A Method of Confidence Evaluation Based on Support Vector Machines

Yang Pan, Shuo Zhang and Quanyin Zhu
Faculty of Computer & Software Engineering, Huaiyin Institute of Technology
Huaian, China
E-mail: hyitzqy@126.com

Abstract. In present, the analysis of students’ behavior is important to instructors and leaders in universities. In order to improve the value of data in the Oran system of Huaiyin Institute of Technology and evaluate the degree of honesty, a method based on Support Vector Machines algorithm is used to evaluate the confidence of the information that about 18000 students filled in the Oran system. Through the experiment, the method achieves good result and achieves a high efficiency, which satisfies the original goal.

1. Introduction
In this information age, it attaches great importance to evaluate people’s behavior especially students’ in universities. In recent years, the education mode has changed. Students' behavior is important for teaching and educational research [1]. For instructors and leaders in universities, the degree of honesty of students also means a lot except the academic achievement.

In this paper, the confidence [2] of students’ data is used to measure the honesty of students. Confidence [3] is referred to the degree of truth of specific object to specific event, which is also referred to the degree of rationality of personal belief [4]. Nowadays, in some applications, not only recognition rate [5] is ought to be as high as possible, but also false acceptance rate is supposed to be as low as possible [6]. The function of confidence is to provide the basis for false acceptance [7]. And at present, the evaluation of confidence mostly covers Support Vector Machines, Naive Bayes, and k-Nearest Neighbor and so on [8].

In this paper, a method based on Support Vector Machines algorithm is designed to evaluate the confidence of students’ data in the Oran system.

The structure of the paper is as follows: section 1 introduces the tendency to the evaluation of students’ behavior and recent situation about evaluating the confidence of the data; section 2 introduces Support Vector Machines algorithm and correction algorithm; section 3 introduces experiment preparations, including data preprocessing and so on; section 4 states the experiments and the relative evaluation results; section 5 draws conclusions and the work that ought to be do in the future.

2. Support Vector Machines Algorithm

2.1. The Concept and Principle of Support Vector Machines Algorithm
The original Support Vector Machines algorithm was invented by Vladimir N. Vapnik in 1963 [9]. In 1992, Bernhard E. Baser, Isabelle M. Guyton and Vladimir N. Vapnik suggested a way to create nonlinear classifiers by applying the kernel trick to maximum-margin hyper planes [10]. The current
standard incarnation (soft margin) was proposed by Corina Cortes and Vapnik in 1993 and published in 1995 [11].

Support Vector Machines algorithm has many unique advantages in solving small sample, nonlinear and high dimensional pattern recognition, and can be applied to other functions such as function fitting machine learning problems. In machine learning, Support Vector Machines are a set of supervised learning methods used for classification, regression and outlier detection [12].

2.2. Evaluation of Confidence Based on Support Vector Machines

The evaluation of confidence based on Support Vector Machines is divided into ‘One-Against-One’ and ‘One–Against-All’, as follows:

One-Against-One is a method of classification which deals with two-category problem. It separately uses two different categories to make up a Support Vector Machines sub classifier, which can make each category formed with other categories into a two-category problem, so that there will be \( k(k-1)/2 \) Support Vector Machines sub classifiers. When forming feature \( i \) and feature \( j \) in Support Vector Machines sub classifiers, select sample data of feature \( i \) and feature \( j \) as training sample data in training sample set. Meanwhile, sign the sample data that attribute to feature \( i \) as positive and sign the sample data that attribute to feature \( j \) as negative. One-Against-One method need to handle the problem about optimization as follows:

\[
\begin{align*}
\text{Minimize} & \quad \phi(\omega, b, z) = \frac{1}{2} \sum_i \sum_j \xi_{ij} \\
\text{s.t.} & \quad (\omega^i)^T \phi(x_i) + b^i \leq -1 + \xi_{ij}, \text{ if } y_i = i \\
& \quad (\omega^j)^T \phi(x_i) + b^j \leq -1 + \xi_{ij}, \text{ if } y_i \neq i \\
& \quad \xi_{ij} \geq 0
\end{align*}
\]

One-Against-All is also a method which deals with two-category classification problem. It will structure a series of two-category classifier, where each classifier will divide the other categories with one category. Meanwhile, set the whole of the other categories as the other category, which will form a two-category classification problem. The method of One-Against-All is to generate \( k \) Support Vector Machines classifiers. When generating feature \( i \) which attributes to sample data as positive feature and feature \( j \) which does not attribute to sample data as negative feature. The problems about optimization which need to handle are:

\[
\begin{align*}
\text{Minimize} & \quad \phi(\omega', b', z') = \frac{1}{2} \sum_i \sum_j \xi_{ij} \\
\text{s.t.} & \quad (\omega^j)^T \phi(x_i) + b^j \geq 1 - \xi_{ij}, \text{ if } y_j = i \\
& \quad (\omega^j)^T \phi(x_i) + b^j \leq -1 + \xi_{ij}, \text{ if } y_j \neq i \\
& \quad \xi_{ij} \geq 0, j = 1, 2, \ldots, l
\end{align*}
\]

After handling these problems, it will achieve \( k \) decision function, as follows:

\[
\begin{align*}
(\omega^1)^T \phi(x) + b^1 \\
\vdots \\
(\omega^k)^T \phi(x) + b^k
\end{align*}
\]
2.3. Correction Algorithm

Considering that the probability that sample data to be measured is whether the same feature with the training samples in the neighborhood affects the confidence of data, as follows:

If the sample $x_i$ is positive sample, the more the number of positive samples, the higher the probability that sample $x_i$ is positive. Therefore, we measure probability that the training sample points of test data are the same category to represent the confidence of data.

Algorithm:

Input: Original training student set $S$, test student set $T = \{x_1, x_2, x_3, ..., x_n\}$, rejection rate $e$.

Output: Partial accepted sample classification results and residual classification results after rejecting distinguishing.

(a) Use optimal training model $M$.
(b) Initialize data, set $i = 1$.
(c) Take training sample $T$ of test sample $\phi$ in turn, if $T \neq \phi$, set $x_i \in T$; otherwise perform step (i).
(d) Use optimal training model $M$ predict test sample, so that it will get the distance $d(x_i)$ that each test sample $x_i$ to Support Vector Machines optimal classification surface and each sample’s classification results to be measured $d_i$.
(e) Take $j$ near neighbors in training sample of each test sample.
(f) Get the probability $\rho_i$ that $j$ training sample attributing to $d_i$ for classification results of test sample.
(g) Calculate the confidence of test sample according to $d(x_i)$, $\rho_i$.
(h) Set $i = i+1$, perform step (c).
(i) Sort the confidence $f(x_1), f(x_2), f(x_3), ..., f(x_n)$ for classification results $g_1, g_2, g_3, ..., g_n$ of test sample from small to large gradually.
(j) Accept the corresponding sample which larger than threshold, output the classification results of the sample. Meanwhile, reject the surplus sample to rejection sample set ACC.
(k) Put the rejection sample to the KNN classifier to reclassify and output the results.

3. Data Preparation

The data that is used in the experiments is all from the information of students in the Oran system of Huaiyin Institute of Technology. The context of the data in the Oran system consists of id, name, sex, id number, telephone number, award, scholarship and so on. The data cover about 18000 students, which is the number after deleting duplicate id’s items.

3.1. Experimental Environment

The machine configuration is Windows 8/10, RAM is 8G. Python 2.7 is used to preprocess and handle the data consisting of normalization and so on.

3.2. Data Preprocessing

Due to the information in the Oran system is very messy, it is necessary to do preprocessing word before training model. For example, in this data, it is serious that there are too many missing values and the labels of some labels are in chaos and are not standard, which will give away to doing the next data analysis work. The steps of preprocessing and points are as follows:

Due to a student only can own one piece of information, it is necessary to delete repeating items according to id.

Considering that some information in the Oran system is not filled in the system on students’ own. Therefore, the data consisting of award, scholarship, political outlook, telephone number, class position, students union’s position, competition level, competition name:
Data = {award, scholarship, political outlook, telephone number1, telephone number2, class position, students union’s position, competition level}

Owing to that the names of some levels are not standard, it is necessary to standardize them. For example, the labels of scholarship ought to be: special-class scholarship, first-class scholarship, second-class scholarship, third-class scholarship; The labels of competition level should be: national level, provincial level, city level, school level, college level. And the concrete arrangement can be seen as follows:

(a) Telephone number1: ‘true’, ‘false’.
(b) Telephone number2: ‘true’, ‘false’.
(c) Political outlook: ‘party member’, ‘probationary party member’, ‘communist youth league member’, ‘the masses’, ‘none’.
(d) Award: ‘excellent student’, ‘outstanding student cadres’, ‘activities award’, ‘none’.
(e) Scholarship: ‘the first prize’, ‘the second prize’, ‘the third prize’, ‘special prize’, ‘none’.
(f) Class position: ‘true’, ‘false’.
(g) Students union’s position: ‘true’, ‘false’.
(h) Competition level: ‘nation’, ‘province’, ‘city’, ‘university’, ‘college’.

It is worth mentioning that ‘Tel number1’ is the telephone number of students, and ‘Tel number2’ is the telephone number of students’ parents.

Due to the positions that students in the class and students’ union are so many and matter less that set the levels of them are ‘True’ and ‘False’, which represent s that the student is whether of the students’ union or not.

Concerning the matter about how to judge whether the telephone number of students and their parents are the real number of not, the method is taken:

1. The length of the mobile phone number is 11 and the first number is ‘1’.
2. The telephone number at home is like ‘XXXXXXXX’ or ‘XXXX-XXXXXXXX’ or ‘XXX-XXXXXXXX’. Therefore, the length of the number is ‘8’ or ‘11’ or ‘7’.

All the missing values are set as ‘None’ or ‘False’.

3.3. Model Training

Considering that award, scholarship, political outlook, class position, competition level are the important information for students that to some degree can represent whether they behave well at school or not. What’s more, this information is relatively reliable that fill in the system. Meanwhile, students fill the telephone number at will and it is hard to judge the telephone numbers that they fill in are real. Therefore, ‘Tel number1’ and ‘Tel number2’ are set the target, and ‘award’, ‘scholarship’, ‘class position’, ‘students union’s position’, ‘competition level’, ‘political outlook’ are used to evaluate the reliability of the target.

In order to improve the result of the experiment, it is important to try to take these methods: select positive sample and negative sample to do training work. The method is as follows:

Selecting negative sample: Only 1 or 2 features can prove that the student behaves well at school. For example, data = {‘First’, ‘None’, ‘Common’, ‘False’, ‘False’, ‘None’, ‘False’, ‘False’}.

Selecting positive sample: At least features can prove that the student behaves well at school. For example, data = {‘First’, ‘Excellent student’, ‘Province’, ‘False’, ‘None’, ‘False’, ‘False’}.

Due to the positive sample is relatively little, the number of negative data is 2154, and the number of positive data is 2250, which almost equals the number of negative sample.

And then, due to the features are discrete variables, it should set continuous variables. Therefore, the weight ought to be put on the labels as follows:

1) Due to ‘Tel number1’, ‘Tel number2’, ‘class position’, ‘students union’s position’ has only two labels. Therefore, set these features as ‘-1’ and ‘1’.

2) The other features originally set the stationary weight that each label of one feature has the same interval. And in the following experiment, the weight of some feature can have some change.

In this paper, the precision of positive sample (P_prec), the recall of positive sample (P_recall), the precision of negative sample (N_prec), the recall of negative sample (N_recall), the precision of sample (P) are used to measure the experiment results, as follows:
(a) \( Tp \) (true positive) represents the true samples that are predicted as positive.
(b) \( Fp \) (false positive) represents the false samples that are predicted as positive.
(c) \( Tn \) (true negative) represents the true samples that are predicted as negative.
(d) \( Fn \) (false negative) represents the false samples that are predicted as negative.

(e) Precision of positive sample (\( P_{\text{prec}} \)) = \( \frac{Tp}{Tp + Fn} \).

(f) Recall of positive sample (\( P_{\text{recall}} \)) = \( \frac{Fn}{Tp + Fn} \).

(g) Precision of negative sample (\( N_{\text{prec}} \)) = \( \frac{Tn}{Tn + Fn} \).

(h) Recall of negative sample (\( N_{\text{recall}} \)) = \( \frac{Tn}{Tn + Fp} \).

(i) The precision of sample (\( P \)) = \( \frac{Tn+Tp}{Tn+Fn+Fp+Fn} \).

And then, use the correction algorithm based on Support Vector Machines algorithm to train the sample, so that we can evaluate the confidence of the students’ data. In order to improve the efficiency, change the rejection rate, which ranges from 1%-10% and the interval is 1%.

4. Experimental Results
The aim of experiment is to the comparison about the confidence of students’ data in the Oran system of Huaiyin Institute of Technology, with the change of rejection from 1% to 10% that uses the correction algorithm based on Support Vector Machines.

Table 1 shows the results of the experiment. From table 1, it is clear that the confidence of sample differ from each other in different rejection rate. And from 1% to 6%, the correction algorithm has the similar classification performance. Meanwhile, the classification performance has a big increase (3%-6%), where the rejection rate is from 7% to 10%.

| Rejection rate | P_{\text{prec}} (%) | P_{\text{recall}} (%) | N_{\text{prec}} (%) | N_{\text{recall}} (%) | P (%) |
|---------------|---------------------|----------------------|---------------------|----------------------|-------|
| 1%            | 83                  | 71                   | 76                  | 67                   | 79    |
| 2%            | 81                  | 73                   | 78                  | 65                   | 78    |
| 3%            | 82                  | 69                   | 76                  | 68                   | 82    |
| 4%            | 84                  | 73                   | 79                  | 66                   | 80    |
| 5%            | 83                  | 70                   | 77                  | 69                   | 81    |
| 6%            | 84                  | 75                   | 77                  | 73                   | 83    |
| 7%            | 88                  | 79                   | 85                  | 75                   | 87    |
| 8%            | 88                  | 81                   | 87                  | 71                   | 89    |
| 9%            | 87                  | 79                   | 87                  | 76                   | 88    |
| 10%           | 89                  | 82                   | 88                  | 78                   | 89    |

5. Conclusion
With more and more students’ information appear, it becomes more and more important for instructors and leaders in universities to evaluate students’ behavior through the information. And the confidence can be used to evaluate he degree of honesty of students. Through the experiment, it is concluded that the confidence can be represented by the precision and recall of sample when using the correction algorithm based on Support Vector Machines. First, we do the data preprocessing work. Then, we use the correction algorithm to train the positive and negative sample to evaluate the confidence of
students’ information. According to the result of experiment, the method achieves higher efficiency where the rejection rate from 7% to 10%.

6. References

[1] Wan H, Yu Q, Ding J, et al. Students' behavior analysis under the Sakai LMS [C]//Teaching, Assessment, and Learning for Engineering (TALE), 2017 IEEE 6th International Conference on. IEEE, 2017: 250 - 255.

[2] Piano X, Wang Z, Liu G. Research on mining positive and negative association rules based on dual confidence [C]//Internet Computing for Science and Engineering (ICICSE), 2010 Fifth International Conference on. IEEE, 2010: 102-105.

[3] Weinzierl W. Attribute Assisted Interpretation Confidence Classification Using Machine Learning [C]// Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on. IEEE, 2017: 55 - 60.

[4] Jiang J, Miyagusuku R, Yamashita A, et al. Glass confidence maps building based on neural networks using laser range-finders for mobile robots[C]//System Integration (SII), 2017 IEEE/SICE International Symposium on. IEEE, 2017: 405 - 410.

[5] Ben-Hur A, Horn D, Siegel Mann H T, et al. Support vector clustering [J]. Journal of machine learning research, 2001, 2 (Dec): 125-137.

[6] Yang L, Ng B, Selah W K G, et al. Deterministic Confidence Interval Estimation of Networking Traffic in SDN [C]//Local Computer Networks (LCN), 2017 IEEE 42nd Conference on. IEEE, 2017: 120 - 127.

[7] Meng X, Zhang Z, Geng Z, et al. MRF-Based Disparity Upsampling Using Stereo Confidence Evaluations [J]. IEEE Signal Processing Letters, 2017.

[8] Kotare T, Monads S, Johor J. Speech corpus generation based on N-gram confidence measure classification [C]//ELMAR, 2017 International Symposium. IEEE, 2017: 149 -152.

[9] Zhang G, Piccard M. Structural SVM with partial ranking for activity segmentation and classification [J]. IEEE Signal Processing Letters, 2015, 22(12): 2344-2348.

[10] Bouncier A, Beaueroy P, Grill E. A straightforward SVM approach for classification with constraints [C]//Signal Processing Conference, 2005 13th European. IEEE, 2005: 1- 4.

[11] Keith S, Sheaved S K, Bhattacharyya C, et al. Improvements to Platt's SMO algorithm for SVM classifier design [J]. Neural computation, 2001, 13(3): 637- 649.