Sigma J Eng & Nat Sci 8 (2), 2017, 169-174

Publications Prepared for the Innovations on Intelligents Systems and Applications Symposium ASYU 2016
Akıllı Sistemler ve Uygulamalardaki Yenilikler Sempozyumu ASYU 2016 için Hazırlanan Yayınlar

Research Article / Araştırma Makalesi

FRAUD RISK MANAGEMENT FOR GSM SYSTEMS

Onur TÜFEKÇİOĞLU, Göksel BİRİCİK, Banu DİRİ*

Yildiz Technical University, Computer Engineering Department, Yıldız-ISTANBUL

Received/Geliş: 05.12.2016 Accepted/Kabul: 11.12.2016

ABSTRACT

GSM operators use manual rule based fraud detection systems in order to identify fraud operations that cause risks. An operator reported very low fraud detection success by using a manual system. In this study, we evaluated the impact of machine learning methods on the performance of fraud detection. Machine learning methods, including K-Nearest Neighbor, Naïve Bayes, Random Forest and Support Vector Machines are utilized in order to increase the performance of the existing manually operated fraud detection. We constructed our dataset by using the cases of the observed GSM operator, where an action is taken upon a suspicious fraud activity. We defined feature set for these samples, which occurred in a period of two months. Besides the mentioned machine learning algorithms, we also performed feature selection and comparatively evaluated their performances. The experimental results show that, machine learning approach doubles the fraud detection performance, even at the worst model.
Keywords: Fraud detection, GSM systems, machine learning, feature selection.

1. INTRODUCTION

The primary purpose of GSM operators is to establish communication between people. Facts like competitive market conditions, developing technology and inflating demands compel the operators to produce new services in order to meet the rising expectations of their customers as well as to create novel income opportunities. Besides these facts, number portability is the most competitive action between GSM operators nowadays. When the customer volume is saturated, the incomes of an operator tend to decrease. Hence, value added services become crucial to protect the profitability via maintaining the satisfaction of the saturated customers.

Value added services were first developed within GSM operators and quickly switched to outsourcing in order to increase the variety of services and reduce risks. In the beginning, the outsource companies were either local or international. Afterwards, the majority turned into global scale, where they can easily handle and overcome legal issues about their provided contents through less strict countries, where the profits of the operation is shared via contracts with the GSM operators. While the operations are moved to the global scale, both of the offered services and payment methods (e.g. mobile payment, credit cards, contactless payment, mobile wallet) expanded, without audit and supervision of the operators. This situation caused serious

* Corresponding Author/Sorumlu Yazar: e-mail/e-ileti: banu@ce.yildiz.edu.tr, tel: (212) 383 57 82

fraud operations [1-3]. In some cases, fraud organizers used fake identity cards for establishing plenty of fake GSM line subscriptions. Using these fake lines, the organizers dial or text the value added service lines where the operator pays to the service provider for the content. The operators in fact have been paid for fraud services, where the fake customers made use of these services.

The huge cost of fraud operations directed the GSM operators to use fraud detection systems. There are many commercial fraud detection implementations on the market that can cost up to million dollar levels. The performance of these systems usually depend on the rule sets, written by the specialists of the GSM operators. The operator which we focused in this study has a 29.7% success rate for fraud determination. As this ratio is very low, we occupied machine learning methods to increase performance. In this study, we present the outcomes and results of the effects of machine learning [4-6] and feature selection methods on fraud classification.

Rest of the paper is organized as follows. In the next section we introduce our dataset. Section 3 describes the fraud detection system which is already used by the selected GSM operator. We introduce our system and methods in Section 4. Finally, we conclude.

2. DATA SET

We investigated 5,641 cases from the selected operator, between August and September of 2014, where one of two actions is taken on the line due to suspicious fraud activity. The first action is voice or message service deactivation for the line. The second action is full deactivation, including voice, message, internet services and all related issues that belong to the line subscription. The lines with an action taken have two usage types: voice and message. The distribution of our cases over action and usage types are given in Table 1.

Table 1. The distribution of fraud cases where an action is taken.

Service	Service deactivated	Full deactivated	Total
Voice	3743	1019	4762
SMS	348	531	879
Total	4091	1550	5641

If a subscriber with a deactivated line pays the fees and reactivates the line, the subscriber is classified as a *candidate*. The subscribers who do not pay their bills are classified as *fraud*. The distribution of the classified subscribers on services and deactivation types is given in Table 2.

Table 2. The distribution of classified subscribers.

		Candidate	Fraud
Service	Voice	3392	1370
	SMS	572	307
Deactivation	Service	2954	1137
	Full	3964	540

Our dataset consists of 5,641 samples, where the subscribers are either classified as a *candidate* or *fraud*. The samples have 76 features. We selected the most effective 10 features in our tests. The selected features are:

- *Subscribed_days*: The number of staying days of the line in the operator
- *Club:* Whether the subscriber is VIP or not
- *Payorder*: Whether the line has automatic payment order via bank or not
- *International_status:* Roaming flag for the line
- *Csi_payment_count:* Number of paid invoices by the registered credit card

- *Csi_payment_amount:* Total payment made by the registered credit card
- *SMS_cdr_number_over_5TL:* Number of SMS usage over 5 TL
- *Number_of_calls:* Total number of calls
- *Total_call_duration_over_50TL:* Total duration of calls that are rated over 50 TL
- *Number_of_call_seconds_rated_over_0.05TL:* Number of calls, that each second is rated over 0.05TL

3. CURRENT SYSTEM PERFORMANCE

The performance evaluation of the fraud system currently used in the selected operator is measured with (1) [5]. The classification success ratios of the current system over action and usage types are given in Table 3. Table 3 shows that average overall fraud classification performance of the current system is 29.7%. This success rate is very low for GSM operators having high number of customers.

$$\text{Success Rate} = 100\times\frac{fraud}{fraud+candidate} \quad (1)$$

Table 3. Success rates of current system.

Service	Service deactivated	Full deactivated	Average
Voice	27.6%	32.8%	28.7%
SMS	29.0%	38.8%	34.9%
Average	27.7%	34.8%	29.7%

After number portability has started, subscribers benefited from this opportunity for factors like better service, campaign advantages, operator dissatisfaction, lower invoices, and better cell reception. Especially, the subscribers who are classified as *candidates* by fraud detectors tend to port their cellular lines to other operators [4] and they are also known as churns. For this reason, it is very important for the GSM operators to protect the profitable level in customers by minimizing the number of false positives in fraud detection.

In order to measure the performance, we randomly selected 38,988 subscribers that have never deactivated as the validation set and labeled these as *normal.* We used (2) to calculate the subscriber loss rate. The churn rates based on subscriber classes are given in Table 4.

$$\text{Subscriber loss rate} = 100\times\frac{\#\ of\ migrated\ subscribers}{total\ subscribers} \quad (2)$$

Table 4. Churn distribution among subscriber classes.

Subscriber classes	# of subscribers	# of churns	Success rate	Ratio
Normal	38,988	36	0.0923	1
Candidate	3,964	179	4.51	49
Fraud	1,677	95	5.66	61
Candidate + Fraud	5,641	274	4.85	52

Table 4 shows that the churn rate of the subscribers on whom an action is taken and labeled as candidates is 49 times higher than the normal subscribers. This ratio proves that false positive samples, where an action is taken on a normal subscriber, is a serious problem for the GSM operators. The new operator information of the churn subscribers are given in Table 5.

Whilst the churn subscribers who belong to normal class migrate equally to alternative operators, operator Z is chosen two times more than operator Y by the *candidate* and *fraud* subscribers. This shows that the prerequisites of Operator Z satisfy the expectations of the churn subscribers.

Table 5. Ratio of the alternative GSM operators chosen by churn subscribers of the selected operator. The names are masked.

Subscriber classes	Operator Y	Operator Z
Normal	50.0%	50.0%
Candidate	35.2%	64.8%
Fraud	32.7%	67.3%
Candidate + Fraud	34.4%	65.6%

4. PROPOSED FRAUD DETECTION SYSTEM

As we have reported in Section 3, the fraud detection performance of the current system is about 30%. This ratio also tells that the ratio of wrong deactivations is about 70%, which causes a very serious subscriber dissatisfaction. We have seen in Section 3 that the churn rate of these subscribers are 49 times higher than the normal subscribers, resulting in a severe customer loss for the operators. For this reason, we occupied four different machine learning methods to find the most suitable model for the reduction of subscriber losses and to take the most suitable action in a possible case of fraud.

4.1. Methods

We used four widely used and well known methods for classification, namely, k-Nearest Neighbors (k-NN), Naïve Bayes (NB), Random Forest (RF) and Support Vector Machines (SVM). Primarily we had 76 features on our models. In order to increase the performance, we applied feature selection and chose 10 features, which were described in Section 2. We used the open-source Weka implementation of Correlation-based Feature Subset Selection for feature selection [3, 7, 8]. While searching the attribute subsets, we used the Best-First search, which is a greedy hill climbing search augmented with a backtracking facility. We selected 10-fold cross validation for performance evaluation.

4.2. Experimental Results

We present the results of our evaluations on action and service basis in the succeeding subsections.

4.2.1. Results Based on Action

The actions taken by the GSM operator are service or full cancellation on the subscriber's line. The success rates of the proposed models are given in Table 6 for 76 features and in Table 7 for 10 selected features.

Table 6. Success rates of the proposed models based on action type, using 76 features.

Methods	Service deactivation	Full deactivation	Service + full deactivation
k-NN	82.03%	62.83%	76.15%
NB	54.14%	63.41%	54.42%
RF	79.78%	68.64%	76.67%
SVM	87.11%	69.87%	81.59%

Table 7. Success rates of the proposed models based on action type, using 10 features.

Methods	Service deactivation	Full deactivation	Service + full deactivation
k-NN	82.74%	61.67%	77.45%
NB	57.95%	57.35%	67.50%
RF	86.94%	64.00%	80.69%
SVM	86.67%	68.77%	80.64%

With the occupied feature selection [9], we chose 10 features out of 76. When we observe the results, we see that SVM and RF are the two most successful methods. We propose to favor RF on SVM, based on their running times. We can also infer that SVM works better with the full feature set, while RF record the best results on service and full deactivations.

4.2.2. Results Based on Service

We investigated the services based on voice, message and their sums. The success rates of the proposed models are given in Table 8 for 76 features and in Table 8 for 10 selected features.

Table 8. Success rates of the proposed models based on services, using 76 features.

Methods	Voice	SMS	Voice + SMS
k-NN	79.73%	63.70%	76.15%
NB	57.29%	64.05%	54.42%
RF	78.24%	68.03%	76.67%
SVM	85.15%	66.89%	81.59%

Table 9. Success rates of the proposed models based on services, using 10 features.

Methods	Voice	SMS	Voice + SMS
k-NN	82.74%	61.67%	77.45%
NB	57.95%	57.35%	67.50%
RF	86.94%	64.00%	80.69%
SVM	86.67%	68.77%	80.64%

Even if we select 10 features out of 76 with [9], there is no significant change in performance with respect to the feature counts. We see that the highest performance is achieved with SVM using 76 features on all service types. SVM also recorded the best accuracy on the reduced feature sets, even if they are lower than the records obtained with 76 features.

4.3. Performance Comparison

When we compare the current system with the proposed machine learning methods using 76 original and 10 selected features, we get the results in Table 10. We note that the current system is operated by specialists, without using the derived features.

Table 10. Performance comparison for the current system and the proposed models.

Model	76 features	10 features
k-NN	76.15%	77.45%
NB	54.42%	67.50%
RF	76.67%	80.69%
SVM	81.59%	80.64%
Current System	29.70%	

We see that even the worst method, NB, performed at least twice better than the current system. RF and SVM has correctly classified the fraud subscribers with accuracy over 80%. These results show that we can prevent the misclassification for about the 50% of the potential churn subscribers.

5. CONCLUSIONS

GSM operators must satisfy their customers in every aspect for the proper operation, while guarding the fraud attempts. Thus, fraud detection is a very important issue for GSM operators in order to decrease churn rate, as the false positive subscribers tend to migrate to alternative operators. In this study, we investigated the fraud detection system of a GSM operator in Turkey and analyzed the churn subscriber data along the manually operated fraud detection system. It is clear that when the number of detected false positive fraud subscribers decrease, customer satisfaction will be convincing in middle to long term projection. What is more, the churn migration that have a monthly average of 100 subscribers will be suppressed. The proposed system also helps to prevent about 1500 unnecessary deactivations in a month, increasing the overall fraud detection performance.

REFERENCES / KAYNAKLAR

[1] Magnify, Fraud Focus Advanced Fraud Detection, White Paper, Chicago, 2002.

[2] R. Bolton, D. Hand, "Statistical Fraud Detection: A Review (With Discussion)", Statistical Science, 17(3): 235-255, 2002.

[3] R. Bolton and D. Hand, "Unsupervised Profiling Methods for Fraud Detection", Credit Scoring and Credit Control VII, 2001.

[4] P. Burge, and J. Shawe-Taylor, J, "An Unsupervised Neural Network Approach to Profiling the Behavior of Mobile Phone Users for Use in Fraud Detection", Journal of Parallel and Distributed Computing, 61: 915-925, 2001.

[5] M. Cahill, F. Chen, D. Lambert, J. Pinheiro and D. Sun, "Detecting Fraud in the Real World. Handbook of Massive Datasets", 911-930, 2002.

[6] B. Bhargava, Y. Zhong and Y. Lu, "Fraud Formalization and Detection", Proc. of DaWaK2003, 330-339, 2003.

[7] R. Wheeler and S. Aitken, "Multiple Algorithms for Fraud Detection", Knowledge-Based Systems, 13(3): 93-99, 2000.

[8] S. Maes, K. Tuyls, B. Vanschoenwinkel and B. Manderick, "Credit Card Fraud Detection using Bayesian and Neural Networks", Proc. of the 1stInternational NAISO Congress on Neuro Fuzzy Technologies, 2002.

[9] D. Foster and R. Stine, "Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy", Journal of American Statistical Association, 99: 303-313, 2004.