Liet. matem. rink., **44**, *spec. nr.*, 2004, 332–338

Data mining process for fraud detection in mobile communication

Jelena MAMČENKO, Regina KULVIETIENĖ (VGTU)
e-mail: jelena@gama.vtu.lt, regina_kulvietiene@gama.vtu.lt

Abstract. Without dependence from a sort of activity (sale, rendering of services, etc.) the using of data mining methods can bring the certain advantage.

Fraud detection methods of data mining can be applied to this problem quite readily. Three important elements of a data mining application/solution are present. These are the ability to handle large amounts of data, suitable methods and algorithms, and the availability of domain expertise.

Keywords: data mining, intelligent miner for data, neural and demographic clustering, fraud.

1. Introduction

Telecoms fraud business is booming. According to the Forum for International Irregular Network Access (FIINA), organised crime is making a staggering $55 billion a year from fraud in telecoms worldwide [1].

Fraud is specifically used here in the sense of 'fraudulent use of telecommunications services' rather than 'fraudulent activities by means of telecommunications networks (hacking etc.)'. The types of telecommunications fraud discussed here are detrimental to the telecommunications operators, as these deliver services which they are not or only partly being paid for, resulting in loss of revenue.

The size of the fraud phenomenon, in terms of financial damage to operators is hard to estimate. Numbers quoted are 3–6% of revenue worldwide [2]. It is clear that growing fraud levels must be a concern to many operators; especially since the margins on telecommunications services are dropping in the liberalised markets [2, 3]. There is a vested interest of telecom operators limiting this type of fraud.

By virtue consumers growth of services of cellular communication, Internet users, expansions of programs of credit cards releasing, telecoms fraud becomes significant and for our country. This conclusion proves to be true precisely shown tendency on migration of criminality in a cyberspace.

Two general types of fraud are [1]:

- Call sell operation/toll fraud.
- Premium Rate Service Fraud.

The main ways of committing fraud are:

- Subscriber fraud. A subscription is obtained through the regular subscription process under a false or stolen identity. It is also possible that employees of the telecom operator participate in this type of fraud, e.g. by deliberately skipping procedures to check a new subscriber's identity.

- Surfing. This includes several forms of unauthorised use of facilities.
  - Cloning of handsets. Identities, telephones or other attributes are being duplicated. This kind of swindle is based on introduction in system of services of the swindler under a kind of the legal user of services. Such swindle is the extremely significant, besides granting cellular communication, also in the industry of credit cards, the Internet services and a number of other areas [4].
  - Calling card fraud. This includes theft of or fraud with PIN-codes and recharging cards.
  - Misuse of hardware. This includes several ways to break into the telecommunications network.
  - Teeing-in other subscriber' s line by physically connecting to the line.
- Mobile fraud in mobile communications open up several new forms of fraud. Specific types of fraud, which are committed using mobile phones, are the following. The simplest form is plain theft of mobile phones. Roaming fraud another form; expensive calls are being made from a foreign country, making use of the delay in billing these calls in the country where the phone is registered. Recharging or copying prepaid cards is also reported. Several types of fraud of connect-through services exist as well.

Data mining involves extracting relevant features from a large and unordered collection of data [5, 6, 12].

Here we describe the early fraud detection carried out by means of premium telephone numbers: a telephone provider is charging the telephone company for calls which never will be paid by the caller. We discuss the method on how data mining supports the detection of possibly fraudulent calls, and present a simple solution of the detection request [7].

The solution described here permits the early detection of fraud carried out by means of premium telephone numbers. By premium (phone) numbers we understand such numbers as offering services like:

- expert hot lines for computers or other technical equipment,
- advice for insurance or juridical questions,
- stock market tips,
- or phone sex.

The prices that are charged, when somebody calls a premium number, are normally higher than "ordinary" phone calls, around two euro per minute. The provider that offers such a service gets paid a high share of the total rate immediately (within days) from the phone company. The phone company itself charges the caller for the whole amount (on a monthly payment).

Given that scenario, the fraud could be carried out in the following way: the (fraudulent) company that offers a premium number service cooperates conspiratorially with a partner. This partner makes very frequent and long phone calls via one or a few other phone numbers to the (expensive) premium number.

Therefore, a large amount of call charges accumulate within a few weeks. Often fraudulent partners perform the phone calls by using a computer or an automatic dial-

ing device multiplying the number of calls and minimizing their efforts. The service provider receives his comparably high share of the call charges from the phone company within a short time. When the phone company tries to recover its expenses from the customer (or the conspiratorial partner), the company becomes aware that the customer used a wrong name, or has disappeared, or denies to pay, and his conspiracy with the service provider cannot be proven.

Besides the detection of fraud based on conspiracy, the telephone company may be also interested in addictive phone call behaviour, mostly occurring with the phone sex service. In such a scenario, the phone company tries to check in time, whether the caller can still pay his high phone fees, by sending to him an additional intermediate invoice.

2. Data mining application

The phone company records for each phone call a Call Detailed Record (CDR). The CDR is stored in a database and normally consists of about 50–100 fields. This database is the starting-point for the detection of fraud.

Table 1 presents the most important fields from a total of about 50 fields in a CDR.

To create data model we have to take the raw data that we collect and convert it into the format required by the data models.

For fraud detection of premium number callers, one table is used, the CDR table. An example is given in Table 2.

In a series of experiments we found out, that a high share of fraud cases can be recognized quite early, for instance, by the end of the first week of operation of a

Table 1. Variables from CDR

Field	Description
CALLER_ID	Identification of the caller
PREMIUM_ID	Identification of the premium number called
Start Date	Date, when phone call has started
Start Time	Time, when phone call has started
Duration	Duration in seconds
Charges	Charges

Table 2. Call detailed record database

CALLER_ID	PREMIUM_ID	START DATE	START TIME	DURATION (SEC)	CHARGES
2561501233	0815691381	2003–01–09	22.40.12	123	4.01
2538366458	0815656545	2003–01–09	10.42.36	37	1.06
2561501233	0815691281	2003–01–09	22.43.00	138	2.21
7857107555	08152232232	2003–01–09	22.28.49	40	0.97
...	...	...	...	...	...

fraudulent premium number. From the CDR fields a weekly connection view containing aggregated data must be generated containing all information useful for indication of fraudulent calls.

Preparing and aggregating the data to build a weekly connection view is done by defining several tables collecting measured values for the whole week (as sum of costs, duration of calls, average duration of calls, and so on.

The connection view will be calculated and derived automatically from the CDRs. It contains the weekly update for every connection. The fields of the view connections are described in Table 3.

The mining technique we use to recognize subsequently connections with a high fraud probability is to perform a demographic clustering of the connections view [4, 8].

The segmentation technique is chosen to identify customer behaviour. Segmentation is also known as the clustering technique, which is mainly used for the customer segmentation [9]. Clustering is a discovery data mining technique. What people usually mean when they talk about discovery data mining is that it does not require any prior knowledge of customer segments to make decisions [11]. This technique groups customers who have similar characteristics, while at the same time maximizes the difference between different groups of customers.

There are two segmentation techniques that you can choose when you are doing segmentation: demographic clustering and neural clustering [7, 12].

Neural clustering is typically used when most of the variables are numeric and *demographic clustering* is generally used when most of the variables are categorical, because demographic clustering has unique functionality to handle categorical attributes without transforming them into numerics. However, These are just general observation based on the characteristic of the algorithm and both can be used to complement

Table 3. Fields used for fraud detection

Field name	Description of the field
SUM_DUR	Whole duration of all calls on a specific connection, from a specific caller to the premium number
NO_CALLS	Number of calls on the connection
REL_DUR	Indication whether the connection has an extraordinarily high share in the turnover generated by all connection with the same premium number.
SUM_COST	Call charges for the connection
MAX_DUR	Duration of the longest call on the connection
VAR_DUR	Variance of the call duration on a connection
NO_CLRS	Number of all different connections to the premium number

Plus, two additional fields are derived directly from the CDR.

Field name	Description of the field
CALLER_ID	ID of the caller
PREMIUM_ID	ID of the premium number

each other. In this case, because we are using only numeric variables to represent the customer behaviour for behavioural segmentation, neural clustering (Kohonen Feature Map) was used.

Demographic clustering has initially been developed to work with demographic data which typically consist of categorical variables. For this reason the technique works best with the categorical variables. *Neural clustering* works best with continuous variables and treats categorical variables as if they are numeric. Another difference is that for neural clustering the user should specify the number of clusters that they want to derive. For demographic clustering the numbers of clusters are automatically decided based on the measure representing how similar the records within the individual clusters should be. These two techniques solve the same problems from different views, both can be used to complement each other and to gain confidence that the segmentation produced has a valid and optimal result. Usually, the clustering algorithm finds a large amount of clusters. In this example we came up with about 50 clusters. For each cluster the value distribution of the derived fields (which were used for the clustering) is shown as well as the two additional fields, *caller_id* and *premium_id* (not used to compute the clusters). Several of these given clusters describe groups of customers with unusual call behaviour.

Experience shows, that mostly smaller clusters contain the connections with a high fraud probability. The details of the connections located in the suspicious clusters can be (optionally) inspected by reporting tools.

A closer look at Fig. 1 where one caller calls a premium number excessively; the duration and the costs of these calls are unusually high, so here we have very likely identified an "addicted caller". Normally many callers of this kind are not able or not willingly to pay their services. Special payment conditions for these people could prevent loss.

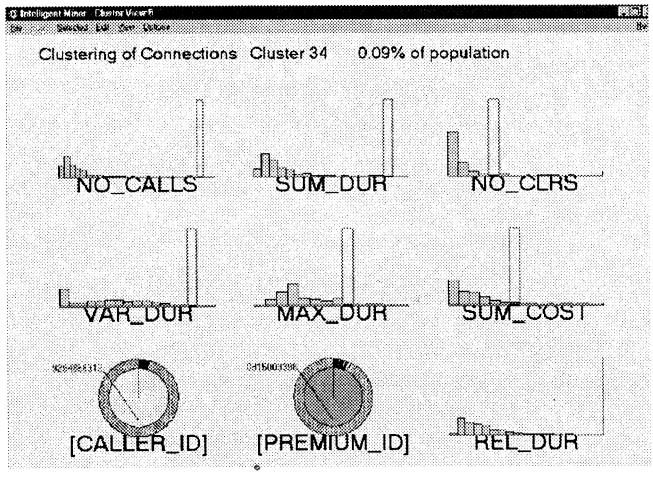

Fig. 1. Cluster 34: connections indicating phone addiction.

Id	Name	Minimum Value	Maximum Value	Mean	Standart Deviation	Distance Unit
5	NO_CLRS	217	217	0	0	117.063

Fig. 2. Cluster 34 ("Phone addiction"): NO_CLRS details.

This cluster contains only one connection, where the premium number has a lot of callers. Again the details view of the field, number of callers ("NO_CLRS"), in Fig. 2, we find the same high minimum and maximum value; here the value is 217. In this example, one of them (CALLER_ ID=9294866312) calls the premium number excessively (NO_CALLS); the whole duration of all his calls (SUM_DUR) as well as the costs (SUM_COST) are very high. A conspiratorial relation between the frequent caller and the provider of the premium number service is less likely due to the existence of other callers, who bring the provider a high revenue.

In this case, it is more likely that the frequent caller is a phone-addicted person. The conspiracy results can be exploited by selecting the customer records indicating possible fraudulent callers: If Intelligent Miner (IM) for Data runs the clustering settings object, where "create output table" (and the field names for the cluster reference, the scoring field and the *confidence* field) have been defined, the miner creates an output table for all records. This output table now contains an additional field in which a *cluster_id* for the best fitting cluster is given. The additional confidence field indicates the reliability of the fit. If the *confidence* value (range 0–1) is below or near 0.5, another grouping for this record could be done as well and maybe it is not so reasonable to expect that the record should be considered to the respective cluster.

Now several treatments could follow:

- charge the callers corresponding to these records are referring to in a weekly billing;
- limit the calls with premium numbers for that caller;
- force these callers to open a deposit account against the billing will be done;
- confirm the correct subscriber address and the employment status.

A second wrap up could be done if historical data about fraudulent callers are available. In that case we could score new customers by their properties, indicating on a semantic level that problems with this customer could arise. For example special connections may already indicate a possible fraud.

Conclusions

Data mining technologies are designed to look for subtle patterns in large data stores. The large amounts of data generated by telecoms transactions are thus not likely to be a problem. In terms of Data Mining methods, there are already numerous algorithms for detecting co-occurrences in datasets. Integrating analytical methods, coupled with domain knowledge can result in a data mining application that can detect and minimize fraudulent practices in the telecoms.

The example presented was part of a real project running with a major European telephone company. The theme of the project was to install an automated detection of

fraudulent behaviour on their premium number services, such as weather and sports scores on demand, or news and financial/investment data.

The operative use of the first version of the solution started at the phone company several weeks after the development of the first mining model. The customer uses, since that time, the solution permanently and with great success in several locations. Each week about 5,000,000 new CDRs are analyzed, and fraud attempts within the scale of tens of thousands of euro are detected and prevented.

References

1. M. Collins, Telecommunications, Crime – Part 3, *Computers & Security*, **19**(2), 141–148 (2000).
2. Telecoms fraud in the cellular market: how much is hype and how much is real? *Computer Fraud & Security*, **6**, 11–14 (1997).
3. P. Hoath, What's new in Telecoms fraud? *Computer Fraud & Security*, **2**, 13–19 (1999).
4. P. Hoath, Telecoms fraud: the gory details, *Computer Fraud & Security*, 10–14 (1998).
5. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, Knowledge discovery and data mining: towards a unifying framework, in: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland (1996), pp. 82–88.
6. R.S. Michalski, K.A. Kaufman, Data mining and knowledge discovery: a review of issues and multistrategy approach, in: *Machine Learning and Data Mining: Methods and Applications*, West Sussex, England (1998), pp. 71–105.
7. K.I. Daskalaki, M. Goudara, N. Avouris, Data mining for decision support on customer insolvency in telecommunications business, *European Journal of Operational Research*, **145**(2), 239–255 (2003).
8. R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A. Verkamo, Fast discovery of association rules, in: U. Fayad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (Eds.), *Advance in Knowledge Discovery and Data Mining*, AAAI Press/MIT Press (1996), pp. 307–328.
9. D. Hand, H. Mannila, P. Smyth, *Principles of Data Mining*, MIT Press (2001).
10. N. Chalamaiah, Data warehousing and data mining for telecommunication, *Computer Communications*, **23**(1) (2000).
11. J. Shawe-Taylor, K. Howker, P. Burge, Detection of fraud in mobile telecommunications, *Information Security Technical Report*, **4**(1), 16–28 (1999).
12. D.J. Hand, Statistics and data mining: interesting disciplines, *ACM SIGKDD Explorations*, **1**(1), 16–19 (1999).

REZIUMĖ

J. Mamčenko, R. Kulvietienė. Duomenų gavybos procesas apgavystės atvejams nustatyti mobilaus ryšio bendrovėse

Straipsnyje nagrinėjamas duomenų gavybos technologijų taikymas telekomunikacinėje bendrovėje, apgavystės atvejų nustatymas ankstyvoje stadijoje, panaudojant klasterizacijos metodą. Pateikiami pagrindiniai duomenų gavybos elementai: galimybė palaikyti didelius duomenų kiekius, tinkami metodai ir algoritmai, analizuojamos srities tinkamumas bei problemos sprendimo būdai.