Fraud Detection without Labels

Masato Nakai*1

*1School of Industrial Technology, Advanced Institute of Industrial Technology

At present,a compliance problem tends to be a company crisis, and the unrecognized fraud tends to result in huge losses. On the other hand, it is difficult for large companies to detect fraud on contracts because they are contracted by various person and in various place. Therefore companies need early to introduce a fraud detection system from the transaction history of contracts. We developed the fraud detection model without supervised label based on the anomaly detection and succeeded in detecting certified fraud contracts in the top rank.

1. Introduction

Although companies which continue the deficit settlement can avoid bankruptcies by lone or selling assets, there are many examples of faced with the bankruptcy crisis if compliance problems occur even if they continue good financial settlement. Therefore, the introduction of the fraud detection system in advance is urgent for many companies. Intentional fraudulent contracts are excessive contracts in cases where a part of the deposit is deceived and a part of the delivery is cashed. Such fraudulent contracts are made from a small scale so as not to be detected at first, but are often large at the time of detection. In large corporations, it is difficult to monitor every contract because of making contract between the various person and various business partners. Therefore, a model for detecting fraud early from a huge amount of transaction history data is required. When there are very few certified detected frauds, these are insufficient as supervised labels, and an anomaly detection model which does not rely on supervised labels is generally used. In general, anomaly detection uses a deviation from the average. However, in the case of fraudulent transactions, it is necessary to detect an anomaly value on the side showing fraud, and the anomaly detection model cannot be applied as it is. We developed the fraud detection model without labels based on the modified anomaly model using the transaction history of contracts, and we succeeded in ranking certified frauds in the top. In this paper, the detection criteria cannot be disclosed in detail due to the role of fraud detection, therefore we present only the adopted methods and results.

2. Fraud Detection Model

Generally there are two type of fraud detection model.

- type I Frauds that can be detected almost by claims due to unauthorized use. (Example) Unauthorized use of another person's credit card
- type II Frauds that can hardly be detected even if they occur. (Examle) Insurance fraud

Contact: Masato Nakai, School of Industrial Technology, b1617mn@aiit.ac.jp

In the case of type I, frauds can be almost detected, so a supervised learning model can be applied and the detection model is easier than type II. On the other hand, since it is difficult to detect frauds in the case of type II, there is no choice but to discriminate with an abnormal data pattern that shows signs of a fraud. For example, an insurance fraud may have a pattern in which an insurance is contracted with some premium that is excessive compared to general contracts, and a large amount of insurance is claimed immediately after the contract. In general, in the case of type II fraud, extraction rules for the fraud has been applied in the past, but in the case of enormous and diverse data, it is not practical to apply these rules. Now a practical way is to apply the extraction rule after narrowing down the target by the anomaly detection model. The problem is that an anomaly model detects both an under side and an over side. Therefore, it is necessary to determine the illegal side in the anomaly detection based on business knowledge.

2.1 Fraud Detection for Contract

This paper is aimed at detecting very few frauds for contracts in the distribution industry with thousands of people. Anomaly detection models may be applied because there are wide variety contracts. The company tends to trade to the limits of organizational rules to accept customer needs. Therefore, there are many cases where splits, changes, and cancellations occur in negotiations with the business partner. It is not easy to recognize whether the contract is valid. Also, because of getting the large contracts, some contracts may not be profitable and it may be difficult to determine the validity of the contract. Such fraudulent transactions tend to repeat differently from general transactions, and may be detected as abnormal transaction.

3. Method for Anomaly Detection

The book[Ide 15] shows lists the anomaly detection methods in Table 1 below. Since the fraud detection model we seek cannot expect supervised label, models using supervised label are excluded, and time series models are also excluded. And there is no guarantee that the data will be a gaussian distribution. As a result the methods that can be applied to this case are shown in the apply column in Table1.

Table 1: Methods of anomaly detection

apply	method	label	time	gauss
△	Mahalanobis dist.	×	×	○
	Naive bayes	○	×	×
◯	K neighbor no label	×	×	×
	K neighbor labeled	○	×	×
△	Mixture distribution	×	×	○
◯	One Class SVM	×	×	×
	Gaussian process	×	○	○
	Partial space	×	○	×
	Graphical model	○	×	○
	Density rate	○	×	×
△	VAE	×	×	○

The Mahalanobis distance is a model which detects peripheral points in the distribution in which difference with various scale of each data are corrected. The K-neighbor no-labeled method detects an anomaly points as sparse group which the zone of the radius space contains K points. However the setting of K and radius is difficult to use because of dependence on experience. A mixed distribution is shown in Fig. 1 below. The Mixed distribution sequential estimation method are detected peripheral points on each distributions. VAE is Variation Auto Encoder[Kingma 14] which can detect peripheral points in space spanned by latent variables. This method is mainly applied for image anomaly detection. Even in mixed distribution the Mahalanobis distance can be applied shown in the Fig. 2. This result indicates that the surrounding outliers can be sufficiently detected except for mixing zone. One Class SVM [Bishop 06] is a model that maps the space in high dimension by the kernel function [Plat 99] so that the peripheral points are as discrete as possible. In the Fig. 3. One Class SVM anomaly detection recognizes two mixed distributions, and it can be seen that the peripheral points in mixed zone of two distributions can also be recognized and have high accuracy. In both figures 2 and 3, the Z-axis indicates the degree of abnormality, and the □ points in both figures indicate 10% high peripheral points.

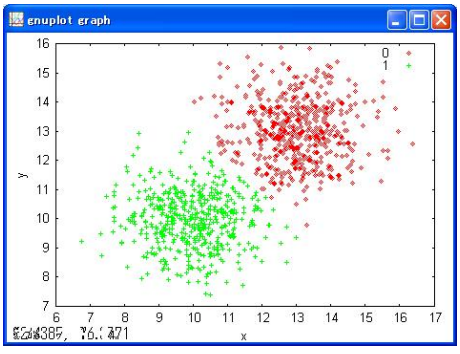

Figure 1: Mixed Gaussian distribution

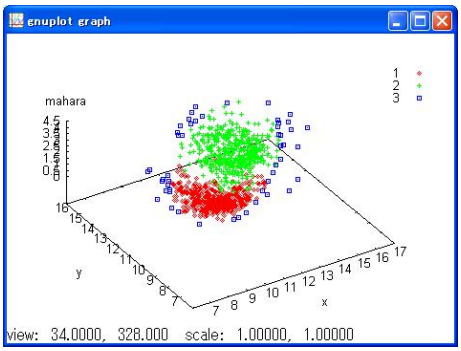

Figure 2: Mahalanobis distance on mixed distribution

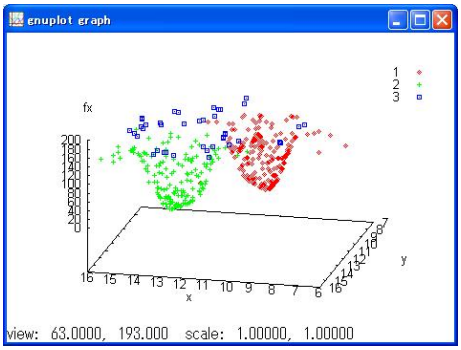

Figure 3: One class SVM on mixed distribution

From the above, one Class SVM is the most appropriate. But if the number of data is 1000 or more, it is practically impossible to resolve and can only be calculated with an approximate solution for large data[Mochihashi 15]. There are thousands of data in this our mission. As a result, even if the data has complex multimodal distribution as shown Fig. 2, the Mahalanobis distance can roughly detect peripheral points, so we finally adopted the Mahalanobis distance as an abnormal value detection model.

4. Fraud Detection Method

There are both wholesale and retail sales in the distribute business. The former is a large-scale transaction with a large partner, and this transaction of contract is systematized, so there is few opportunity for frauds. On the other hand, in the latter counter parties are small companies or individuals, so there are some opportunity for frauds in the negotiation process. In general, fraudulent contracts are occurred by the complex trade in retail and are done inconspicuously. Here, it is required using statical models to detect frauds that cannot be detected by humans.

4.1 Preprocessing and Selection features signing Frauds

Contract and transaction history data are distributed in RDB of the huge business system. Preprocessing is required to collect data showing signs for frauds from RDB and edit to appropriate data that can be easily analyzed for data consistency and abnormality.

As a result 4 items ware selected for feature values signing frauds in the preprocessing. Since there was no supervised label, such a selection was based on inconsistencies and abnormality recognized by business knowledge, The selected criterion are not disclosed in detail due to the role of the fraud detection. We only show the selected features in about as follows

- Inconsistency between shipped items and contract amount
- Abnormal volume in similar contract
- Inconsistency between volume and partner size
- Abnormal trade span in similar contract

5. Resutl of Fraud Detection

We calculated above 4 feature values of all contracts and the following two methods were applied as shown below and Fig. 4.

- Mahalanobis method : Ranking according to Mahalanobis distance using 4-dimensional data composed by 4 features indicating signs of fraud. However, since Mahalanobis distance is evaluated equally for both under and over distance, we selected only contract which has larger amount than the average in each organization. The result is shown in Table 2
- Overall ranking method : Mahalanobis distance is applied to each of the 4 feature values, and ranking is applied according to each distance. But we ignored the detected lower side because of safe side. And these 4 rankings were totaled to make the overall ranking as shown in rightmost column in Table 3. We sorted descending overall ranking and ranked contracts as shown in leftmost column.

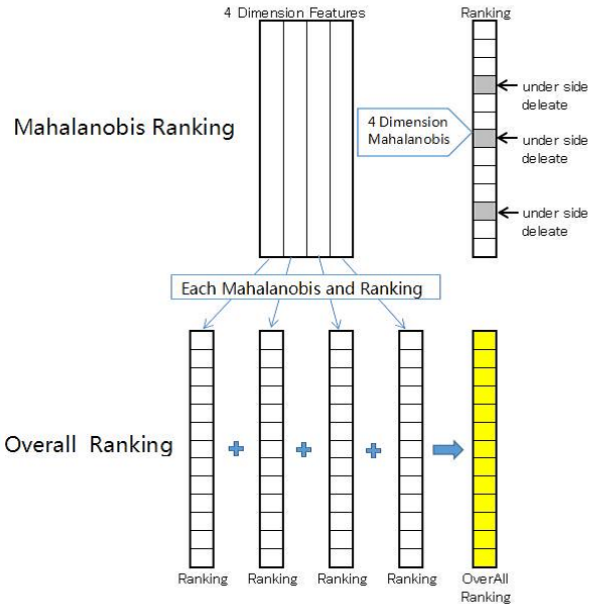

Figure 4: Method of Fraud Detection by Mahalanobis distance

In these tables higher rank shows that the possibility of fraud is higher. As indicated by ⊗ , there are two fraud cases that are currently certified within the company. We evaluated whether these methods are ranked higher.

Table 2: ranking by mahalanobis distance

rank	contract	mahalanobis
1	08263	22.234
2	40882	20.037
3	31153	17.098
15	77728⊗	8.217
83	50364⊗	3.557
1054	78000	0.320
1054	55520	0.329

Table 3: overall ranking of each mahalanobis rank

rank	contract	sd	wc	qph	mis	all
1	77728⊗	2780	2765	2761	2341	10603
2	50364⊗	2779	2514	2696	2307	10296
3	02359	2766	2727	2466	2270	10229
4	37759	2740	2777	2697	1811	10095
5	52233	2758	2514	2471	2337	9909
6	26155	2576	2765	2647	1921	9884
2788	68865	2	12	3	0	17
2789	91703	2	12	0	0	14

Overall ranking method showed that two certified fraud was ranked top.

6. Consideration

Fraud detection at Mahalanobis distance using 4 dimension shows lower ranking of certified fraud contracts. The peripheral value of Mahalanobis distance is calculated as a deviation from the average, and over and under are treated equally. On the other hand, the overall ranking method applies the Mahalanobis distance for each variable and ranks by the sign of fraud ignoring lower side, and the sum is thought to make the sign of fraud more prominent. For overall evaluation of multiple rankings, it is appropriate to weight each feature amount. However, since there are very few fraudulent cases at present, it is considered difficult to estimate an appropriate weight.

7. Conclusion

If the number of fraudulent transactions is extremely few, modeling based on supervised label is impossible, so we selected features sining frauds by business knowledge, and ranked each feature by using unsupervised anomaly model, and totaled these rank as overall rank. As a result we could rank certified fraud cases at the top. On the other hand the detection of anomalies such as Mahalanobis distance alone evaluated the overs and unders equally, so it did not become a significant model.
The remaining issues are as follows.

- Applying one Class SVM for large-scale data
- wholesale has large transaction volumes, so if there is an illegal contract, the damage will be great. It is a future subject whether abnormality detection model can be applied also in this field.
- It has been found that most of contract are legitimate contracts even if they are ranked high in our fraud detection model. It is necessary to consider the legitimate reason and reflect this reason in our model to eliminate legitimate contracts and to make higher accuracy of fraud detection.

References

[Mochihashi 15] Mochihasi.D,Base of Gaussian Process and Unsupervised Lerning,2015

[Ide 15] Ide T,Anormaly Detection and Change Detection,Kodunsha,2015

[Bishop 06] Bishop C.M, Pattern Recognition and Machine Learning 7.1.3,Springer, 2006

[Plat 99] Platt, J. C. Fast Training of Support Vector Machines using Sequential Minimal Optimization,1999

[Kingma 14] Kingma D. P, Welling M. Auto-Encoding Variational Bayes,arXiv:1312.6114,2014

[Ishigima 18] Ishigima T. Method of data analysis corresponding to cases of accounting fraud at overseas subsidiaries,2018