programming model of spark platform, the parallel SVM implementation scheme under the spark environment is designed (SP-PCA-SVM). Before the model training begins, upload the intrusion data which sets to be trained to the HDFS distributed file storage system. The task scheduling of the Spark cluster divides the dataset into K pieces. Each piece creates a new task in the Executor, and allocates computing resources. And then PCA data processing on the spark cluster and SVM training in parallel until the training is completed to obtain K pieces of model. Each model is used to predict dataset to be tested and finally the predicted results are combined by voting. The flow chart is shown in Fig. 3 and Fig. 4.

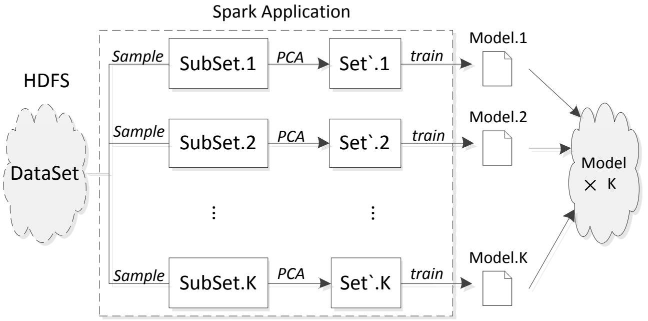

Figure 3. Parallel SVM Training Process on Spark

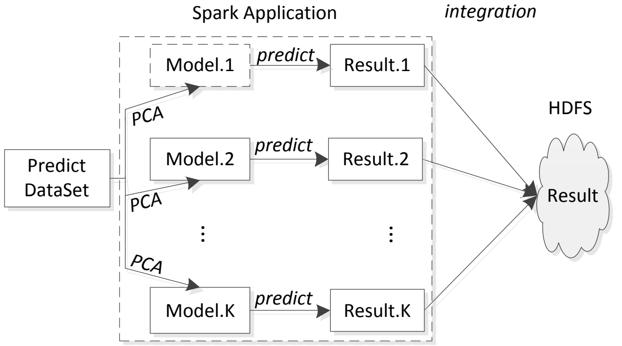

Figure 4. Parallel SVM Predicting Process on Spark

IV. EXPERIMENT AND ANALYSIS

The Spark cluster used in this experiment is composed of five physical machines, including a Master node and four Slave nodes. The task scheduling mode for the cluster is standalone. Each machine's configuration is: 2 core CPU, 4GB memory, 118G of storage space. Operating system environment is Centos 6.3. Hadoop version is 2.7.1 and Spark version is 2.0.0. Spark MLlib SVMWithSGD is selected for SVM.

The experimental data sets are 10% of KDD CUP99. Each connection consists of 41 tagged features and a markup that is normal or not. Specific features are not described here, see KDD official website. These attacks are broadly classified into four categories: 1) Scan attack (Probe); 2) Denial of Service attacks (Dos); 3) Unauthorized use of local super privilege attack (U2R); 4) Remote user unauthorized access attacks (R2L). This paper tests the four types of data separately. The experiment was carried out in three groups: single PCA-SVM test, parallel SVM test and parallel SP-PCA-SVM test. The compositions of the experimental data are shown in Tab. Ⅰ.

The pre-processed samples are stored in HDFS and the training and predicting time-consuming and predicting accuracy of the SP-PCA_SVM is obtained as shown in Fig. 3 and Fig. 4. In the meanwhile, the experiment of the parallel SVM was implemented referencing reference 2. For the single

PCA-SVM scheme in reference 3, the experimental comparison is also given. Among them, the parameter of PCA is referred to reference 3 and a number of experimental options. The above experiments are performed five times, and the average experimental results are shown in Tab. Ⅱ.

TABLE I. COMPOSITION OF EXPERIMENTAL SAMPLE

Category	Number of training samples	Proportion of invasive samples	Number of predicting samples
Probe	8000	50%	54000
Dos	8000	50%	54000
U2R	1000	5%	2050
R2L	2400	42%	11000

TABLE II. COMPARISON OF EXPERIMENTAL RESULTS

Category	Single PCA-SVM		Parallel SVM		SP-PCA-SVM	
	Accuracy	*Time*	*Accuracy*	*Time*	*Accuracy*	*Time*
Probe	92.28%	155.6s	93.56%	46.4s	93.40%	35.8s
Dos	86.31%	708.5s	87.72%	188.3s	90.24%	127.6s
U2R	96.4%	855.4s	96.2%	191.0s	96.7%	136.2s
R2L	88.1%	612.8s	87.3%	150.9s	89.6%	109.1s

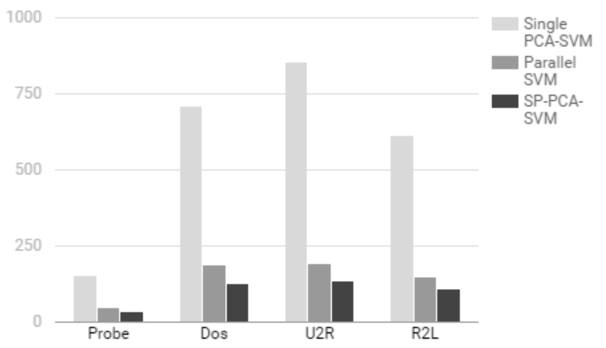

Figure 5. The time-consuming of the three schemes

Experiments show that the accuracies of classifications of the three schemes are basically the same, but SP-PCA-SVM training and predicting time is significantly shorter than the other schemes. Single PCA-SVM has the longest training and predicting time, which is about 5 times as long as that of SP-PCA-SVM. For four types of training and predicting, the time of SP-PCA-SVM is lower than that of parallel SVM because the data are processed by the PCA. The training and predicting time of the classifier will be shortened and the efficiency is improved about 20% compared to the parallel SVM. The time-consuming histogram of the three schemes are shown in Fig. 5. In general, the SP-PCA-SVM scheme effectively reduces the classification time without substantially affecting the classification accuracy.

CONCLUSION

There are numerous changes for network attack, and the Support Vector Machine has unique advantages in learning and pattern recognition under the small sample, which makes it have a wide application prospect in the field of intrusion detection. It is time-consuming to use only SVM to train large amounts of data. In this paper, PCA is used to reduce the sample data, then parallelize SVMs, and finally the scheme is implemented on spark platform. The experimental results show that the proposed scheme can shorten the classification time of the classifier under the condition that the accuracy rate is not significantly reduced. At the same time, we should also see the problem of the method that when the scale of data is not large, the optimization effect is not obvious; The parameter selection of PCA dimensionality reduction is still in the tentative stage, and the classification accuracy will be reduced when the parameter is set incorrectly. al footnote at the bottom of the column in which it was cited. Do not put footnotes in the reference list. Use letters for table footnotes.

In the next study, the training structure of the algorithm will be further optimized. A more accurate value for the dimensionality reduction parameter will be discussed. In general, an effective optimization scheme is proposed in this paper for intrusion detection problem with SVM.

ACKNOWLEDGMENT

This work is supported by National Science & Technology Pillar Program (No. 2014BAH26F03).

REFERENCES

[1] YANG Zhi-jun, TIAN Di, MA Jun-xiao, SUI Xin, and ZHOU Bin, “Survey of intrusion detection technology,” Computer Engineering and Design, vol. 27, No. 12, pp. 2119–2123, June 2006.

[2] Liu, Chang, et al. “Multiple submodels parallel support vector machine on spark,” Big Data (Big Data), 2016 IEEE International Conference on. IEEE, 2016.

[3] QI Ming - yu, LIU Ming, YU Fu – ming, “Research on SVM Network Intrusion Detection Based on PCA”, Information Network Security 2 (2015): 15-18.

[4] Nguyen, Tu Dinh, et al. “Distributed data augmented support vector machine on spark,” Pattern Recognition (ICPR), 2016 23rd International Conference on. IEEE, 2016.

[5] Zuo Xiaojun, Dong Lizhong, Qu Wu, “Distributed intrusion detection approach based on the Spark framework,” Computer Engineering and Design, vol. 36, No. 7, pp. 1720-1726, July 2015.

[6] Zhang Pengxiang, Liu Limin, Ma Zhiqiang, “Research on Cascade-Grouping parallel SVM algorithm based on MapReduce”, Computer Applications and Software, vol. 32, No. 3, pp. 172-176, Mar. 2015.

[7] YE Fei, LUO Jing-qing, YU Zhi-fu. “An Improved Parallel Processing SVM Learning Algorithm,” Microelectronics and Computers, vol. 26, No. 2, pp. 40-43, 2009.

[8] Sun, Zhanquan, and Geoffrey Fox. “Study on parallel SVM based on MapReduce,” Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012.

[9] LIU Ze-shen, PAN Zhi-song, “Research on Parallel SVM Algorithm Based on Spark”, Computer Science 43.5 (2016): 238-242.

[10] Zhang jianpei, Cheng Lili, and Ma Jun, “A Network Intrusion Detection Method Based on Parallel Support Vector Machine,” Computer Engineering and Applications, vol. 43, No. 4, pp. 137-139, 2007.

[11] ZHANG Xue-qin, GU Chun-hua and WU Ji-yi, “Fast intrusion detection algorithm based on reduced support vector machine,” Journal of South China University of Technology (Natural Science Edition), vol. 29, No. 2, pp. 108-112, Feb. 2011.

[12] He Ming, LI Guo-zheng, et al, “Bagging Integrated Learning Method Based on Principal Component Analysis”, Journal of Shanghai University (Natural Science Edition), vol. 12, No. 4, pp. 415-418, 2016.