

# Discovering Fraud Behaviour in Call Detailed Records

Christoph Schommer  
University Luxembourg, Dept. of Computer Science and Communication  
ILIAS Laboratory, MINE Research Group  
6 Rue Coudenhove-Kalergi, L-1359 Luxembourg  
christoph.schommer@uni.lu

## Abstract

*Since years, the efficiency and potential behind Data Mining stimulates the hope of detecting hidden but useful findings in the data. The explorative concept is attractive and an excellent initial situation regarding its application in one of the today's most challenging application fields: security, trust, and reliability. The following contribution concerns with the interactive application of Data Mining in one of its most sensible scenarios: fraud. We present a real project and describe an early fraud detection that has been 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. A method on how Data Mining supports the detection of possibly fraudulent calls is presented and a simple solution of the detection request introduced.*

**Keywords:** Data Mining, Fraud Detection, Call Detailed Records (CDR).

## 1 Introduction

Since more than 2000 years, the concept of *Data Mining* is successfully applied in many industrial and academic applications that concern with behavioural analysis, profiling, and scoring. On the basis of qualitative data, *Data Mining* concerns with the explorative analysis of (masses of) data to find hidden, but useful information. With a reliable, knowledgeable, and stimulating interpretation of even this information, *Data Mining* yields in a life-cycle process, which loops it back to the collection of data through the deployment of discovered findings. Application examples are an analytical examination of evidences of thefts – which may contribute to the dissolving of criminal cases –, the detection of insider trading in stock exchanges, and the characterization of intruders in a computer network.

Major steps in the process of *Data Mining* are a certain quality assurance of data, the usage of intelligent algorithms and

concepts to explore and to understand the data, the collaboration with diverse methods to visualize the data, information, and findings, and finally, the interpretation and deployment of the findings. A corresponding domain knowledge is indispensable.

An important and challenging target application for *Data Mining* is in the area of Security. With that, a diverse number of disciplines inside security have been established in the last years, which concern e.g. with the analysis and prevention of crime, the detection of intruders inside a computer-networks, the characterization and justification of plagiarism, the detection of fraud, Insider Trading on the stock market, Data Privacy, Forensic linguistics, and many more. But although the discipline of computer science offers a wide range of intelligent software solutions, the data itself is the key of success: if “nothing” is in the data then there is “nothing” to find. A second point is that the human user with his natural creativity and intelligence regarding an inherently interpretation, thinking, and strategically planning is often concealed and that *Data Mining* is often equated with the development of software, especially machine learning or visualization algorithms. This is a misbelief, since a strategic-thinking and autonomous operating analytical software with a domain knowledge does not exist. With respect to this, the concept of *Data Mining* is less a framework of technical achievements but even more an exemplar fostering on human users using software as a matter of exploration.

## 2 Brief Literature Review

The preoccupation with fraud is an important discipline in industry and it is undoubted that *Data Mining* plays a key role in it. *Fraud Auditing* concerns with the analysis of establishments that carries risk, for example divisions or processes. It includes the control of internal affairs and the ascertainment of loss. *Fraud Prevention* refers to solutions are developed to preclude fraud and defalcation. The goal is to minimize the probability of a consequential loss. In

e-commerce, it might be necessary to have such a preventing solution in real-time in order to extirpate the fraud right from the beginning. In the field of a mobile telecommunication, the prevention of fraud is answered by *Premium Rate Services*. *Fraud Detection* encompasses the identification of risks regarding deceptive and fraudulent activities. The following paper therefore addresses exemplarily the relevance of *Data Mining* in this discipline by introducing a fraud detection scenario.

Fraud detection has been concerned in many applications. Due to credit card transaction proportions, *Brause* presents a new concept that is developed and tested on real credit card data. The work claims how advanced *Data Mining* techniques and more specifically, Neural Network algorithms, can become successfully combined to obtain a high fraud coverage combined with a low false alarm rate of existing real credit card transactions (Gesellschaft für Zahlungssystem (GZS), Eurocard/Mastercard). In [7], the usage of process mining to reduce the risk of internal fraud is discussed. [1] present a comprehensive framework that mines and detects fraudulent transactions of Card-Not-Present in the e-payment systems with a high degree of accuracy. [10] introduce an analytical method and show how it can be used in a real-world auction scenario: a working Java prototype is developed, which allows users to query the legitimacy of e-Bay users. In [21], a new method for avoiding telephone frauds using a method is proposed (*CAPTCHA = Completely Automatic Public Turing Test to Tell Computer and Human Apart*). *Alaric* is a card fraud detection and prevention system that uses a proprietary inference techniques based on Bayesian methods. *PATTERN:Detect* aims at uncovering fraud and anomalies, such as fraudulent credit card transactions or network intrusions. *Nestor* offers a risk management products, including credit card fraud detection. An interesting software library is offered with *Neural Technologies Decider*, which is a suite of solutions for the finance industry for advanced modelling and scorecard development for detecting bad debt and application fraud. *StatConsulting* is a fraud detection system based on customer behavior modeling using latest *Data Mining* methods together with traditional statistics and *Xtract Fraud Detector* uses adaptive neural nets to analyze customer behavior and detects insurance claims fraud and payment card fraud. Finally, *Wizrule* find unexpected rules in data and other applications for fraud detection.

### 3 Detection of fraud carried out by means of premium telephone numbers

The following project has been performed on authentic data for a major European telephone company. The theme of the project has been to install an automated detection of fraudulent behavior on their premium number services.

#### 3.1 Problem Description

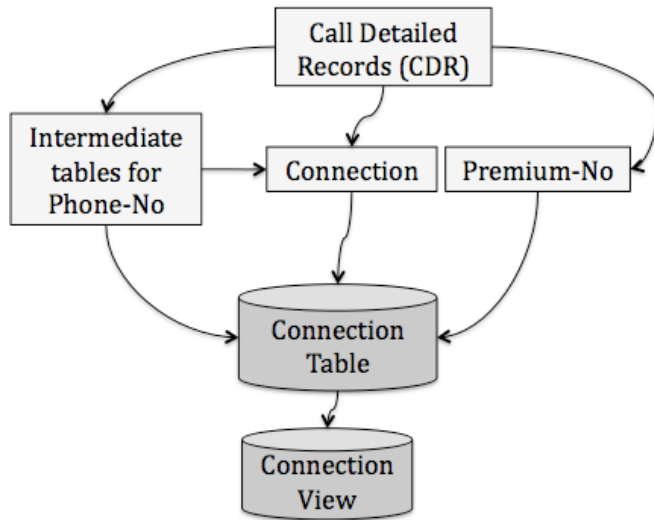
The solution described here permits the early detection of fraud carried out by means of premium telephone numbers. By premium (phone) numbers, we understand such numbers as offering services like *expert hotlines for computers or other technical equipment, advice for insurance or juridical questions, Stock market tips, phone sex*, etc. The prices that are charged when somebody calls a premium number are normally higher than “ordinary” phone calls, ranging from 0,49 Cents per minute to two Euro (or more). 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 dialing 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 caller (or the conspiratorial partner), the company becomes aware that the caller 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 behavior, 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.

#### 3.2 The Data Situation

The (real) phone company records for each phone call are called *Call Detailed Record (CDR)*. The Call Detailed Records are stored in a database table Call Detail Record that consists of more than 50 attributes and that is being updated on a daily basis.

With respect to fraud detection, we preprocess the raw data and create a data model on the basis of the Call Detail Records only (Figure 1). Most important attributes are *CALLED-ID* (identification of the caller), *PREMIUM-ID* (identification of premium number), *START Date* (date, when the phone call has started), *START Time* (time, when the phone call has started), *Duration* (duration in seconds), and *Charges* (charges, in Euro).



**Figure 1. The connection view as a subset of the connection table, which is made of the intermediate tables for phone-no and the premium-no table.**

CALLER_ID	PREMIUM_ID	Start Date	Start Time	Duration (secs)	Charges
2561501233	0815691381	2000-01-09	22.40.12	123	4.01
2538366458	0815656545	2000-01-09	10.42.36	37	1.06
2561501233	0815691281	2000-01-09	22.43.00	138	2.21
7857107555	0815223223	2000-01-09	22.28.49	40	0.97

**Figure 2. Example of a Call Detailed Records (CDR) database table.**

In a series of experiments we find 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 fraudulent premium number. From the Call Detail Record attributes, a weekly connection view containing aggregated data must be generated that contains 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 is calculated and derived automatically from the Call Detail Records. It contains the weekly update for every connection:

- *SUM-DUR*: Whole duration of all calls on a specific connection, from a specific caller to the premium number.
- *NO-CALLS*: Number of all calls on the connection.

- *REL-DUR*: Indication whether the connection has an extraordinarily high share in the turnover generated by all connections with the same premium number. Defined by the relationship between the whole duration of all calls on the specific connection and the average whole duration of all different connections 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.

as well as *CALLER-ID* and *PREMIUM-ID*. As an important feature, Call Detail Records are generated automatically and do not contain any missing values. It is therefore useful to take advantage of statistical functions like univariate or bivariate statistics for a better understanding of all attributes. However, the hope that these attributes also may contain further valuable information indicating potential fraudulent calls has not been come true.

### 3.3 Demographic Clustering

The Demographic clustering learning algorithm discovers the number of clusters automatically on a given user-defined similarity threshold, which typically lies in the range of [0, 1]. A value of 1 refers to a complete identity and 0 to a complete difference. Two telephone callers are similar enough, then they are candidates for being put into the same cluster. Here, the similarity between two callers is calculated by comparing each caller's attribute and giving a score for how closely the attributes match. The scores are then summed and divided by the number of attributes that are compared. If all the attributes are of a categorical data type, then this is quite simple as two identical values contribute with a score of 1.0 (if they are different, they contribute a score of 0, respectively). So, if a caller is described by  $k$  categorical attributes and if we compare two callers with a match of  $\frac{k}{2}$  values, then the two caller share a similarity of 50% and are grouped together (in case that the similarity threshold is lower, of course). For numerical attributes, the concept of being similar is slightly different. If the values are identical, then they contribute a score of 1.0, but if the values differ, then they get a score based on the degree of difference. Typically, if the difference is expressed in terms of the number of standard deviation of the attribute for all callers and the score is calculated such that if the two values are 0.5 of a standard deviation apart, the score is

### Database Table

	SUM-DUR	NO-CALL	SUM-COSTS	MAX-DUR
A	100-200	<50	<1000	<30
B	0-50	<50	>=1000	30-40
C	50-100	>=50	<1000	<30
D	>1000	<50	<1000	>50
E	100-200	>=50	<1000	<30
F	0-50	<50	>=1000	>40

### Clustering Algorithm

	A	B	C	D	E	F
A	4	1	2	2	3	1
B	1	4	0	1	0	3
C	2	0	4	1	3	0
D	2	1	1	4	1	2
E	3	0	3	1	4	0
F	1	3	0	2	0	4

	A	C	E	B	F	D
A	4	2	3	1	1	2
C	2	4	3	0	0	1
E	3	3	4	0	0	1
B	1	0	0	4	3	1
F	1	0	0	3	4	2
D	2	1	1	1	2	4

### Result: 3 typical clusters:

	SUM-DUR	NO-CALL	SUM-COSTS	MAX-DUR
A	100-200	<50	<1000	<30
C	50-100	>=50	<1000	<30
E	100-200	>=50	<1000	<30

	SUM-DUR	NO-CALL	SUM-COSTS	MAX-DUR
B	0-50	<50	>=1000	30-40
F	0-50	<50	>=1000	>40

	SUM-DUR	NO-CALL	SUM-COSTS	MAX-DUR
D	>1000	<50	<1000	>50

**Figure 3. Demographic Clustering:** Each data record is compared with each other yielding in a similarity matrix. The generation of clusters is then performed on the basis of maximizing the Condorcet, which is the sum of all similarity values inside each cluster divided by its best match (here 4), respectively. With that, this clustering owns a Condorcet of  $\frac{1}{3}(\frac{28}{36} + \frac{14}{16} + \frac{4}{4}) = 0.88$ .

0.5. Although the standard deviation is normally used, other measures can be defined. If the similarity threshold is 0.5, then callers can potentially be grouped together when the sum of the scores for each attribute divided by the number of attributes is greater than 50%.

An important aspect is that the clustering algorithm has to consider that even if a caller has an acceptable similarity to an existing group of callers, this does not automatically mean that they will be put into the same cluster. This clustering technique finds optimum combinations of callers that maximize their similarity within each cluster, while at the same time maximizing the dissimilarity between different clusters. To decide that this it tries to maximize the value of a statistic that it calculates called the *Condorcet* value.

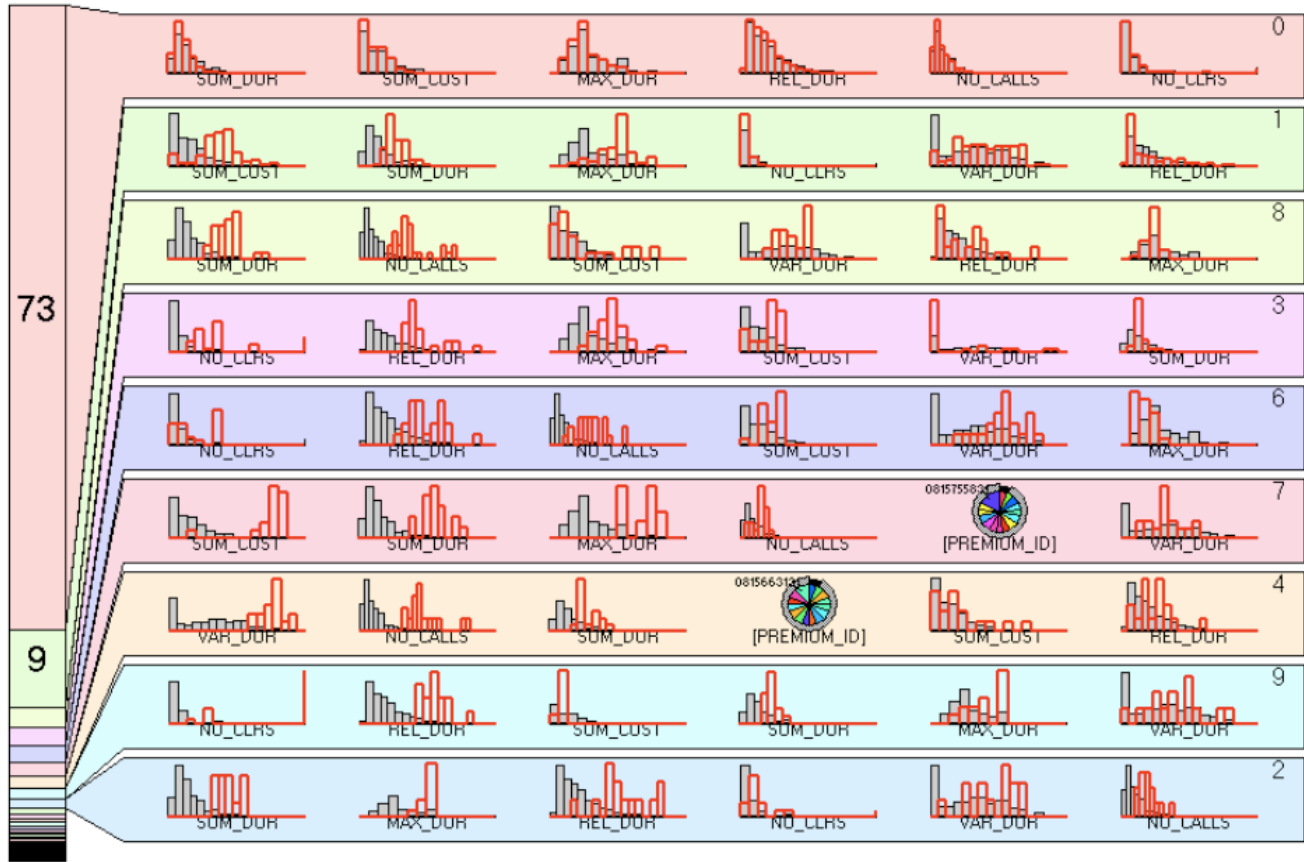
If we give the clustering algorithm a similarity threshold and do not limit the number of clusters that it can produce, it keeps on trying to find the minimum number of clusters that satisfy the similarity threshold since this will also maximize the condorcet value. If we do not know what similarity threshold to select, an internal mechanism can be used to discover what the optimum value may be. The *Condorcet* value is a value in  $[0, 1]$  that is a measure of how similar a caller is to other callers within the cluster, and how dissimilar they are to callers in other clusters. It has a value of 1 when all callers in a cluster are identical and where there are

no callers outside the cluster that have the same characteristic. A value of 0 indicates that the callers are distributed randomly among the clusters. The condorcet value can be calculated for all the caller attributes or for each attribute separately.

### 3.4 Findings

Usually and unless the number of clusters is restricted to a certain number, a clustering algorithm returns many clusters of different characteristics. Regarding the detection of fraudulent behavior, the unlimited concern is of importance since a possible fraud may be given rather in small, tight clusters, comprising exactly the unusual connections (Figure 4). We therefore have come up with about 50 clusters and have ordered them descending by size. For each cluster, the value distribution of the derived attributes (which are used for the clustering) is shown as well as the two additional attributes *Caller-Id* and *Premium-Id* (where are not used to compute the clusters). Several of these given clusters describe groups of callers with unusual call behavior. Experience shows, that mostly smaller clusters contain the connections with a high fraud probability.

Performing clustering for the created data mart, we have generated a segmentation of all connections described in



**Figure 4. Clustering Result with about 50 clusters, which are ordered depending on the number of data records that are in (= the size, in percent). Regarding the numerical distribution (histograms), the grey bars correspond to the distribution of all data whereas the red bars corresponds to the data inside the cluster. With respect to the categorical distributions (circular chart), the outer ring corresponds to all data, the inner ring to the data inside the cluster. All variables inside each cluster are ordered according to the  $\chi$ -square test: the more diverse the distribution of the variable inside a cluster versus the overall population is, the higher the appearance in this cluster.**

Figure 4. Looking closely through all the clusters we find clusters describing normal connections as well as sundry segmentations indicating fraudulent calls. Usually, after having clustered the connection data, the largest cluster (Figure 5) comprises the normal connections. This is because the value distribution of the different features within the cluster (inner ring for categorical attributes, red bars for numerical attributes) is very similar to their distribution in the whole population (outer ring for categorical attributes, grey bars for numerical attributes). This cluster comprises about 73% of all connections which is indicated at the bottom of the visualized result.

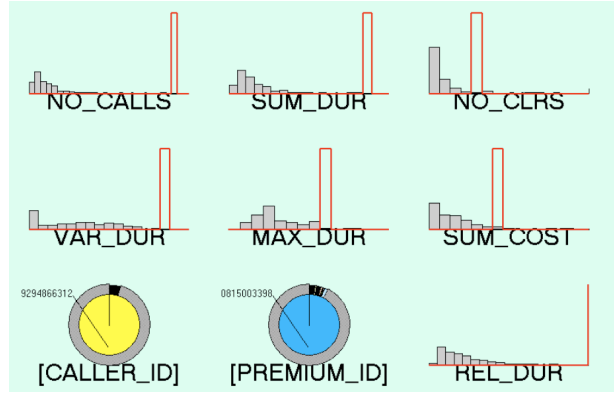
If a set of premium numbers are called again and again by the same caller, this may indicate a case of conspiracy. This cluster comprises three one-to-one connections, which

means that each of the three premium numbers is called by one, and only one, caller. If we display the details for the attribute *NO-CLRS* (number of callers) we observe that the minimum and maximum value is exactly 1 (Figure 5). The costs for each of these connections (*SUM-COST*), the whole duration of all calls (*SUM-DUR*), and the number of calls (*NO-CALLS*) are comparably very high. These are strong indicators, that this cluster depicts connections based on conspiracy between the caller and the provider of the premium number service with a high probability.

We also have found a cluster (Figure 6), where one caller calls a premium number excessively: the duration and the costs of these calls are unusually high, so we probably have identified an “addicted caller”. Normally, many callers of this kind are not able or not willingly to pay their services.



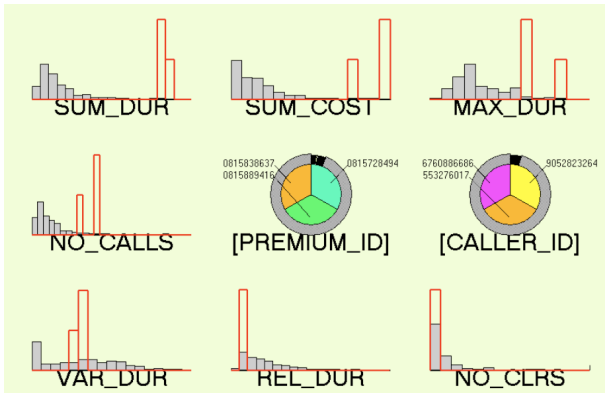
**Figure 5. Cluster depicting normal connections (Cluster 0, 72.96% of all telephone calls).**



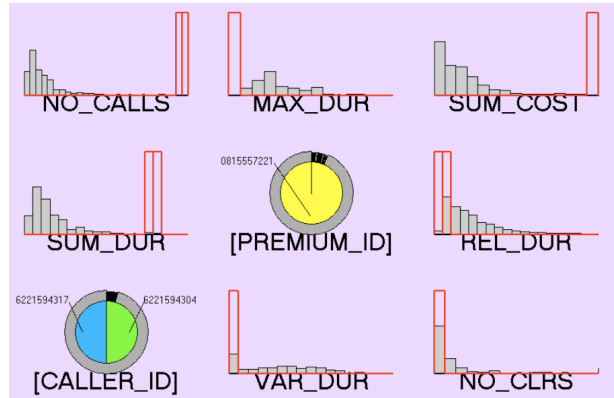
**Figure 7. Cluster depicting connections indicating phone addiction (Cluster 34, 0.09% of all telephone calls).**

So, special payment conditions for these people could prevent loss. Cluster 34 contains only one connection, where the premium number has a lot of callers. For the number of callers (*NO-CLRS*), we observe the same high minimum and maximum value, which is 217.

single phone calls (*MAX-DUR*) is very low, the length of the calls do not vary at all (the details for the attribute *VAR-DUR* show the minimum and maximum value 0). These are also strong indicators, that the cluster contains connections indicating that a fraudulent conspiracy between caller and premium number service provider is highly probable.



**Figure 6. Cluster depicting connections indicating conspiracy (Cluster 26, 0.27% of all telephone calls).**



**Figure 8. Cluster depicting connections indicating conspiracy using automated devices (Cluster 28, 0.18% of all telephone calls).**

Concerning the usage of automated dialing devices, we have found a cluster as presented in Figure 7. The cluster contains exactly two connections with two callers and the same premium number. If we display the details for the attribute the number of callers (*NO-CLRS*), we observe that a minimum and maximum value is exactly 2, which refers to the situation there no other callers of this premium number exist. Both callers call the premium number excessively (*NO-CALLS*) and the costs (*SUM-COST*) and the whole duration of all calls (*SUM-DUR*) are extraordinarily high, respectively. Whereas the maximum duration of the

In contrast to the cluster 26, the calling of the premium number is performed by the aid of a computer or an automatic dialing device, because the duration of the phone calls do not vary at all. Figure 8 illustrates the inspection of details of a cluster, namely phone calls belonging to cluster 26.

## 4 Conclusions

The potential behind an explorative analysis with *Data Mining* for security applications has been proved on a real project dealing with the detection of fraud. Preprocessing the raw data to a *Data Mart* (= Connection Table) has been the fundament for any further explorations that finally has led to the discovery and characterization of phone connections indicating phone addiction, conspiracy, and conspiracy using automated devices.

The operative use of this initial solution concept has been implemented several weeks after the presentation of the analytical results. Since this time, the telephone company has used these observations/findings permanently and with great success in several locations. Week by week, about 5 million new Call Detail Records have been analyzed, and fraud attempts within the scale of tens of thousands of Euro have been detected and prevented. The return on investment has been achieved already after six months (2001).

## Acknowledgement

This project has been performed in 2001 and is documented in the book "Mining your own Business. Vol. 3 Telecommunications" [2]. I am very thankful to my former IBM colleagues Christian M. Andersen, Corinne Baragoin, Graham Bent, Jieun Lee, and Stephan Bayerl for the warm, harmonious, and fruitful time at the IBM Almaden Research Center.

## References

- [1] A. M. Al-Khatib, E. Hattab. Mining Fraudulent Transactions in e-payment Systems. iiWAS 2007. (2006).
- [2] C. M. Andersen, S. Bayerl, G. Bent, J. Lee, C. Schommer: Mining your own Business. IBM Redbook Series. Vol. 3 Telecommunications. IBM Press. (2001).
- [3] V. Arnold. Advances in Accounting Behavioral Research. Emerald Group Publishing. (2008).
- [4] F. Bonchi, F. Giannotti, G. Mainetto, D. Pedreschi. Using *Data Mining* Techniques in Fiscal Fraud Detection. DaWaK. (1999).
- [5] R. Brause, T. Langsdorf, M. Hepp. Neural *Data Mining* for Credit Card Fraud Detection. ICTAI. (1999).
- [6] P. Ferreira, R. Alves, O. Belo, L. Cortesao. Establishing Fraud Detection Patterns Based on Signatures. Industrial Conference on *Data Mining* (2006).
- [7] M. Jans, N. Lybaert, K. Vanhoof. Internal Fraud Risk Reduction - Results of a *Data Mining* Case Study. ICEIS 2008.161-166
- [8] E. Kirkos, C. Spathis, Y. Manolopoulos. *Data Mining* techniques for the detection of fraudulent financial statements. Expert Syst. Appl. (ESWA) 32(4). (2007).
- [9] H. Koesmarno, W. Graco. Targeted Fraud Detection Use of a Taxonic Method, Subspace Clustering and Knowledge Acquisition to Develop a Classification Model Ensemble. Industrial Conference on *Data Mining* - Industry Proceedings. (2009).
- [10] Y. Ku, Y. Chen, C. Chiu. A Proposed *Data Mining* Approach for Internet Auction Fraud Detection. PAISI (2007).
- [11] M. Lek, B. Anandarajah, N. Cerpa, R. Jamieson. *Data Mining* Prototype for Detecting E-Commerce Fraud.
- [12] P. A. Ortega, C. J. Figueroa, G. A. Ruz. A Medical Claim Fraud/Abuse Detection System based on *Data Mining*. A Case Study in Chile. DMIN (2006).
- [13] F. S. Park, C. Gangakhedkar, P. Traynor. Leveraging Cellular Infrastructure to Improve Fraud Prevention. ACSAC 2009.
- [14] S. Rping, N. Punko, B. Gnter, H. Grosskreutz. Procurement Fraud Discovery using Similarity Measure Learning. Industrial Conference on *Data Mining* - Posters and Workshops 2008.
- [15] B. P. Veldkamp, T. de Vries. Identification of Bankruptcy Fraud in Dutch Organizations. IADIS European Conf. *Data Mining* 2008.
- [16] S. Viaene, D. Van Gheel, Mercedes Ayuso, Montserrat Guillen. Cost-Sensitive Design of Claim Fraud Screens. Industrial Conference on *Data Mining* (2004).
- [17] M. Weatherford. IEEE Intelligent Systems. Intelligencer - Mining for Fraud. IEEE Distributed Systems Online (DSONLINE) 3(7) (2002).
- [18] C. Westphal. *Data Mining* for Intelligence, Fraud & Criminal Detection. Advanced Analytics & Information Sharing Technologies. Auerbach Publications, 2008.
- [19] J. Xu, A. H. Sung, Q. Liu, S. Mukkamala. Fraud Detection System Based on Behavior Mining and Anomaly Detection. IICAI (2005).
- [20] W. Yang, S. Hwang. A process-mining framework for the detection of healthcare fraud and abuse. Expert Syst. Appl. (ESWA) 31(1).56-68 (2006).
- [21] H. Zhang, X. Wen, P. He, W. Zheng. Dealing with Telephone Fraud Using CAPTCHA. ACIS-ICIS 2009.