Research on SQL injection detection technology based on SVM

Zhuang Chen¹, Min Guo², Lin Zhou³

¹Software Engineering School, Chongqing University of Posts and Telecommunication, Chongqing, China

²Software Engineering School, Chongqing University of Posts and Telecommunication, Chongqing, China

³Software Engineering School, Chongqing University of Posts and Telecommunication, Chongqing, China

Abstract. SQL injection, which has the characteristics of great harm and fast variation, has always ranked the top of the OWASP TOP 10, which has always been a hot spot in the research of web security. In view of the difficulty of detecting unknown attacks by the existing rule matching method, a method of SQL injection detection based on machine learning is proposed. And the author analyses the method of SQL injection feature extraction, finally, the word2vec method is selected to process the text data of the HTTP request, which can effectively represent the SQL injection features containing the attack payload. Training and classification of processed samples with SVM algorithm, The experiment shows that this method effectively solves the problem of SQL injection to the mutation and the high leakage rate of the rule matching. By comparing with the classification results of statistical features, this SQL injection classification model has a higher detection rate.

1 INTRODUCTION

With the rapid development of the Internet, the web 2.0 and various new Internet products, web security has become one of the hot topics in information security research in recent years. web system security issues have become increasingly prominent, SQL injection is one of the most damaging vulnerabilities in web security. In Owasp Top Ten Project [1], SQL injection was at the top of the list with its great harmfulness and rapid mutation. "2016 China Website Security Vulnerability Analysis Report" [2] shows that the SQL injection ranked third in the Web vulnerability, you can see a very high degree of harmfulness, which was released by 360. The existing methods of SQL injection detection are mostly based on the existing security knowledge, using rule matching method to detect, this detection method is powerless to unknown attack.

In article [3], authors proposed a detection method based on parse tree, which is to detect SQL injection attack by using comparative between safety parse tree and parse tree to be detected, to make up for the shortcomings of the classical method. In article [4], authors proposed a second-order SQL injection defense model based on improved parameterization, which prevents SQL injection through filtering input, index replacement, syntax comparison and parameter replacement. In article [5], authors proposed a decision tree-based SQL injection defense model, after 1000 attack load training using machine learning, got a better classification model. In article [6], authors proposed the use of web log to detect SQL injection detection model, combined with machine learning and pattern matching in log analysis, the offline log training using Bayesian

algorithm, through the pattern matching on the SQL injection testing. Author proposed the use of web log to detect SQL injection detection model, the use of Bayesian log off-line training, through the pattern matching to detect SQL injection. In article [7], author proposes a Web attack detection technology based on SVM, extracts the features of attack request through existing security knowledge and statistical features, and uses SVM algorithm to train and classify.

According to the above literature, the existing SQL injection detection technology has the following problems: the rules matching methods are mostly based on existing security knowledge, and these methods are overly dependent on the rules so that they can't do anything about the unknown attack, resulting in underreporting Case. However, the methods of machine learning and the process of feature extraction also need safety knowledge, which needs to be manually extracted from a large amount of data to carry the characteristics of the original attack load.

In this paper, a machine learning based detection algorithm is proposed, which uses the text expression method of word vector, processing HTTP requests with word2vec features, and classifying SQL injection with SVM classification model, which achieves very good classification results.

2 SQL INJECTION

2.1 The concept of SQL injection

SQL injection is a vulnerability of web application, which is defined by OWASP top 10: When sending

untrusted data to command or a part of query, it will generate injection defects of SQL injection, NoSQL, OS injection and LDAP injection. The attacker's malicious data can induce the parser to execute unintended commands or access data without appropriate authorization. In fact, an attacker inserts a malicious SQL script into a web page and directly into the database to query it.

2.2 Characteristics of SQL injection

The SQL injection attack has the following features [8]:

(1) Universality. As long as it is a Web application that uses SQL syntax, it is easy to generate SQL injection without any processing of the input.

(2) Technical difficulty is not high. The attack process is simple. At present, many SQL injection tools are popular on the Internet. By using these tools, attackers can quickly attack or destroy the target websites.

(3) Harmfulness. Because of the defects of the web language itself, and the few developers with secure programming, most of the web application system has been SQL injection attacks, and once the attack is successful, attacker can control the whole web application system of data make any modifications or steal, damage to the extreme.

(4) Fast variation. An experienced attacker will manually adjust the parameters of the attack, the attack data is nonenumerative, which leads to the traditional feature matching method can only recognize the very few attacks. Or the most conventional attack, which is difficult to prevent.

2.3 The principle and harm of SQL injection

SQL injection is only related to the database, the principle is that the related parameters accepted are not processed directly into the database query operation, For example, there are the following query statements in a typical login form:

Select * from user where username= 'admin' and password = '123'

If the parameter admin is constructed in the form:

Admin 'or 1=1 -- , then the query statement becomes:

Select * from user where username= 'admin' or 1=1 -and password = '123'

Because in the database, '--' means annotations. So the condition of judgment is always true, and this will bypass the login.

The harm of SQL injection is great, which can cause the following harm:

(1) Unauthorized operation of data in the database.
(2) Maliciously tamper with the content of the web page.

(3) Get web shell to raise permissions, and so on.

The following figure is the result of using the sqlmap to inject a web page with a SQL injection vulnerability, which directly exploits the current database.

```
[17:23:44] [INFO] fetching database names
available databases [3]:
[*] information_schema
[*] news
[*] test
```

Fig.1. The results of sqlmap injection

3 Support vector machine

SVM is one of the most widely used algorithms in the field of machine learning, Generally used for classification. it is especially suitable for the analysis of the problem of non-black and white in the field of security.

Fig. 2. SVM classification schematic

Sql injection detection is actually problem of two classification. As shown in Figure 1, assuming only two dimensional feature vectors, we need to solve a classification problem, need to distinguish between normal users and hackers. If it is true that it can be distinguished by a straight line, then this problem is called linear separable, if not linear, it is called linear inseparable. The simplest case is discussed, assuming that the classification problem can be linearly differentiated, the differentiated line is called hyperplane, and the problem is transformed to find the best hyperplane. In figure 2, the hyperplane can be described by following:

$$wx + b = 0 \quad (1)$$

It is recorded as hyperplane:

$$(w, b) \quad (2)$$

The distance from any point x to the hyperplane in the sample space can be written as:

$$r = \frac{|wx + b|}{\|w\|} \quad (3)$$

Suppose the hyperplane (w, b) can classify the training samples correctly, That for $(x_i, y_i) \in D$,

$$\text{if } y_i = +1 \quad (4)$$

$$\text{then } wx_i + b > 0 \quad (5)$$

$$\text{if } y_i = -1 \quad (6)$$

$$\text{then } wx_i + b < 0 \quad (7)$$

$$\text{order:} \quad \begin{cases} wx_i + b \geq +1, y_i = +1 \\ wx_i + b \leq -1, y_i = -1 \end{cases} \quad (8)$$

The nearest sample of the distance hyperplane is called the support vector. The sum of the distance between two dissimilar support vectors to the hyperplane is a:

$$\gamma = \frac{2}{\|w\|} \quad (9)$$

It is called the margin.

To find the best hyperplane with "the maximum margin", It is to find the constraint parameters w and b which can satisfy the formula (9), so that the γ is maximum, that is:

$$\begin{aligned} & \max_{w,b} \frac{2}{\|w\|} \\ & s.t. \ y_i(wx_i + b) \geq 1, \quad i = 1, 2, \dots, m. \end{aligned} \quad (10)$$

Obviously, in order to maximize the margin, only the maximization of $\|w\|^{-1}$, which is equivalent to the minimization of $\|w\|^2$, then:

$$\begin{aligned} & \min_{w,b} \frac{1}{2} \|w\|^2 \\ & s.t \ y_i(w^T x_i + b) \geq 1, \quad i = 1, 2, \dots, m. \end{aligned} \quad (11)$$

This is the support vector machine.

4 SQL injection detection algorithm based on SVM

4.1 Flow chart of the algorithm

SVM is a supervised learning algorithm, So the flow chart of this algorithm is as follows:

Fig. 3. Algorithm flowchart

4.2 Participle

The dataset is composed of 2 parts, black samples and white samples. It includes 1000 samples of SQL injected in GitHub and exploit-db website, and 1000 samples of normal HTTP request as negative examples. According to the rules of SQL injection, it is generally only in the request part, so only the request part of the sample is retained.

The sample data for SQL injection is as follows:

```
page=2%25%27%20UNION%20ALL%20SELECT%20
NULL%2C%20NULL%2C%20NULL%2C%20NULL
%2C%20NULL%2C%20NULL%2C%20NULL%2C%2
0NULL%2C%20NULL%2C%20NULL%2C%20NULL
%2C%20NULL%2C%20NULL%2C%20NULL%2C%2
0NULL--%20
```

The sample data that is normally accessed is as follows:

action=http%3A//ytequocte.com/nhan-biet-dau-hieu-
trieu-chung-benh-phu-
khoa/%3Fkeyword%3Dphu%2520khoa%26matchtype%
3De%26adposition%3D1t4%26device%3Dc%26mtk%3
Dsearch%26gclid%3DCLfa6O7Pj9MCFdAHKgodr5YO
EA

In this paper, a word vector method[9] is used to classify text, So the text must be participle. By Observing the above sample data and the characteristics based SQL injection ,Using regular expressions('%','?', 'Space'...) for participle. The samples after the participle are as follows:

Positive sample: ['page=', '2', 'union', 'all', 'select',
'null', 'null', 'null', 'null', 'null', 'null', 'null', 'null',
'null', 'null', 'null', 'null', 'null', 'null']

Negative sample: ['action=', 'http://ytequocite.com', 'nhan', 'biet', 'dau', 'hieu', 'trieu', 'chung', 'benh', 'phu', 'khoa', 'keyword=', 'phu', 'khoa', 'matchtype=', 'e', 'adposition=', 'lt4', 'device=', 'c', 'mtk=', 'search', 'gclid=', 'clfa6o7pj9mcfdahkgodr5yoea']

4.3 Feature extraction

For the samples after segmentation, human can understand the meaning of each label, but the machine can't. So the next is to vectorization of the sample. Translating the text after the word segmentation into a machine learning problem. The first step is to find a way to quantify these words. The most common method is one-hot code[10]. This method is to express the word list as a very long vector, only one dimension is 1, the others are 0. For example, the word "select" is expressed as [0,0,1,0,0,0,...]. An important problem in this method is that the vectors that constitute text are extremely sparse, and words and words are independent of each other. Machine learning can't understand the semantics of words. Word2vec [11] is a way to express text features as N dimensional vectors, which is to transform words in text into dense vectors that computers can understand, so that the distances between semantically similar words in space are close.

Here we use the embedded word vector model to create a semantic model of SQL injection, so that the machine can understand tags such as “union”, “select”,

“and” and “or”. The 300 words that have the most times in the black samples are used to form a vocabulary, and the other words are marked as "None". Modeling with the word2vec of the gensim module. The following table 1 gives a part of the word list:

As you can see from the table 1, most of the words are SQL injection commonly used characters.

Table 1. Part of the word list

)	select	union	null
and	end	all	then
when	case	from	by
count	floor	group	or
extravalue	char(	order	undatexml

4.4 Vectorization and Data marking

Through the established word vector model, the space vector can be used to express a text, and then the SVM algorithm is used to classify. The marking part, because we already know which samples are SQL injection and which are normal samples. In this paper, all the positive samples are marked as 1, and the negative sample is 0.

5 Experimental results and analysis

In this paper, we use Python machine learning library scikit-learn to train text space with SQL injection and square access. We get the best result by adjusting SVM kernel function and parameters.

In order to evaluate the accuracy of classification models, we used the positive classification rate (TPR), the negative classification rate (FPR), recall rate (Recall), and the area of the ROC curve to evaluate the classifier. The following is a description of several evaluation parameters:

TPR: The rate of SQL injection or normal request to correct classification.

$$TPR = \frac{TP}{TP + FN}$$

TPR: The rate of SQL injection or normal request to error classification.

$$FPR = \frac{FP}{FP + TN}$$

Recall Rate: The rate of the predicted request to the total sample.

$$Recall\ Rate = \frac{TP}{TP + FN}$$

ROC curve : A comprehensive index of classification rate and negative classification rate, the closer to 1, the better the effect of classification.

The experiment uses three way cross validation method to mix all SQL injected samples and normal access sample data into 3 parts, of which 2 are used as training samples and one is taken as test samples. Through the experiment, the final selection of the penalty parameter $C=1$, the kernel function selection linear, the classification effect is best. Table 2 shows the classification effects of selecting statistical features and

word2vec features. The experimental results show that the word2vec features used in this paper are improved compared with the statistical characteristics.

Table 2. Classification results

Features	TPR	FPR	Recall	ROC curve area
Statistical characteristics	0.920	0.058	0.923	0.933
word2vec	0.954	0.041	0.946	0.982

Can be seen from the table 2, the comparison and statistical features, showing a high TPR and low FPR. The prediction rate reached more than 90%, because the data is limited, if there are more samples, the classification model of the effect of training will be better.

6 CONCLUSION

This paper presents a detection algorithm based on machine learning, the use of word vector text representation method, the use of word2vec features http request processing, combined with svm classification model sql injection, getting a good classification results.

This paper also proposes a SQL injection detection technology based on word2vec feature and SVM, which applied to security detection, effectively solves the problem of SQL injection mutation and overcomes the defects of the existing rule matching methods. And the use of word2vec features, without the existing security knowledge, you can effectively express the SQL injection of text features, the relative statistical characteristics of the classification has also been improved. However, there are still some problems in practical application as follows:

(1) In order to identify malicious behavior from a large number of labels and combinations of functions, the high-quality data sets Need to provided and the experiment required a large number of known sql injection samples and normal requested samples.

(2) Word2vec saves the text vector into memory. The memory required for model training depends on the size of the data set.

References

- OWASP TOP Ten. [EB/OL], (2017)
- 2016 China Website Security Vulnerability Analysis Report. [EB/OL], (2016)
- Tao Han. Research on SQL injection detection method based on analytic tree[D]. Harbin Institute of Technology, (2013)
- Tian Y J, Zhao Z M, Zhang H C. Second-order SQL Injection Attack Defense Model, Netinfo Security, **11**(2014)
- Wahid Rajeh, Alshreef Abed. A novel three-tier SQLi detectin and mitigation scheme for cloud enviroments, ICECIS, 33-37(2017)

6. Suraj Natarajan, Melody Moh. Recommending News Based on Hybrid User Profile Popularity Trends and Location, CTS,204-211(2016)
7. Wu Shao hua, Chen Shu bao.Web Attack Dtection Method Based on Support Vector Mechines, Computer Science, **42**(2015)
8. Chen xiao bing, Zhang Han yu, Luo Li ming, Research on technology of SQL injection attacks and detection. Comput Eng Appl, **43**(2007)