Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005

DEVELOPING AN INTELLIGENT DATA DISCRIMINATING SYSTEM OF ANTI-MONEY LAUNDERING BASED ON SVM

JUN TANG [1], JIAN YIN [2]

[1]Information Technology School, Zhongnan University of Economics and Law, Wuhan 430074, China
[2]Department of Computer Science, Zhongshan University, Guangzhou 510275,China
E-MAIL: issjyin@zsu.edu.cn, tangjunyes@tom.com

Abstract:

Statistical learning theory (SLT) is introduced to improve the embarrassments of anti-money laundering (AML) intelligence collection. A set of unusual behavior detection algorithm is presented in this paper based on support vector machine (SVM) in order to take the place of traditional predefined-rule suspicious transaction data filtering system. It could efficiently surmount the worst forms of suspicious data analyzing and reporting mechanism among bank branches including enormous data volume, dimensionality disorder with massive variances and feature overload.

Keywords:

Anti-money laundering; SLT; support vector machine; pattern recognition

1. Introduction

Anti-money laundering (AML) is attracting a highly focused attention among mainland China financial and judicial fields nowadays. It has become a symbol of the abilities of the central government and the faith of the nation, as well as an important measure fighting organized crimes and forestalling the soaring nation-crossed corruption. However, the construction of an effective AML mechanism is just at its startup which is far from perfection. From 2003 the regulator PBC(The People's Bank of China) has promulgated stipulations to monitor the suspicious transactions related with money laundering. Every branch of commercial banks is responsible to regularly report large value and suspicious transaction data to the intelligence department. A set of computer-aided transaction monitoring system based on pre-defined judging rules has also been developed and took its trial in some provinces at the end of 2004. But the data reporting system has been embarrassed by its terrible data volume and the high false positive rate which shows poor intelligence merits. The current monitoring system works mainly by establishing a fixed threshold or by finding a known money laundering pattern. Such method is accompanied with two connatural limits: First, pre-established rules may be easily eluded by the money launders while those normal businesses may frequently trigger the thresholds resulting false positive reports. Second, it could only discover the specific money laundering behavior that has been found before, while the new appeared pattern would be missed.

Urgent requirements of developing new generation solutions with the ability of self-learning, self adaptive and self-decision making are put forward by the regulators. Some related research work has been done among the field of customer behavior pattern recognition. Suspicious data may be derived out by comparing the unusual transaction records with the normal behavior norms. It is reported a relative better result has been gained during trial running: the system has got approximate 1-in-14 false-positive rate, generates alerts on average at a rate of 0.00001[11]. However, due to the late startup of domestic AML system developing, no self-learning and self-adaptive research results have been found in our literature searching while similar applications in intrusion detection of system security are flooding which may become good references.

The finance business environments give many differences compared with their western colleagues, for examples, Chinese people prefer cash payments which takes a proportion of more than 80%. Credit payments level in small medium cities and rural areas is even far lower than that in big cities. It's very difficult to trace the sources of cash currencies, those bank branches and private underground illegal banks dispersed in lower level areas are main choices for the money laundering activities. And each bank adopts different IS database structure which brings about the difficulties of data pre-cleansing and training data sets obtaining. So, it is essential to study the unsupervised learning and detection method in addition to the supervised method.

The paper has presented an unusual customer behavior detection method. We construct a RBF kernel function based on the definition of HVDM distance, which extends the supervised C-SVM algorithm and unsupervised one-class SVM algorithm over heterogeneous data sets. The rest of the paper is organized as follows: section 2 discusses the mechanism to apply pattern recognition technologies to AML data discrimination, and then introduces the details of C-SVM and One-class SVM algorithms. Section 3 discusses SVM application over heterogeneous data sets. Section 4 gives the verification experiment prototype, methods and results based on the simulated data sets provided by Wuhan branch of China Agriculture Bank, a commercial bank in south central China. Section 5 is the conclusion.

2. Unusual customer behavior detection using SVM

2.1. SVM is suitable for AML data filtering

The problem each bank posed is far more challenging---millions transactions per day, millions individual accounts and customers, hundreds of product types, money laundering could range from a single transaction to the culmination of months of complex transactional activity. The current AML data reporting system is mainly running on the rule-based mechanism, which pre-establishes a set of fixed threshold or pattern matching those money laundering cases discovered before. Such model could be called the first generation AML system. Enormous data volume and high false-positive rate bring about urgent requirements of developing new generation system which could understand and measure what is usual for a customer. The AML system must acquire an empirical understanding of the norm at the account level, and then examine all dimensions of unusual behavior. So the problem becomes a pattern classification to separate the unusual from the normal and only those unusual records would be reported. This method could efficiently solve the main problems existing in the current data reporting system.

In practice, AML data reporting now has become a pattern classification task to determine whether the transaction appears normal. SVM is a classifier based on small example training sets and is not sensitive to the dimensionality disorder. It could also be utilized to density estimation and outlier discovery. Thus SVM is suitable for the classifier designing and unusual discovery among high dimensionality heterogeneous data sets.

2.2. C-SVM algorithm

Considering the simplest two-class classification task, it is given the training datasets (xi,yi) with underlying probability distribution F(x,y):

$$(x1,y1),\ldots,(xm,ym)\in X\times\{\pm 1\}, \quad (1)$$

X should not be null without any other extra requirement, try to design an optimum classifier f(x): X→{-1,+1},in order to estimate the probability distribution of given test dataset.

If X belongs to linear separable R^n, the task will become a problem of seeking a generalized optimum separating plane with a maximal separating margin. This problem could be converted to a dual optimization question:

$$\text{Max } Q(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j (x_i\bullet x_j),$$

$$\text{Subject to } \alpha_i\ge 0, i=1,\ldots,n \text{ and } \sum_{i=1}^{n} y_i\alpha_i=0, \quad (2)$$

We get an optimum classification function by solving above question:

$$f(x)=\operatorname{sgn}\{(\omega\bullet\chi)+b\}=\operatorname{sgn}|\sum_{i=1}^{n}\alpha_i^* y_i (x_i\bullet\chi)+b^*| \quad (3)$$

If the input space X is linear inseparable, a tradeoff should be taken between the optimum classification plane and the minimum mistaken classification examples. That is, to construct a classification hyper plane with a soft separating margin.

If the input space is non-linear, the statistical learning theory (SLT) adopts a kernel function to map the input space into a high dimensioned feature space, and then constructs an optimum classification plane among the feature space. The kernel function only need meet the Mercer condition.

Let the kernel function $K(x_i, x_j)$, the corresponding dual optimization problem is:

$$Q(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i\bullet x_j),$$

$$\text{Subject to } \alpha_i\ge 0, i=1,\ldots,n \text{ and } \sum_{i=1}^{n} y_i\alpha_i=0, \quad (4)$$

The corresponding final decision function is:

$$f(x)=\operatorname{sgn}\{(\omega\bullet\chi)+b\}=\operatorname{sgn}|\sum_{i=1}^{n}\alpha_i^* y_i K(x_i\bullet\chi)+b^*| \quad (5)$$

2.3. One-class SVM algorithm to discover the outlier

Traditional C-SVM or V-SVM is supervised learning algorithms requiring labeled training data to create the classification rule. One class SVM algorithm presented by Scholkopf[7] could pose those unlabeled dataset. It is an unsupervised learning algorithm mainly used to the unusual detection and outlier discovery which is especially suitable for customer behavior observation.

The outlier discovery problem can be formulated as follows:

Given a unlabeled training dataset $x_1,...,x_l \in X$, and X is such a set where most data have a common feature while a small proportion belongs to outlier. One-class SVM attempts to find a function f(x) where most of the data f(x) takes the value +1, else it would take the value -1 on outlier. The main idea is that the algorithm will create a kernel function to transform the input space into a high dimensional feature space, and then find an optimum classification hyper plane to separate the data points from the origin with maximal margin, where the value of f(x) corresponding to each point will be decided by its location on which side of the classification plane. Similar to C-SVM, the solution to the problem is decided by the following quadratic optimum formula:

$$\min_{\xi \in R, \rho \in R} (\frac{1}{2}\|\omega\|^2 + \frac{1}{vl}\sum_i^l \xi_i - \rho, \quad (6)$$

Subject to $\omega \bullet \phi(x_i) \geq \rho - \xi_i, \xi_i, \xi_i \geq 0$

Where $v \in (0,1)$ is a parameter that controls the tradeoff between maximizing the distance from the origin and containing most of the data in the region created by the hyper plane and corresponds to the ratio of "outliers" in the training dataset.

Then the decision function for each point x_i,

$$f(x) = \mathrm{sgn}((\omega \bullet \varphi(x) - \rho) \quad (7)$$

will be positive for most examples x_i contained in the training in the training set.

If we introduce a Lagrange multiplier and rewrite formula (6) in terms of the Lagrange multipliers α_i, we can represent the formula (6) as

$$\min \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j K_\varphi(x_i, x_j),$$

subject to $0 \leq \alpha_i \leq \frac{1}{vl}, \sum_i \alpha_i = 1$ (8)

In terms of the Lagrange multipliers, the decision function is

$$f(x) = \mathrm{sgn}(\sum_i \alpha_i K(x_i, x) - \rho) \quad (9)$$

At the optimum, ρ can be computed from the Lagrange multipliers for any x_i such that the corresponding Lagrange multiplier α_i satisfies $0 < \alpha_i < \frac{1}{vl}$,

$$\rho = \sum_j \alpha_j K_\varphi(x_j, x_i) \quad (10)$$

One property of the optimization is that for the majority of the data points, $\alpha_i = 0$, which makes the decision function efficient to compute.

In the experiment we used the LIBSVM[8].

3. Apply the SVM over heterogeneous data sets

3.1. Definition of HVDM

A lot of high dimensional and heterogeneous data sets will be encountered while posing the bank transaction records. It is essential to solve the problem which inner product could not be defined over heterogeneous data sets so as that the SVM can not be applied directly. We proposed an improved RBF kernel based on the research results of Bernald Scolkopf[9] combining to the definition of distinct distance presented by D.Randall Wilson,et al. A kernel function based on the HVDM distance definition over heterogeneous data sets could extend the SVM to heterogeneous data sets. The definition is discussed as follows:

Def 1: heterogeneous data sets

Given data set X, $x \in X$,each x has m properties, each property is represented $x_i(i = 1,...m)$, and $x_i(i = 1,...l)$ takes continuous value while $x_i(i = l+1,...m)$ takes discrete value, such dataset is defined as heterogeneous data sets.

Def 2: Normalized difference

Two data x,y,x_l,y_l over heterogeneous data set are the lth continuous properties, the difference between the two points over the lth properties is defined as following:

$$normalized - diff_l(x,y) = \frac{|x-y|}{4\sigma_l} \quad (11)$$

where σ_l is the variance of the lth property over the data set.

Def 3: VDM (value difference metric)

Two data x,y,x_α,y_α over heterogeneous data set are the α th discrete properties, the distance between the two points over the α th properties is defined as following:

$$\text{normalized}-\text{vdm}_\alpha(x,y)=\sqrt{\sum_{c=1}^{C}\left|\frac{N_{\alpha,}x,c}{N_{\alpha,}x}-\frac{N_{\alpha,}y,c}{N_{\alpha,}y}\right|^2} \quad (12)$$

where $N_{\alpha,}x$ is the count of the data whose α th properties take the value x_α over the data set X, and $N_{\alpha,}x,c$ is the count of the data over the data set X whose α th properties take the value x_α as well as the output type is C. C is the whole output type of the data set.

Def 4 Heterogeneous distance function (HVDM)

Let $x,y\in X$, the HVDM distance between x,y is defined as

$$d_\alpha=\begin{cases}H(x,y)=\sqrt{\sum_{\alpha=1}^{m}d_\alpha^{\,2}(x_\alpha,y_\alpha,)},1,\\ \text{if } x_\alpha \text{ or } y_\alpha \text{ is undetermined}\\ normalized-vdm_\alpha(x,y),\\ \text{if the }\alpha\text{th property belongs to discrete value}\\ normalized-\text{diff}_\alpha(x,y),\\ \text{if the }\alpha\text{th property belongs to continuous value}\end{cases} \quad (13)$$

Such HVDM distance definition would expect a better effect and more efficient computation over heterogeneous datasets tested by D. Randal Wilson et al.[10]

3.2. An improved RBF kernel function based on HVDM

An RBF kernel function is defined over a non-null data set X as:

$$K(x,x_i)=\exp\left(-\frac{|x-x_i|^2}{\sigma^2}\right) \quad (14)$$

which takes the place of traditional norm value with distance, and maps X into a real number Hilbert space H:

$$f(\bullet)=\sum_{i=1}^{m}\alpha_i k(\bullet,x_i),g(\bullet)=\sum_{j=1}^{m'}\beta_j k(\bullet,x_j')$$

$$(m,m'\in N,\alpha_i,\beta_j\in R\ ,x_i,x_j'\in X) \quad (15)$$

An inner product is defined over the linear space:

$$<f,g>:=\sum_{i=1}^{m}\sum_{j=1}^{m'}\alpha_i\beta_j k(x_i x_j') \quad (16)$$

Since the improved RBF kernel definition involves the heterogeneous distance, and the HVDM distance formula (13) expects a more efficient and accurate measurement of distance, we adopt it as the distance computation, and the final kernel definition is:

$$K(x,x_i)=\exp\left(-\frac{H(x,x_i)}{\sigma^2}\right) \quad (17)$$

4. Experiment

The main embarrassments of Anti-money laundering data reporting system include its high false-positive rate, enormous data scale and unable to adapt to the changing situation due to its pre-established rule filtering mechanism. To improve it, we introduce a new method of customer behavior pattern recognition which the SVM is suitable to be applied. The main features to build a customer behavior profile include the transaction sum, frequencies of transaction, the business cycle, the alternation of business types, the change of the co-operating partners, etc. The essential purpose of intelligent data system is to determine the behavior is usual or unusual. We design the experiment to test the ability of the above algorithms to recognize the outliers from majority normal data.

4.1. Data set description

A real financial transaction record database set acquired from Wuhan Branch of Agriculture Bank in south central China is adopted in the experiment. It comprises 5000 accounts, 1.2 million records over 7 months. In this data set we mixed with 80 deliberately simulated unusual accounts whose business features are obviously deviated from those of peer groups. Since no real money laundering data could be acquired, and the purpose of intelligent system is not to find out the real money laundering action but the suspicious transaction for the staff to investigate.

The main structure of database is as below:

Table 1 The database structure of raw data sets

account	time	sum	virement	customer id
00001	10/15/02	108030	050310	13420101
00002	10/15/02	5600	120101	25090803
…	…	…	…	…

Some properties of customer such as the registered capital sum, the major business type, etc. may be added to the table by the customer id.

4.2. Training sets

To test the ability of unusualness detection, three key features vector is conceived including sum, frequency and business fluctuation. Frequency is the accumulation of transaction of every month. Business fluctuation may be acquired by accumulation of every month transaction sum.

We picked 32000 instances as the training set from the raw data. In this training set there are 318070 normal instances and 30 unusual instances. We divided them into three groups for 3 types of customer behavior including sum, frequency and punctuation.

4.3. Experiment Results

Before the experiment, we trained UC and SVM algorithm by using preprocessed training data. We used Libsvm software kit as the experiment tool, and improved the RBF kernel in order to adapt to the unusualness detection. Our attention is focused on the detection rate(DR) and the false positive rate(FPR) which are the urgent problems of current AML data reporting system. Here the detection rate is defined as the number of unusual instances detected by the system divided by the total number of unusual instances presented in the test set. The false positive rate is defined as the total number of normal instances that are incorrectly classified as unusual divided by the total number of normal instances.

However, due to the lack of good methods in choosing parameters, we could only use trial-and-error method. Two parameters C(incorrectly classification punishment parameter) and control factor g(g=$\frac{1}{\sigma^2}$) are assembled in the test. Table 2 shows the experiment results.

Table2 Unusual Detection Experiment Results

C	g	DR	FPR
50	0.5	63.27%	6.8%
100	1	69.13%	5.4%
200	2	67.8%	3.4%

5. Conclusions

The experiment result indicates that SVM is efficient for AML data reporting system reconstruction. The algorithm can get a fast speed and high accuracy by HVDM distance definition using RBF kernel. It does not rely on labeled and filtered training sets. Furthermore, the algorithm has a good performance in detecting unknown money laundering method which is essential in AML practicing. Finally, since it is simple and rapid, it can be used in real world system without a lot of modification.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (60205007), Natural Science Foundation of Guangdong Province (031558, 04300462), Research Foundation of National Science and Technology Plan Project (2004BA721A02), Research Foundation of Science and Technology Plan Project in Guangdong Province (2003C50118) and Research Foundation of Science and Technology Plan Project in Guangzhou City(2002Z3-E0017).

References

[1] The People's Bank of China, Administrative rules for the reporting of large-value and suspicious foreign exchange transactions, the People's Bank of China stipulation [2003] No 3

[2] Guo Jin Sheng, An analyzing to promote the AML efficiency in root bank branches, Journal of Jinan Finance, June, pp. 34-37,2004

[3] Li Dong Rong, AML and IS applications in Australia, Journal of Foreign Exchange Administration in China, August,2004,pp56-60,

[4] Richard J. Bolton and David J. Hand, Statistical Fraud Detection: A Review , http://stats.ma.ic.ac.uk/~rjbolton/fraudstatscipaper.pdf

[5] ZhangYan, A study on data mining technologies applied in finance, Journal of Computer Engineering and Applications, Vol 41,No,8, pp. 1413-1417,Aug.2004

[6] Deng Chao, A review to the anti-money laundering system in China, Journal of Southwest Nation College, Vol 21, No. 5, pp. 264-323,Oct.2004.

[7] Scholkopf.B,A short tutorial on kernels. Microsoft Research, Rech Rep:MSR-TR-2000-6t,2000

[8] http://www.csie.ntu.edu.tw/~cjlin/libsvm

[9] Bernhand Scholkopf, John Plattz. Estimating the support of a high dimensional distribution. Neural Computation,2001,13(7), pp1443-1472

[10] D. Randall Wilson, Tony R Martinez. Improved Heterogeneous distance functions. Journal of Artificial Intelligence Research, 1997,6(1), pp1-34

[11] Jason Kingdon, AI fighting money laundering, www.computer.org/intelligent, June,2004