ON THE CLASSIFICATION BASED ON THE STOCHASITC METHOD

Takeshi Matsuda[1]

Abstract: Nowadays, a lot of classification techniques such as pattern recognition, machine learning and statistical inference have been studied. It is very important that the classifier can classify unknown data appropriately. Namely, the generalization error is very important quantity. However, even if the expectation value of the generalization error is small, there is no guarantee a classifier can classify unknown data correctly. The classifier is judged by results of the unknown data. In this study, we proposed a new classifier that fluctuate the classification threshold stochastically using a stochastic differential method. We showed that the expectation value of the classification threshold is equal to the initial threshold under some condition. Moreover, we applied our proposed model to web application attack detection problem and showed the effectiveness of our proposed method.

Keywords: Classifier, Stochastic Differential Equation, Web Application Attack.

1. INTRODUCTION

Recently, a lot of classification method based on the machine learning such as support vector machine and Naive Bayes have been devised [1,2]. Support vector machine is one of the methods in the statistical learning theory and it was proposed by Bernhard E Boser, Isab elle M Guyon and Vladimir N Vapnik [3, 4]. In the learning theory, it is very important to classify an unknown data correctly. Machine learning models learn from observed sample data and make a classification rule. And machine learning models classify an unknown data based on the obtained classification rule. There are no guarantees that machine learning models can classify correctly. So, the performance of machine learning models is judged by the false classification.

To measure the performance of machine learning models, the generalization error which expresses information on the false classification is used. It is known that the support vector machine allows for smaller generalization error.

Therefore, support vector machine has been used in various fields. Naive Bayes is also well used in many areas.

In particular, the application for the filtering of the spam mail is very famous. Support vector machine and Naive Bayes may be effectiveness for the data which does not change dramatically. However, there exists a data which changes dramatically. For example, a number of vulnerabilities concerning web application are still uncovering now. An unknown attack may be made from uncovered vulnerability, and have different feature from known attack. In this case, the generalization performance of machine learning models cannot be expected. Therefore, it is important to extract features abstractly in the case above [5, 6].

In this study, we consider the abstraction of the feature concerning web application attacks, and classifier that fluctuates the classification threshold stochastically under learning from training data. Classifier has threshold for the classification in general. Some of data near by the threshold may be misjudged by determining classification threshold. Our proposed method classifies the data near by the threshold stochastically. We will show that that the expectation value of the classification threshold is equal to the initial threshold under some condition. Finally, we will apply our proposed classifier to web application attack detection problem and showed the effectiveness of our proposed method.

2. RELATED WORK

In this section, we summarize the outline of support vector machine. Support vector machine separates the data plotted in n-dimensional Euclidean space $\boldsymbol{R}^n$ by determining some hyperplane. Even if data cannot be separated by linear hyperplane, support vector machine can be also separated by non-linear hyperplane by using kernel trick method.

From now on, we will discribe the derivation method of the separating hyperplane. Let $\boldsymbol{x}$ be a n-dimensional vector in $\boldsymbol{R}^n$ belonging to either of two different class C_1 and C_2.

Let y be a random variable which takes the binary value -1 or 1. We assume $y = 1$ (resp. $y = -1$) if $x \in C_1$ (resp. $x \in C_2$).

To classify M training data $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_M, y_M)$, we seek for the following decision function

$$\begin{cases} x \in C_1 & D(x) > 0 \\ x \in C_2 & D(x) \leq 0 \end{cases}$$

The decision function is defined by

$$D(x) = w\varphi(x) + c,$$

where $\varphi(x)$ is the function that represents x to other feature space and c is a bias. Since

$$y = \begin{cases} 1 & \text{x} \in C_1 \\ -1 & \text{x} \in C_2, \end{cases}$$

we have $y_i(w\varphi(x) + c) - 1 \geq 0$.

Moreover, the distance between the hyperplane $D(x) = 0$ and training data x is given by

$$\frac{|D(x)|}{\|w\|} \cdots (1)$$

The goal of the support vector machine is to maximize Eq.(1) under the conditions $y_i(w\varphi(x)+c)-1\geq 0$, for $i=1,2,\cdots,M$.

By using Lagrange's method undetermined multipliers, this maximization problem can be written by the minimization problem of the function

$$L(w,c,\lambda)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{M}\lambda_i(y_iD(\boldsymbol{x}_i)-1)$$

by the variables w.

Here, $\lambda=(\lambda_1,\lambda_2,\cdots,\lambda_M)$ are non-negative real valued vector. Support vector machine is not probabilistic model, but it can be deal with a random variable data.

Therefore, false classification may be caused by fixing a separating hyperplane. In this paper, for a random variable data which fluctuates drastically, we propose the classification method by applying the concept of the stochastic differential equation.

3. APPLICATION OF STOCHASTIC DIFFERENTIAL EQUATION FOR CLASSIFICATION PROBLEM

In this section, we propose the classification algorithm that the classification threshold changes stochastically. Let $X_M=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{M}$ be a n training sample vector. In this paper, we set $\boldsymbol{x_i}=(x_{i1},x_{i2})\in\boldsymbol{R}_{\geq 0}^2$ and $y_i\in\{-1,1\}$ for all $i=1,2,\cdots,M$. Our goal is to classify au unknown data $\boldsymbol{x}^*\in\boldsymbol{R}^2$ from a training sample X_M. Let $\{(\boldsymbol{x},1)\}\subseteq R^2\times\{-1,1\}$ and $\{(\boldsymbol{x},-1)\}\subseteq R^2\times\{-1,1\}$ be a set of data belonging to the class C_1 and the class C_2, respectively. Our proposed algorithm is composed of the following four steps.

1. Prepare m_1 and $m_2$sapmles from the class C_1 and C_2, respectively.

2. Compute

$$\begin{cases} a=max_{1\leq i\leq m_1}\left\{\dfrac{x_{i2}}{\sqrt{x_{i1}^2+x_{i2}^2}}\right\} & \text{for } (x_i,1)\in C_1 \\ b=max_{1\leq j\leq m_2}\left\{\dfrac{x_{j2}}{\sqrt{x_{j1}^2+x_{j2}^2}}\right\} & \text{for } (x_j,-1)\in C_2 \end{cases}$$

3. Choose $\varepsilon\in\boldsymbol{R}$ satisfying $0\leq\varepsilon\leq|\theta_a-\theta_b|$,and compute

$$\begin{cases}\theta(0)=\varepsilon+\min\{\theta_a,\theta_b\}.\\ d\theta(t)=\mu(t)dt+\sigma(t)dB(t).\end{cases}$$

4. If $\sin\theta(t) \geq \frac{x_2^*}{\sqrt{x_1^{*2}+x_2^{*2}}}$ (resp. $\sin\theta(t) < \frac{x_2^*}{\sqrt{x_1^{*2}+x_2^{*2}}}$), then $y^* = 1$ (resp. $y^* = -1$), for an unknown data $(\boldsymbol{x}^*, y^*) \in \boldsymbol{R}^2 \times \{1, -1\}$.

Here, we set $\boldsymbol{a} = \left(\sqrt{1-a^2}, a\right)$ and $\boldsymbol{b} = \left(\sqrt{1-b^2}, b\right)$, then $\boldsymbol{a}$ and $\boldsymbol{b}$ are the vectors on the first quadrant. Moreover, we define $\theta_{\boldsymbol{a}}$ (resp. $\theta_{\boldsymbol{b}}$) as the angle between $\boldsymbol{a}$ and $\boldsymbol{e}_1 = (1,0)$ (resp. between $\boldsymbol{b}$ and $\boldsymbol{e}_2 = (0,1)$).

The equation

$$d\theta(t) = \mu(t)\theta(t) + \sigma(t)dB(t) \cdots (2)$$

is a stochastic differential equation and $B(t)$ denotes Brownian motion. The Brownian motion $B(t)$ is defined in the following way.

(1) $B(0) = 0$.

(2) For $0 = t_0 < t_1 < t_2 < \cdots < t_n < \cdots$,

$B(t_1) - B(t_0), B(t_2) - B(t_1), \cdots B(t_{n+1}) - B(t_n)$

are all independently.

Futhermore, if $s < t$, then the probability that $B(t) - B(s)$ belonging in the area $A \subset \boldsymbol{R}$ is given by

$$P\big(B(t) - B(s)\big) = \frac{1}{\sqrt{2\pi(t-s)}} \int_A e^{-\frac{x^2}{2(t-s)}} dx$$

Let $\mathbf{Err}(\boldsymbol{x}^*, y^*)$ be a probability of the false classification of proposed algorithm. Then, the following theorem holds for the proposed algorithm.

[Theorem]

We assume $\mu(t) = \mu\theta(t)$ and $\sigma(t) = \sigma\theta(t)$. Let $\mu, \sigma \in \boldsymbol{R}$.

Under the condition that is given consideration to the expectation value of $\theta(t)$ by the probability P defined in the definition of Brownian motion, the probability of

$\mathbf{Err}(\boldsymbol{x}^*, y^*)$ is given by

$$\mathbf{Err}(\boldsymbol{x}^*, y^*) = \frac{|\theta_1 - \theta_2|}{45}$$

if $\mu = 0$.

To prove this theorem, the following Ito's lemma is required.

Let F_t be a minimum completely additive class that the random variable sequence $\theta(1,\omega),\cdots,\theta(t,\omega),\cdots,\theta(T,\omega)$ are all measurable, where $t \in [0,T] \subset \boldsymbol{R}$.

We prepare two continuous stochastic process $\mu(t,\omega)$ and $\sigma(t,\omega)$, and assume that $\mu(t,\omega)$ and $\sigma(t,\omega)$ are square integrable and F_t-measurable. .

[Lemma]

For the stochastic process

$$\theta(t) = \theta(0) + \int_0^t \mu(s,\omega)ds + \int_0^t \sigma(s,\omega)dB(s),$$

the following equation holds.

$$f\big(\theta(t)\big) = f\big(\theta(0)\big) + \int_0^t \frac{df\big(\theta(s)\big)}{dx}\mu(s,\omega)ds + \int_0^t \frac{df\big(\theta(s)\big)}{dx}\sigma(s,\omega)dB(s) + \frac{1}{2}\int_0^t \frac{d^2f\big(\theta(s)\big)}{dx^2}\sigma(s,\omega)^2 ds,$$

where $f: \boldsymbol{R} \to \boldsymbol{R}$be a C^2-class function (We denote $f(x)$), and both $\frac{df}{dx}$ and $\frac{d^2f}{dx^2}$ are bounded.

4. PROOF OF THEOREM

Firstly, we will give the proof that the expectation value of $\theta(t)$ by the probability P is equal to $e^{\mu t}\theta(0)$. This proof is well known, but it has very important role in our proposed algorithm. Secondly, we prove that the probability of the false classification of proposed algorithm is equal to $\mathbf{Err}(\boldsymbol{x}^*, y^*) = \frac{|\theta_1-\theta_2|}{45}$ by using the expectation value of $\theta(t)$ by the probability P.

The stochastic process $\theta(t)$ has the following stochastic differential equation.

$$d\theta(t) = \mu\theta(t)dt + \sigma(t)dB(t).$$

From Ito's lemma, we have

$$f\big(\theta(t)\big) = f\big(\theta(0)\big) + \int_0^t \frac{df\big(\theta(s)\big)}{dx}\mu\theta(s)ds + \int_0^t \frac{df\big(\theta(s)\big)}{dx}\sigma\theta(s)dB(s) + \frac{1}{2}\int_0^t \frac{d^2f\big(\theta(s)\big)}{dx^2}(\sigma\theta(s))^2 ds,$$

Here, we put $f(x) = \log x$. Then, we get

$$\log\frac{\theta(t)}{\theta(0)} = \mu t + \sigma B(t) - \frac{1}{2}\sigma^2 t.$$

Therefore, we have

$$\theta(t) = \theta(0)e^{-\frac{1}{2}\sigma^2 t + \sigma t + \mu t}$$

Now, let us compute the expectation value of $\theta(t)$ by the probability P.

We define the expectation value of $\theta(t)$ by the probability P as $\mathrm{E}_P[\theta(t)]$.

Since $B(t)$ is a Brownian motion that has mean 0 and variance t, we obtain

$$\begin{aligned}\mathrm{E}_P[\theta(t)] &= \frac{1}{\sqrt{2\pi t}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\sigma^2 t + \sigma x + \mu t - \frac{x^2}{2t}}\,dx \\ &= \frac{1}{\sqrt{2\pi t}} e^{\mu t}\theta(0)\int_{-\infty}^{\infty} e^{-\frac{1}{2t}(x-t\sigma)^2}\,dx\end{aligned}$$

Hence, we have $\mathrm{E}_P[\theta(t)] = e^{\mu t}\theta(0)$. Therefore, we have $\mathrm{E}_P[\theta(t)] = \theta(0)$ if $\mu = 0$. This means that the expectation value $\mathrm{E}_P[\theta(t)]$ does not depend on t.

Finally, we compute that the probability of the false classification of proposed algorithm. The type of the false probability is as follows.

Type A :

True is $(\boldsymbol{x}^*, y^*) \in C_1$, but the algorithm judges $(\boldsymbol{x}^*, y^*) \in C_2$

Type B :

True is $(\boldsymbol{x}^*, y^*) \in C_2$, but the algorithm judges $(\boldsymbol{x}^*, y^*) \in C_1$

If the unknown data $(\boldsymbol{x}^*, y^*)$ takes the value in the region

$$Z_1 = \left\{ (\boldsymbol{x}^*, y^*) \middle| \sin\theta^* \le \frac{x_2^*}{\sqrt{x_1^{*2} + x_2^{*2}}} \le \sin\theta_1 \right\}.$$

Then the false classification of type A is caused,Similarly, if $(\boldsymbol{x}^*, y^*)$ takes the value in the region

$$Z_2 = \left\{ (\boldsymbol{x}^*, y^*) \middle| \sin\theta_1 \le \frac{x_2^*}{\sqrt{x_1^{*2} + x_2^{*2}}} \le \sin\theta^* \right\}.$$

Then the false classification of type B is caused. We denote the area of the region Z_l by Z_l $(l = 1,2)$. Since the expectation value $\mathrm{E}_P[\theta(t)]$ is equal to the initial value $\theta(t)$,

the probability $\mathbf{Err}(\boldsymbol{x}^*, y^*)$ under the expectation value $\mathrm{E}_P[\theta(t)]$ equals to

$$\mathbf{Err}(\boldsymbol{x}^*, y^*) = \frac{|Z_1| + |Z_2|}{\frac{\pi}{4}} = \frac{|\theta_1 - \theta_2|}{45}.$$

5. EXAMPLE

In this section, we will treat simple example of the classification problem. From now on, let us detect SQL injection attack by using support vector machine and our proposed method which is introduced in section 2 and section 3, respectively. SQL injection attack is executed via the form on website, and attackers steal personal information from the database of web application. We show a specific example of SQL injection attack.

SELECT * FROM TableX WHERE ID='7' '7'='7'";

Since 7=7 is always true, if the SQL sentence above was executed, then all of records in TableX are outputted. As the result, attackers can get the data which must not be able to be originally acquired. In [5, 6], we proposed the detection model of SQL injection attacks by computing the content rate of attack feature symbols. If we detect SQL injection attack by using the symbols space, single quote and semi-colon as attack feature symbols, the content rate of the example above is equal to $\frac{14}{43} = 0.32558\cdots$. If we set the detection threshold α as $0 < \alpha < \frac{14}{43}$, we can detect the SQL injection attack described above. In our previous study [5, 6], the content rate of most attack data were greater than 0.09. Likewise, the content rate of most normal data was equal to 0. However, some normal data includes space, so the false classification might be caused in such cases.

In this paper, we will apply our proposed method and support vector machine for the detection of SQL injection attack. And we show the effectiveness of our proposed model by comparing the result of these two methods. To detect SQL injection attack, we use the information of the content rate x of the attack feature symbols including in an input string. From an input data, we compute $(x_1, x_2, y) = (x_1, 1 - x_1, y) \in \boldsymbol{R}^2 \times \{-1,1\}$. If an input is attack (resp. normal), then we define $y = 1$ (resp. $y = -1$). We use the learning data

$$\{(0.2, 0.8, 1), (0.1, 0.9, 1), (0, 1, -1), (0.01, 0.09, -1)\}$$

and prepare the testing data $\{(0.0588, 0.9412, -1)\}$.

Note that for the sake of ease, we treat simple example. We used three symbols, space, single quote and semi-colon, as the attack feature symbols.

5.1. Detection by Suppor Vector Machine

By using the learning data, we see that the support vectors are

$$\{(0.1, 0.9, 1), (0.01, 0.09, -1)\}$$

and the separating hyperplane is

$$D(x_1, x_2) = -\frac{100}{9}x_1 + \frac{100}{9}x_2 - \frac{89}{9} = 0$$

Note that $D(0.055, 0.045) = 0$. If $D(x_1, x_2) \geq 0$, then we assume that attack is observed. Therefore, if a normal data whose content rate is equal to and greater than 0.055 were observed, then the support vector machine cannot classify correctly. For example, the content rate of the normal input data

Dr.Takeshi Matsuda is equal to $0.05882\cdots$.

5.2. Detection by Proposed Method

From the learning data $\{(0.2, 0.8, 1), (0.1, 0.9, 1), (0, 1, -1), (0.01, 0.09, -1)\}$, we have $\theta_a = 1.460139, \theta_b = 1.560696$.

Let us define $\theta(0) = 1.512661$. This is based on the classification point $(0.055, 0.945)$ of the support vector machine above. We set $\mu(t) = 0$ and $\sigma(t) = \sigma\theta(t)$ and $t_n - t_{n-1} = 0.1$ in Eq.(2). In this experiment, we set $\sigma = 1, 0.1, 0.01, 0.001, 0.000$ and classified the 100 normal data $(0.0588, 0.9412)$. The result is shown in Table 1.

σ	1	0.1	0.01	0.001	0.0001
Rate	14.18%	51.57%	63.41%	89.75%	100%

Table 1: Average rate of the false classification of proposed method

5.3. Discussion: In this experiment, we used the learning data $\{(0.2, 0.8, 1), (0.1, 0.9, 1), (0, 1, -1), (0.01, 0.09, -1)\}$and prepare the testing data $\{(0.0588, 0.9412, -1)\}$. These sample data feature the most attack and normal data in SQL injection attack. We may expect that most attack data are classified correctly from the result of the previous work [6]. However, some normal data including space character are classified as attack data such as the example of 5.1.

Therefore, we assume that (x_1^*, x_2^*) is classified as attack if $\sin\theta_a \geq \frac{x_2^*}{\sqrt{x_1^{*2} + x_2^{*2}}}$. In the experiment of this study, we assumed that 100 normal data such that the examle of 5.1 are observed, and classified these data using the support vector machine and our proposed method. As the result, the average of the false classification probability of our proposed method is 14.18% (when $\sigma = 1$), though the support vector machine

cannot classify. As shown in Table 1, we see that the average of false classification increases with the decreasingσ. This experiment seems to be a little successful, but there are some problems. For example, we used the data in the first quadrant, but there are some case that $\theta(t)$ is not the angle in the first quadrant. To control the motion of $\theta(t)$ is our future work.

6. CONCLUSION

In this study, we proposed a new classifier that fluctuates the classification threshold stochastically. To find out the optimal stochastic differential equation for the classification of web application attack is our future work.

7. ACKNOWLEDGMENT

This research is partially supported by the 44th Kurata Grants of the Kurata Memorial Hitachi Science and Technology Foundation.

8. REFERENCES

1. F. Nentwich, N. Jovanovic, E. Kirda, C. Kruegel, G. Vigna, " Cross-Site Scripting Prevention with Dynamic Data Tainting and Static Analysis," In Proceeding of the Network and Distributed System Security Symposium NDSS, 2007.
2. Komiya. R, Paik. I, Hisada. M., Wadsworth, "Classification of malicious web code by machine learning,". IEEE 2011 3rd International Conference on Awareness Science and Technology (iCAST), 406- 411, 2011.
3. B.E. Boser et al, "A Training Algorithm for Optimal Margin Classifiers," Proceedings of the Fifth Annual Workshop on Computational Learning Theory 5 144-152, Pittsburgh, 1992.
4. V. Vapnik, "The Nature of Statistical Learning Theory," 2edition, Springer, 1999.
5. Takeshi Matsuda, Daiki Koizumi, Michio Sonoda, and Shigeichi Hirasawa, "On Predictive Errors of SQL Injection Attack Detection by the Feature of the Single Character," Proceeding of 2011 IEEE International Conference on Systems, Man, and Cybernetics (SMC2011), pp. 1722-1727, 2011.
6. Michio Sonoda, Takeshi Matsuda, Daiki Koizumi, and Shigeichi Hirasawa, "On Automatic Detection of SQL Injection Attacks by the Feature Extraction of the Single Character," Proceedings of the 4th International Conference on Security of Information and Networks (SIN2011), ANM, pp.81-86, 2011.

* * * * *

---

[1]*Takeshi Matsuda 4F, 1-11 Kitayamabushi-cho, Shinjuku-ku, Tokyo , Japan /Cyber Univeristy/takeshi_matsuda@cyber-u.ac.jp*