BI-LEVEL CLUSTERING IN TELECOMMUNICATION FRAUD

Luis Pedro Mendes[1], Joana Dias[1] and Pedro Godinho[2]

[1]*Faculty of Economics of the University of Coimbra and INESC Coimbra, Coimbra, Portugal*

[2]*Faculty of Economics of the University of Coimbra and GEMF Coimbra, Coimbra, Portugal*

Keywords: Telecommunication fraud, Data-mining, Clustering, Categorical data.

Abstract: In this paper we describe a fraud detection clustering algorithm applied to the telecom industry. This is an ongoing work that is being developed in collaboration with a leading telecom operator. The choice of clustering algorithms is justified by the need of identifying clients' abnormal behaviors through the analysis of huge amounts of data. We propose a novel bi-level clustering methodology, where the first level is concerned with the clustering of transactional data and the second level gathers data from the first phase, along with other information, to build high-level clusters.

1 INTRODUCTION

Telecommunication industry processes a very substantial amount of data per unit of time. Data is of type transactional as it refers to the interaction between clients and an operator. For an operator, it is infeasible to verify the goodness of each transaction exclusively with human resources. Fraud in telecommunication industry should be addressed effectively in order to reduce costs of illegitimate usage of the network. Computer automation fed by intelligent algorithms is the only viable solution to a problem of this scale. Several methods have been employed to track suspicious behavior of clients, to classify them or to analyze how they relate to each other.

This paper presents ongoing research on a fraud detection system undertaken in a joint agreement with a leading national network operator. Work is being carried out with real data provided by the operator. Data was received from a major national telecom provider that consists of some database tables, including only masked and truncated data in order to ensure the protection of personal data and confidential information. Therefore, the data provided by the telecom operator and used in this paper do not involve the disclosure of any personal data related to the telecom company subscribers or confidential information, ensuring full compliance with the applicable data protection legal framework. A prototype developed in the first stages of research make use of a relatively small sample of data. For later stages, a great amount of data will be made available by the operator. Besides the aim for effectiveness, developed algorithms must take into consideration the scale factor and performance efficiency.

In the first section, an overall overview of the problem of fraud in the telecom industry is presented. Several methodologies to combat this problem are referred in section two. In section three, we describe the general structure of our method for detecting fraud. Concluding remarks end the document.

2 FRAUD IN TELECOMS

Fraud can be defined[1] as a deliberate deception, trickery, or cheating intended to gain an advantage. In the telecom industry, fraud constitutes itself as a major threat to profit margins. Not only does it mean less revenues for not paid services, but can also increase direct or indirect costs.

If not properly assessed, fraud can become a critical issue for a telecom provider. As for subscription fraud, Estévez et al. (2006) refer to a 2.2% rate for a major telecom company in Chile. It is possible that this number is a lower bound to the true value of losses due to the reluctance of these companies to assume that their systems are so vulnerable to fraud.

The telecommunications industry generates and stores huge amounts of data regarding calls, SMS (Short Messages Service), MMS (Multimedia Mes-

[1] According to Collins English Dictionary - Complete and Unabridged.

Pedro Mendes L., Dias J. and Godinho P..
BI-LEVEL CLUSTERING IN TELECOMMUNICATION FRAUD.
DOI: 10.5220/0003718901260131
In *Proceedings of the 1st International Conference on Operations Research and Enterprise Systems* (ICORES-2012), pages 126-131
ISBN: 978-989-8425-97-3

saging Service) and Internet services of clients. Due to such an amount of transactions, only automated fraud detection systems (FDS) have enough power to skim over these data and select cases of possible anomalies. As there is no way to know the intention of people behind each of the transactions, algorithms must check for signs of fraud.

3 FRAUD DETECTION METHODOLOGIES

Different types of indicators have been used to identify fraud activity in the cell phone field. Moreau et al. (1996) divide these indicators in three types: 1) Usage indicators - related to the way in which a mobile telephone is used; 2) Mobility indicators - related to the mobility of the telephone. 3) Deductive indicators - which arise as a by-product of fraudulent behavior (e.g., overlapping calls and velocity checks).

The typical behavior of a given user can be called a *signature* (Cortes and Pregibon, 2001). Since it is not possible to analyze every single transaction on a real time basis, the signature tries to build on the idea that a user's behavior will not change much in a short time period. New data can then be compared to the signature of the user and if they are dissimilar, then a flag can be raised. As time evolves, so do typical behaviors of users, which implies that signatures have to be updated.

In event-driven updating, every new record is used to refresh the signature, eventually discarding its older record or giving an ever decreasing weight to it. Time-driven updating is less demanding in computation effort. The signature updating process is done using data collected during a time interval.

Another approach to detect cases of fraud can be summarized as "guilt by association" (Cortes et al., 2002). The idea behind this concept is that fraudsters tend to be closer to other fraudsters than they are to random accounts. As such, the authors consider a dynamic graph that changes in time where nodes represent the transactors and edges represent the interactions between pairs of transactors. Their paper shows that the probability that an account is fraudulent is an increasing function of the number of fraudulent nodes in its *community of interest* - union of sub-graphs centered on the account node.

3.1 Training and Test Data

Although a fraud detection system is meant to reduce costs for telecom companies, such a system can cost more in investigating false alarms than what it may save by reducing fraud. In order to address this problem, (Barse et al., 2003) propose to generate synthetic test data for fraud detection in an IP based video-on-demand-service. Synthetic data is defined as data that are generated by simulated users in a simulated system, performing simulated actions and presenting some advantages over authentic data:

- Some FDS need huge amounts of data for training that are not available in authentic data and can be synthetically generated.
- To be able to test the FDS to check how well it responds to variations of known frauds or how the detection rate is affected by new frauds.
- To be possible to compare several FDS in a benchmarking situation.

The norm is that fraud detection data is highly skewed or imbalanced (Phua et al., 2004). Since there are much more legitimate examples than fraudulent ones in a data set, an algorithm may have a high success rate without detecting any fraud. The authors propose two ways to address this problem:

1. Apply different algorithms (meta-learning). Each algorithm can be best used in particular data instances in accordance to its strengths.
2. Manipulate class distribution in such a way that the proportion of fraudulent minority class of data is increased. This may raise the chances for the algorithm to make correct predictions.

3.2 Machine Learning

Machine learning is a field of research devoted to the study of learning systems. It encompasses several fields, building upon ideas on statistics, mathematics, biology, engineering, cognitive science and other disciplines. There are two major methodologies of machine learning that can be used in fraud detection: *supervised* and *unsupervised* learning. By the former, it is meant that a classifier function grows knowing both the input data and the result. After the training is done, the classifier function should be able to predict the output of new input data that is fed to it.

Although classification methodologies can be effective at detecting fraud cases, several problems may arise: Since the algorithm was trained with labeled data, it is only sensible to types of fraud that where present in training data. Another problem refers to the fact that data may be mistakenly labeled as fraud and thus biasing later analysis. A third drawback may arise because of the necessity to have relatively large amounts of data that may be difficult or expensive to obtain.

Unsupervised learning focuses on finding hidden patterns in data. Data contains no output values which means that the purpose of these algorithms is to find patterns in the data that can help to give a structured representation of what could be firstly seen as noise.

When there is no need to know how a predictive solution has been reached, *neural networks* are a natural choice for a machine learning technique. Like other machine learning techniques, neural networks have been inspired in the biological world, in this case of the human brain. As Takagi (1991) states, a neuron consists of a cell body that is connected to other neurons by synapses. A neural network is the network of all these connected neurons that corresponds biologically to how the human brain operates. An artificial neural network (ANN) simulates the biological counterpart, where information input to the network produces an output. Like in the biological brain, it is expected that a learning process takes place, through the adaptation of synaptic weights, that can help achieve very good prediction results for unseen data. Krenker et al. (2009) proposed a system for mobile phone fraud detection based on a bidirectional ANN. The authors aim at predicting the behavior of individual mobile phone users and detecting fraud using both offline and real time processing. They report a 90% success rate at predicting time series that describe the behavior of a mobile phone user in optimal configuration. Although the output quality may be measured, neural networks lack explicative power.

3.3 Clustering

In telecommunications industry, the decision of characterizing fraudulent behavior is most of the time not a clear cut, as there is no way to guess a user's intentions. Many telecom providers, if not all, have human resources assigned to the task of deciding whether or not to consider the alerts raised by automatic algorithmic systems. This makes it necessary for human operators to understand and interpret the results provided by the algorithmic tools. One very common machine learning process that addresses this necessity is *clustering*. Clustering is a data mining tool that tries to join similar objects in homogeneous groups based in the values of their attributes.

The categorization of clustering algorithms is neither straightforward, nor canonical (Berkhin, 2006). A brief classification of clustering techniques can be as follows:

- Partitioning - Clusters are usually found in one pass over the data. As iteration progresses, points are allocated to existing clusters if similar, or they start a new cluster.
- Hierarchical - Clusters are built in a tree representation also called a dendogram. In agglomerative technique, clustering starts considering each point a cluster. The process keeps merging the more similar clusters until a stop criterion is reached. In divisive clustering, the logic is symmetric. The process begins by considering only one cluster and continues by subdividing the clusters into finer groupings.
- Density based - This kind of algorithms are capable of discovering clusters of any shape because they follow density paths. Although these techniques have an advantage in the consideration of outliers, they lack some interpretability.
- Grid based - A multi-dimensional space is divided into a large number of hyper-rectangular regions. Making use also of the concept of density, regions that are adjacent are merged until final clusters are found.

The k-means algorithm is probably the most popular clustering technique. There were several contributions for its development (MacQueen et al., 1967). Given a set of points and a number *k* of clusters, the k-means algorithm searches for a partition of the points into clusters that minimizes the within groups sum of squared errors. The algorithm starts by considering *k* observations from the data set and uses these as the initial means. *k* clusters are then created and surrounding values are assigned to each of these clusters. A value is assigned to the nearest cluster, i.e., to the one where the distance function between the point and the mean of the cluster is minimal. The algorithm proceeds in an iterative way until convergence has been reached. The centroid of each cluster becomes the new mean. Points are reassigned to clusters accordingly to the new centroids. Some of its properties are: 1) It is efficient in processing large data sets; 2) It often terminates at a local optimum; 3) The clusters have convex spherical shapes; 4) The clusters are expected to be of similar size. This algorithm has a severe constraint as it only works on numeric values.

3.4 Clustering with Categorical Data

Real world data as well as more specifically telecommunication data, consist of many values that are not numerical by nature, but categorical. The measure of distance between objects and clusters loses its significance when applied to categorical values. There are a number of algorithms that consider categorical data for clustering purposes. This paper will briefly present two of them. One, k-modes algorithm, (Huang, 1998), tries to extend the popular k-means algorithm to the realm of categorical values. The three main differences to k-means algorithm are:

- The numerical distance is substituted by a simple dissimilarity measure for categorical objects;
- K-modes uses modes instead of means for clusters;
- To minimize the clustering cost function, a frequency-based method to update modes is used.

The authors emphasize the scalability of the k-modes algorithm. A main shortcoming of this technique is that it needs the number of clusters to be defined a priori.

A more elaborate algorithm for dealing with categorical data is the ROCK algorithm (Guha et al., 2000). This is a hierarchical clustering algorithm that employs *links* and not distances when merging clusters. The Jaccard coefficient[2] (JC) has been used to measure the similarity between points. The authors argue against the Jaccard coefficient and justify their option for links. The JC is a measure of the similarity between only two points in question at a time, it does not take into consideration the neighborhood of points. As such, JC fails to capture the natural clustering of "not so well-separated" data sets with categorical attributes. For the ROCK algorithm, if the similarity between a pair of points exceeds a certain threshold then they are considered *neighbors*. The number of links between a pair of points is then the number of common neighbors of the points. Points belonging to a single cluster will in general have a large number of common neighbors, and consequently more links. The link-based approach adopts a global view to the clustering problem. It captures the global knowledge of neighboring data points into the relationship between individual pairs of points. The algorithm starts by considering a random sample of points from the data set. A hierarchical algorithm that employs links is applied to the sampled points. Finally, the clusters involving only the sampled points are used to assign the remaining data points to the appropriate clusters.

3.5 Evaluation Criteria

As previously said in the beginning of this section, fraud detection can be based on event or time driven methodologies. In a time driven assessment, one has to acknowledge that fraud may only be detected at the end of the time window that starts an instance of the FDS. In the literature, there are several performance measures for a FDS used by different authors: 1) Accuracy - Percentage of correctly predicted fraud instances; 2) True positive rate - Correctly detected fraud divided by actual fraud; 3) Receiver Operating Characteristic - *False positive rate = False Positives/(True Negatives + False Positives)* versus *true positive rate = True Positives/(True Positives + False Negatives)*; 4) Area under the Receiver Operating Curve (as in (Viaene et al., 2004)) - Single-figure summary measure of ROC performance; 5) Minimize Cross Entropy (Bishop, 1995) - How close predicted scores are to target scores; 6) Minimize Brier score (Hand, 1997) - Mean squared error of predictions.

4 PROPOSED ALGORITHM

4.1 Existing Telecom FDS

The telecom operator has a FDS based on alerts that are raised when suspicious behavior is detected (Cortesão et al., 2005). After these cases are flagged, they are dealt by fraud analysts that investigate all relevant information, regarding alert details, account information, and others. Cases that are classified as fraudulent by fraud analysts are then forwarded to a case manager to initiate consequent bureaucratic processes.

4.2 Data Sample

Data was received from a major national telecom provider that consists of some database tables, including only masked and truncated data in order to ensure the protection of personal data and confidential information. Therefore, the data provided by the telecom operator and used in this paper do not involve the disclosure of any personal data related to the telecom company subscribers or confidential information, ensuring full compliance with the applicable data protection legal framework. For numeric attributes data transformations occurred as follows. One thousand bins were created for each attribute. Each bin corresponds to a quantile of the distribution of the attribute (0.1%). The "nth" percentile of an observation variable is the value that cuts off the first *n* percent of the data values when it is sorted in ascending order. Each bin is labeled with an integer sequential number starting from zero (label 0 = 0.1%, 1 = 0.2%, ...). After completing the label - value dictionary, values of the attribute are changed for those of their corresponding bin label in the vector of values. In order to reduce the number of bins, sequential bins that have the same

[2] The Jaccard coefficient for similarity between transactions T1 and T2 is

$$\frac{|T_1 \cap T_2|}{|T_1 \cup T_2|}$$

value are aggregated. For each categorical attribute, a unique set of values is considered. Each unique value is then assigned a sequential integer value beginning in zero, which constitutes its label. A substitution in the original vector is performed analogously to that of numerical attributes.

4.3 Ongoing Research

This subsection aims at presenting ongoing work regarding a framework for detecting fraud in the telecom industry context.

4.3.1 First Level Clustering

Clustering was chosen as the tool to identify fraudulent behavior in data, due, mainly, to its explicative power. Although one of the database tables, concerning some client information, contains a field signalling detected frauds, the proposed methodology will follow the path of unsupervised learning. Clusters built in an unsupervised way are then compared to the fraud information contained in the mentioned table to analyze the viability of the current methodology. On the contrary, to use a supervised clustering algorithm with not so many available records would possibly restrict hidden insights available in data.

The clustering algorithm follows a bi-level structure. In the first level of clustering, a partition algorithm is used to separate records into clusters. Due to its effectiveness, an implementation of the Rock algorithm is used to find clusters in each of the transaction tables.

When this work is done, results of the partial clustering runs are aggregated to each record of the client information data set. This dataset is augmented in such a way that it will contain as many columns more as the total of clusters found in the previous level. For example, if four new clusters are found in the data set concerning voice calls, four new attributes will be added, one for each of these clusters. For each record, each of these attributes contains the value of the percentage of times that transactions belong to that cluster.

In telecom industry, many times, fraud can be characterized by strange behavior, which means that some records will be found as *outliers*. An attribute considering the percentage of times a call is considered an outlier (not belonging to any defined cluster) may be added, as well, to the cards data set. And similar logic should be considered to accommodate outliers of the remaining transaction tables.

4.3.2 Second Level Clustering

After the results of the first level clustering are aggregated to the client information data set the methodology proceeds to the second level of clustering. The number of attributes may now be very large, so may not be meaningful to try to extract knowledge from hidden patterns in such a high dimensional space. Combining all dimensions brings more noise into consideration. We may think about the hypothesis of a client performing fraud in voice calls, but not in SMS or Internet access.

One other factor that must be taken into consideration is that several clusters may cohabit for the same record. Since available data encompasses roughly five and a half weeks, the chosen algorithm must make room to the fact that fraud may begin at some time during that interval. An account may be completely legitimate until some date and, afterwards, start behaving in a fraudulent way. In order to take into account these different patterns of behavior, we will make use of a subspace clustering algorithm.

Traditional clustering techniques consider all dimensions of a data set in an attempt to get the most possible information about each point. But when data has a great number of attributes, generally more than a couple dozens, many of them become irrelevant, for each cluster. As Parsons et al. (2004) refer, in a high dimensional space, objects are very near of each other in what is called the *curse of dimensionality*.

Subspace clustering is a method that is able to uncover clusters by avoiding taking into consideration noise promoted by not meaningful attributes in each cluster. For example, regarding different types of fraud, we may find out different relevant clusters in the voice calls data set. One cluster may be related to abuse of international calling while another may show intensive national calling. For a subspace clustering algorithm, the same point may belong to different clusters. Continuing the example, the algorithm may find that a record belongs to a cluster where call intensity is low and *SMS* texting is high, and at the same time is part of another cluster of null Internet activity. Therefore, the algorithm must find all relevant clusters in all subspaces in order to discover hidden patterns in data.

The second level of clustering is meant to be performed on regular time interval basis. The updating process should therefore provide the operator with sufficient knowledge of trends in consumer profiles. Some of these may consist of new types of fraud mechanisms. In the meantime between two successive subspace clustering runs, transactional real time data keep being produced by the network operator operating system and should be verified. In order to process all these data, we decided to build a classifying algorithm. This algorithm should classify incoming data according to clusters defined by the subspace clustering method. Client's data that deviates from the "normal" clusters defining his previous behavior may be subject to further investigations by the network FDS. Suspects may also arise when the classification algorithm does not seem able to make new acquired data fit previous defined clusters. Also, data that is classified to clusters previously identified by being acquainted with fraud should be forwarded to the next step of the FDS. On the contrary, data that is classified as belonging to previously defined non risky clusters may be considered safe. Or, for a straight client when his new data corresponds to his clustering profile, no further measures should be taken, as data presents close to null risk.

5 CONCLUSIONS

Fraud presents itself as a major concern for telecommunication providers in today's competitive market. As competition tends to decrease operating margins, telecom companies try to cut in costs, such as those caused by fraudsters. Due to the great amount of data generated by each customer transaction, fraud detection cannot be addressed only by humans. Many methodologies have been presented in the literature. Unsupervised clustering has been used to automatically group transactions into clusters of similar records. When run in regular intervals of time, this tool allows a telecom to keep up to date with the ever evolving fraud dynamics.

This paper proposes a bi-level clustering algorithm to address fraud in telecommunications. In the first level, transactional records are grouped into clusters for each one of those services. Once this procedure is done, data is aggregated for each SIM card belonging to clients. As fraud, or just suspicious behavior, may be performed in only some of the services provided by the telecom, the clustering algorithm applied to the second level should be of the subspace type. Current research is concerned with the first level clustering.

REFERENCES

Barse, E., Kvarnstrom, H., and Jonsson, E. (2003). Synthesizing test data for fraud detection systems. In *Proceedings of the 19th Annual Computer Security Applications Conference* (ACSAC 2003). Citeseer.

Berkhin, P. (2006). A survey of clustering data mining techniques. *Grouping Multidimensional Data*, pages 25-71.

Bishop, C. (1995). *Neural networks for pattern recognition*. Oxford university press.

Cortes, C. and Pregibon, D. (2001). Signature-based methods for data streams. *Data Mining and Knowledge Discovery*, 5(3):167-182.

Cortes, C., Pregibon, D., and Volinsky, C. (2002). Communities of interest. *Intelligent Data Analysis*, 6(3):211-219.

Cortesão, L., Martins, F., Rosa, A., and Carvalho, P. (2005). Fraud management systems in telecommunications: a practical approach. *In Proceeding of ICT*.

Estévez, P., Held, C., and Perez, C. (2006). Subscription fraud prevention in telecommunications using fuzzy rules and neural networks. *Expert Systems with Applications*, 31(2):337-344.

Guha, S., Rastogi, R., and Shim, K. (2000). Rock: A robust clustering algorithm for categorical attributes* 1. *Information Systems*, 25(5):345-366.

Hand, D. (1997). *Construction and assessment of classification rules*, volume 15. Wiley.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2(3):283-304.

Krenker, A., Volk, M., Sedlar, U., Better, J., and Kos, A. (2009). Bidirectional Artificial Neural Networks for Mobile-Phone Fraud Detection. *Etri Journal*, 31(1):92-94.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281-297.

Moreau, Y., Preneel, B., Burge, P., Shawe-Taylor, J., Stoermann, C., and Cooke, C. (1996). Novel techniques for fraud detection in mobile telecommunication networks. *Proceedings of ACTS Mobile Telecommunications Summit*, Granada, Spain.

Parsons, L., Haque, E., and Liu, H. (2004). Subspace clustering for high dimensional data: a review. *ACM SIGKDD Explorations Newsletter*, 6(1):90-105.

Phua, C., Alahakoon, D., and Lee, V. (2004). Minority report in fraud detection: classification of skewed data. *ACM SIGKDD Explorations Newsletter*, 6(1):50-59.

Takagi, H. (1991). Introduction to fuzzy systems, neural networks, and genetic algorithms. *Intelligent Hybrid Systems: Fuzzy Logic, Neural Networks, and Genetic Algorithms*, pages 405-468.

Viaene, S., Derrig, R., and Dedene, G. (2004). A case study of applying boosting Naive Bayes to claim fraud diagnosis. *IEEE Transactions on Knowledge and Data Engineering*, 16(5):612-620.