A Model for Rule Based Fraud Detection in Telecommunications

1. Smt.S.Rajani, research scholar, S.VUniversity, Tirupati.

2. Prof.M. Padmavathamma, Head, Dept of Computer Science & Applications, SVUCCMIS, S.VUniversity, Tirupati.

Abstract:

Telecommunications fraud is a worldwide problem that deprives operators of enormous sums of money every year.. Fraud detection is an increasingly important and difficult task in today's technological environment. Several data mining applications are described and together they demonstrate that data mining can be used to identify telecommunication fraud, improve marketing effectiveness, and identify network faults. In this paper we propose a rule based for fraud detection in telecommunication system.

Key words: Telecommunications, fraud detection, rule based system

1. Introduction:

The Telecommunications industry generates and stores a tremendous amount of data .The amount of data is so great that manual analysis of the data is difficult, if not impossible. The need to handle such large volumes of data led to the development of knowledge-based expert systems. These automated systems performed important functions such as identifying fraudulent phone calls and identifying network faults. The problem with this approach is that it is time consuming to obtain the knowledge from human experts and, in many cases; the experts do not have the requisite knowledge. As consumers are putting more of their personal information online and transacting much more business over computers, Telecommunications fraud (*non revenue fraud*) includes:

- To avoid or reduce payment of services used

the potential for losses from fraud is in the billions of dollars, not to mention the damage done by identity theft.

. This paper focuses on the problem of finding fraudulent customers using rule based systems, and gives the specific method to forecast the behaviour of malicious arrearage.

1.2 Objectives

This paper has the following main objectives:

- Expose the fraud problem within the scope of telecommunications, enumerating the main causes for telecommunication fraud and the impact on the operators;
- Analyze the actual fraud detection solution, explaining the methods used by the fraud solution to detect fraud and analyzing how the solution can evolve;
- Define a new approach in order to evolve the actual fraud detection solution, defining new methods;
- Propose a solution model, which supports the methods previously defined.

2. The Nature of Fraud:

The most difficult aspect of fighting fraud is identifying it. In the context of telecommunications, a fraudulent phone call is one in which there is no intent to pay—theft of service. The main motivation to commit fraud is to make money (*revenue fraud*). This can be achieved by selling fraudulently obtained services at cheap rates or by selling critical company information to other criminals. Other reasons to commit

- To maintain anonymity while committing other crimes
- To demonstrate ability to outmanoeuvre the operator's system security.

Our goal was to create a fraud management system that was powerful enough to handle the many different types of fraud that we encountered and flexible enough to potentially apply to things we had not seen yet. We next provide examples of some common varieties of fraud in the telecommunications world.

2.1. Subscription fraud.

Subscription fraud happens when someone signs up for service (e.g., a new phone, extra lines) with no intent to pay. In this case, all calls associated with the given fraudulent line are fraudulent but are consistent with the profile of the user.

2.2. Intrusion fraud.

This occurs when an existing, otherwise legitimate account, typically a business, is compromised in some way by an intruder, who subsequently makes or sells calls on this account. In contrast to subscription calls, the legitimate calls may be interspersed with fraudulent calls, calling for an anomaly detection algorithm.

2.3. Fraud based on loopholes in technology.

Consider voice mail systems as an example. Voice mail can be configured in such a way that calls can be made out of the voice mail system (e.g., to return a call after listening to a message), as a convenience for the user. However, if inadequate passwords are used to secure the mailboxes, it creates vulnerability

2.4. Social engineering.

Instead of exploiting technological loopholes, social engineering exploits human interaction with the system. In this case the fraudster pretends to be someone he or she is not, such as the account holder, or a phone repair person, to access a customer's account.

2.5. Fraud based on new technology.

Using new technology, such as Voice Over IP fraudsters realized that they could purchase the service at a low price and then resell it illegally at a higher price to consumers who were unaware of the new service, unable to get it themselves, or technologically unsophisticated. Detecting this requires monitoring and correlating telephony usage, IP traffic and ordering systems.

2.6. Fraud based on new regulation.

In 1996, the FCC modified payphone compensation rules, requiring payphone operators to be compensated by the telecommunication providers. This spawned a new type of fraud—payphone owners or their associates placing spurious calls from payphones to toll-free numbers simply to bring in compensation income from the carriers.

2.7. Masquerading as another user.

Credit card numbers can be stolen by various means (e.g., "shoulder surfing"— looking over someone's shoulder at a bank of payphones, say) and used to place calls masquerading as the cardholder.

3. Data mining for fraud detection

Call detail records are generated in real time and therefore will be available almost immediately for data mining. This can be contrasted with billing data, which is typically made available only once per month. Call detail records are not used directly for data mining, since the goal of data mining applications is to extract knowledge at the customer level, not at the level of individual phone calls. Thus, the call detail records associated with a customer must be summarized into a single record that describes the customer's calling behaviour. The choice of summary variables is critical in order to obtain a useful description of the customer. Below is a list of features that one might use when generating a summary description of a customer based on the calls they originate and receive over some time period P:

1. Average call duration
2. Percentage of no-answer calls
3. Percentage of calls to/from a different area code
4. Percentage of weekday calls (Monday – Friday)
5. Percentage of daytime calls (9am – 5pm)
6. Average number of calls received per day
7. Average number of calls originated per day
8. Number of unique area codes called during P

These eight features can be used to build a customer profile. Such a profile has many potential applications

4. Requirement of knowledge based expert system in telecom:

Issues to look at when designing our Fraud Management System are:

1. The collection and the format of the input data
2. The identification of fraud indicators
3. The fraud detection technique

4.1. The collection and the format of the input data

Collecting data for analysis is the first step in the fraud detection process. Typically, CDRs (call detail records) generated by network elements such as telephone switches for billing purposes, are the main source of input data for current FRAUD MANAGEMENT SYSTEM

4.2. The identification of fraud indicators

Fraud indicators are details about the service usage that may indicate that fraud is perpetrated. In the traditional voice networks usual indicators of

fraud include long duration calls, large number of calls from the same account and calls to blacklisted numbers. These indicators are used to create fraud rules or signatures that are characteristics of a fraud type. Fraud rules need to be updated continuously as fraud types evolve. An alternative to defining fraud signatures is the creation of customer profiles. A customer profile defines the individual pattern of normal usage for a customer. By comparing the current usage to the stored profile, fraud can be detected without the need for specifying rules for specific fraud scenarios. Our suggestion is therefore to create a *service* profile that describes how the service is normally used by the average user. This service profile is also used to create service specific fraud rules used to detect suspicious events. The profile will answer questions such as: how much is usually spent on this service, at what time and for how long is the service usually used? Answers to these questions enable the creation of groups of users for a specific service. For instance, it is possible to create different profiles for different times of the day or week (e.g. peak time, night, and week-end). The billing records are then sent to the relevant group profile based on the time of the service usage. The service profiles are stored in modules that can be added and removed from the Fraud Management System as needed.

4.3 The fraud detection technique

Various data analysis techniques are in use by Fraud Management Systems. The most recurrent techniques are **threshold-based, rules-based and the use of neural networks**. In threshold-based fraud analysis, details about the call (e.g. call duration) are compared to fixed criteria called triggers. If the value of the call detail exceeds that trigger, an alarm is generated. Threshold based detection tools are simple, efficient but only work well for detecting the extremes of fraudulent events as triggers are usually set to high values. In rules-based analysis, fraud patterns are defined as rules and call records are analysed against these rules to spot fraud. Call detail records include sufficient information to describe the important characteristics of each call. At a minimum, each call detail record will include the originating and terminating phone numbers, the date and time of**.**

5. Rule Based Fraud Detection**:**

Many commercial fraud analysis applications based on rules. In a rule based fraud detection system, fraud patterns are defined as rules. Rules may consist of one or more conditions. When all conditions are met, an alert is raised. Three **types** of data may participate in rule conditions**: call details, customer details and behaviour monitors**. Behaviour monitors are summations of number, duration or rated value of calls over a certain time window (e.g., the daily number of calls to mobile phones at off-peak hours). Any population of calls can be monitored. For identifying superimposed fraud, "normalized" monitors can be used. These monitors denote the measured value in terms of standard deviations from the average value. High value of such monitor indicates an extreme increase in usage, and can be used in a **superimposed fraud.**

In the fraud analysis context, the generated rules will be used as alarm-setters for suspected fraud. Therefore, we would like to generate rules that are appropriate for this task, rather than for standard machine learning tasks such as classification or scoring.

In rule-based fraud management systems, the alarms (or alerts) are usually not treated individually but rather combined at the customer level into "cases" of suspected fraud. Thus, K alerts generated for the same customer result in only one case being created, while K alerts generated for K different customers, result in K different cases being created. If we just count the number of true alarms (i.e., alerts that are actually fraudulent) and false alarms, the two situations would be identical. Thus, it is generally true that accuracy should be computed at the customer (case) level - the "higher" of the two levels mentioned above. The success of a fraud rule is determined by how many really fraudulent cases were identified and how many cases were false alarms.

Example for rules may be:
Credit-rating=C AND daily international calls duration> 2hrs => alert
Deposit= X AND normalized-daily-duration standard deviations >4 = > alert

The alerts are gathered into cases (a case for each account) together with account data and Call Detail Records. The cases are the starting point of the manual investigation process, where a human analyst determines for each case whether it is actually fraudulent or not. Within a rule based system the performance of each individual rule is secondary in importance. The main issue is, of course, the performance of the rule-set selected for use in the system. Our ultimate goal in the rule-discovery process should be to select a rule-set**.**

5.1 Frame work for Rule based model

We must first tackle the problem of "where the patterns live". There are at least two separate levels of data, and sometimes more. **One level is the customer data**, Examples of such attributes are customer's age, ethnicity and family status, price plan and telephone model. **The second level is what we have termed "behavior'‘-level data**. This term refers to usage characteristics in a short time frame (typically a single day). Typical behavior-level attributes are the number of international calls in a day and total duration of all calls in a day. They may also include "normalized" behaviour monitors detecting changes in behaviour relative to the history of usage by this particular customer. **Our goal is to find patterns combining elements from both levels,** giving rules such as the following: "People who have a particular price plan that makes international calls expensive and who display a sharp rise in international calls are likely the victims of customer longetevity fraud".

There are several possible approaches to constructing correct bi-level rules. One is based on standard rule-generation procedures completely in favour of simpler ad-hoc methods. For example, we could use a standard procedure to build rules on customer attributes only, using amount balances with one record per customer, and then run a separate second stage with one record per "behaviour sample" to add behaviour attributes to the rules. This naive approach is unlikely to give good results, as it would be limited in its ability to find "interactions" between customer-level and behaviour-level attributes (e.g., that customers in a certain area are likely to be fraudulent if they make many international calls).

Another approach is to modify the existing algorithms to ensure that they count the records correctly, taking into account the issue of bi-level data. We have taken this approach, and have built a rule generator based on a modification of the **C4.5 algorithm**.

The relevant changes in the **c4.5** algorithm are concentrated in three areas- **splitting criterion and stopping rule for tree construction and pruning significance tests**. The splitting criterion is used to select the "best" greedy split in each stage during tree construction. It is based on calculating the "information content" of each of the suggested splits with regard to the class distribution and choosing the one with the highest content. The stopping rule dictates the size of groups we are willing to accept as "leaves" in the tree. The goal of using a stopping rule is to prevent the system from creating rules representing small samples with no statistical generalization ability. For both of these areas the key to working on bi-level data is that the "size of groups" concept has to be defined with respect to the level at which the attribute being split belongs. So, when splitting on a customer-level attribute, the amount of customers of each class found in each "leaf" is counted. When splitting on a behaviour-level attribute we should count the number of instances of behaviour (i.e. the number of "records") of each class in each "leaf".

5.2 Categories of Subscribers

In order to generate a database of known fraudulent/legitimate cases, it was necessary to formalize the definition of subscribers' categories. Consequently the following four categories of subscribers were defined:

- **Subscription fraudulent**. Most of the users in this category do not pay their bills at all, but if they do, the debt/payment ratio is very high. The line is typically blocked due to suspicious behaviour in long distance calls within 6 months after the installation date.
- **Otherwise fraudulent**. Subscribers for more than a year who present a sudden change in their calling behaviour, generating an abnormal rise in their newer billing accounts.
- **Insolvent**. Subscribers with a total debt of less than 10 times their monthly payments, having two or more unpaid bills. This category includes new customers that have never paid their bills but whose monthly expenditures are similar to average residential lines.
- **Normal**. Customers with their bills up to date or at most a single unpaid bill for less than 30 days after the due date

First, 700 cases were drawn from the repository and classified manually into the four categories described above. The manual classification procedure was assisted by an expert with many years of experience in telecommunication fraud
management. This was a time-consuming procedure since for each case, all the information available in the repository had to be examined on the computer screen.

The classification module was designed with a hierarchical tree structure, including three layers and five nodes, as shown
in Fig. 1.

- The first layer consists of the root node, which discriminates between fraudulent and normal subscribers, but assigns the

insolvent subscribers to any of the two groups.

- The second layer has two nodes. Node N/I discriminate between normal and insolvent cases. Node F/I discriminate between fraudulent and insolvent cases.
- The third layer has two nodes that discriminate among subscription fraudulent, otherwise fraudulent and insolvent cases. Node I/O distinguishes between insolvent and otherwise fraudulent. Node S/O discriminates between subscription fraudulent and otherwise fraudulent.

The data set of 700 cases was used to select the variables of the classification module, as well as to design fuzzy rules to discriminate among the categories. As an example, for continuous variables, three Gaussian-like fuzzy membership functions were defined to measure low-risk (LR), medium-risk (MR) and high-risk (HR) of subscription fraud. A total of 54 fuzzy rules were defined for the classification module, using 17 variables. Here we present some examples. At the root node of the tree-classifier shown in Fig. 1, the first three rules generated were:

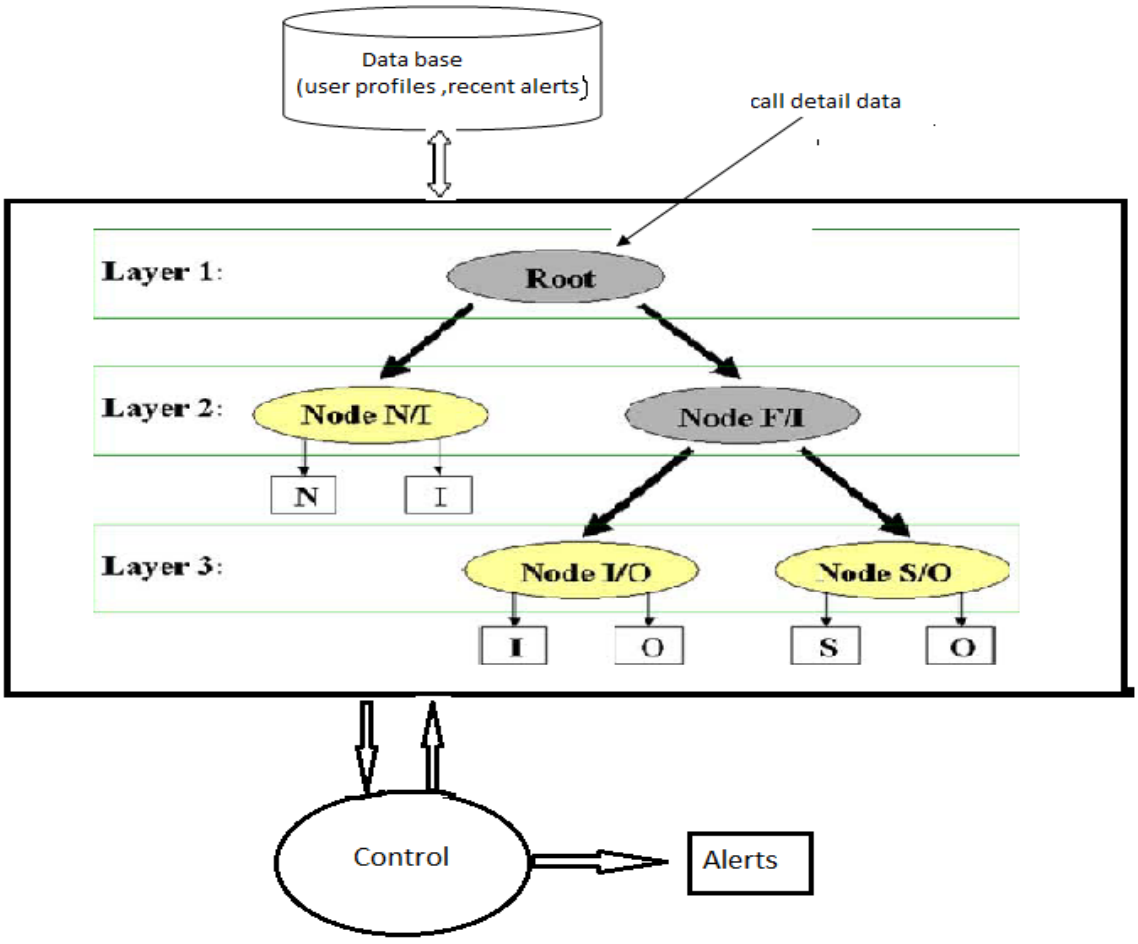

Fig.1

Rule 1: IF (cutomer longetivity is LR) AND (elapsed time between installation and blocking data is LR) AND (debt/payment ratio is LR) AND (phone blocked flag is LR) AND (amount balance is LR) THEN (Output_RootNode is Node N/I).

Rule 2: IF (cutomer longetivity is HR) AND (elapsed time between installation and blocking data is HR) AND (debt/payment ratio is HR) AND (phone blocked flag is HR) AND (amount balance is HR) THEN (Output_RootNode is Node F/I).

Rule 3: IF (amount balance is LR) AND (number of days with unpaid bills is LR) THEN (Output_RootNode is Node N/I).

The first three rules generated at the F(Fraudulent)/I(Insolvent) node were:

Rule 4: IF (maximum debt with carriers is HR) AND (elapsed time between installation and blocking data is HR) AND (debt/payment ratio is HR) AND (phone blocked flag is HR) AND (amount balance is HR) THEN (Output F/I Node is Node S/O).

Rule 5: IF (maximum debt with carriers is MR) AND (debt/payment ratio is MR) AND (amount balance is MR) THEN (Output F/I Node is Node I/O).

Rule 6: IF (call forwarding traffic is HR) THEN (Output F/I Node is Node S/O).

The proposed model contains a data base server which collects bi-level data and a control unit to generate alarms for fraudulent cases.

Conclusion:

Most of today's fraud detection tools are either rule-based or at least comprise a rule-based detection component. The proposed model allows detecting the definite frauds with a low rate of false alarms. Moreover, this rule-based model can easily provide reasons for an alarm being raised. The rule-based tool uses the profiling strategy described above and features similar to those of the supervised neural network. The rules for the triggering of an alarm are designed manually by an expert.. In the case of rule discovery for fraud, we believe that understanding the unique features and identifying the points at which the standard tools were falling short were the key steps to suggesting a successful alternative approach.

References

Data Mining in the Telecommunications Gary M. Weiss

A data mining framework for detecting subscription fraud in telecommunication, Hamid Farvaresha and Mohammad Mehdi Sepehri

Cerebrus Solutions. (November 2002). Fraud Primer. Issue 2.3. Available: http://cerebrussolutions.com/pdf/Fraud_Primer-Nov02.pdf

O. Brad. Cyber Crime: How Technology Makes It Easy and What to Do About It. Information Systems Security, vol. 9, issue 6, pp.45-51, Jan/Feb2001 CFCA. (March 2003).

"Communications Fraud Control Association (CFCA) announces results of worldwide telecom fraud survey". Available: http://cfca.org/pressrelease/FraudLoss%20%20press %20release%203-03.doc.

Breiman, L., J. H. Friedman, R. A. Olshen and C. J. Stone (1984). Classification and Regression Trees. Chapman Hall.

Burge, P. and J. Shawe-Taylor (1997). Detecting Cellular Fraud Using Adaptive Prototypes. Proceedings of AAAI-97 Workshop on AI Approaches to Fraud Detection and Risk Management, Providence, RI, 9- 13.

Fawcett, T. and F. Provost (1997). Adaptive Fraud Detection. Data Mining and Knowledge Discovery. U. Fayyad, H. Mannila and G. Piatetsky-Shapiro (Eds.), Kluwer Academic Publishers, Boston, CA. voll,291-316.

Kokkinaki, A. I. (1997). On Atypical Database Transactions: Identification of Probable Fraud using Machine Learning for User Profiling. Proceedings of IEEE Knowledge and Data Engineering Exchange Workshop, 107-l 13.

Hoath, P. (1998). Telecoms fraud, the gory details. *Computer Fraud & Security*, 1998(1),10-14.

Hong, S. J., & Weiss, S. M. (2001). Advances in predictive models for data mining. *Pattern Recognition Letters*, 22, 55-61.

Kou, Y., Lu, C.T., Sirwongwattana S., & Huang Y.P. (2004). Survey of fraud detection techniques. *Proceedings of the IEEE International Conference on Networking, Sensing and Control* (pp. 749-754). Taipei, Taiwan.

Liao, S. H. (2005). Expert systems methodologies and applications - a decade review from 1995 to 2004. *Expert Systems with Applications*, 28, 93-103

Shawe-Taylor, J., Howker, K., & Burge, P. (1999). Detection of fraud in mobile telecommunications. *Information Security Technical Report*, 4(1), 16-28.