International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:10, 2007

A Rule-based Approach for Anomaly Detection in Subscriber Usage Pattern

Rupesh K. Gopal, and Saroj K. Meher

***Abstract*—**In this report we present a rule-based approach to detect anomalous telephone calls. The method described here uses subscriber usage CDR (call detail record) data sampled over two observation periods: study period and test period. The study period contains call records of customers' non-anomalous behaviour. Customers are first grouped according to their similar usage behaviour (like, average number of local calls per week, etc). For customers in each group, we develop a probabilistic model to describe their usage. Next, we use maximum likelihood estimation (MLE) to estimate the parameters of the calling behaviour. Then we determine thresholds by calculating acceptable change within a group. MLE is used on the data in the test period to estimate the parameters of the calling behaviour. These parameters are compared against thresholds. Any deviation beyond the threshold is used to raise an alarm. This method has the advantage of identifying local anomalies as compared to techniques which identify global anomalies. The method is tested for 90 days of study data and 10 days of test data of telecom customers. For medium to large deviations in the data in test window, the method is able to identify 90% of anomalous usage with less than 1% false alarm rate.

***Keywords*—**Subscription fraud, fraud detection, anomaly detection, maximum likelihood estimation, rule based systems.

I. Introduction

THE mobile telecommunication industry has expanded dramatically in the last decade with the development of affordable mobile phone technology. With the increasing number of mobile phone users, global mobile phone fraud is also set to rise. Telecommunication fraud has identified itself as the single biggest cause of revenue loss for telecom carriers. According to a recent press release by Communications Fraud Control Association (CFCA), the annual global loss due to fraud is $ 55-60 billion USD [1].

Shawe-Taylor *et al* [2] distinguish six different types of fraud scenarios: subscription fraud, manipulation of private branch exchange (PBX) facilities or dial through fraud, freephone fraud, premium rate service fraud, handset theft and roaming fraud. Subscription fraud, which is defined as the use of telephone service without the intention of paying, is the most significant and prevalent worldwide telecom fraud [3,4]. In subscription fraud, the typical behavior of fraudsters is to abuse services by making significant usage of telecom services (for example, calling, messaging, internet, etc) before the bill is served. Customer applications are sometimes rejected by the company at the time of application if they find that it is risky to entertain customers who are likely to hold bad debt. Estevez *et al* [4] propose fuzzy rules and neural network model to detect subscription fraud at the time of application. Nevertheless, fraudsters can get into the network by providing false information or disguising as somebody else, known as identity theft. Another type of fraud, called superimposed fraud, is where a legitimate account is taken over by a fraudster (for example, SIM cloning). Both subscription fraud and superimposed fraud can be detected by effective user profiling and looking for patterns which are significantly different from normal patterns [5,6].

Fawcett and Provost [5,7] describe a user profiling method for fraud detection. They create account-specific thresholds rather than universal thresholds. This procedure takes 30 days of fraud-free traffic activity followed by a period of fraud. For each account a set of rules that distinguish fraud from non-fraud is developed. Pruning is done to get a set of rules that cover many accounts with different thresholds. Thus account specific thresholds are derived for detecting fraud activity on any given account day of a customer.

Taniguchi *et al* [6] and Hollmen [8] suggest a method of estimating the probability density function of subscribers' past usage behavior and then compute the probability of current usage with the model. They use a Gaussian mixture model for modeling the probability density function. The parameters of the Gaussian mixture model are estimated using online estimation (partial estimation) by the online version of the EM algorithm. The features using this model were daily number of calls and length of calls for national and international usage. Hence this approach performs statistical modeling of past behavior and produces a novelty measure of current usage as negative log likelihood of current usage.

Burge and Shawe-Taylor [9,10] use recurrent neural network to develop user profiles. They define two spans over the call data records – current behavior profile (CBP) and behavior profile history (BPH). They use second maximal entropy principle [11] to create statistical profiles and Hellinger distance to calculate the distance between CBP and BPH. If this distance is greater than some pre-determined threshold, alarm is raised. Moreau *et al* [12,13] use multilayer perceptron to classify fraud and non-fraud examples.

Cortes *et al* [14,15,16] use statistical summaries, called as signature, of users over two time windows similar to [9]. When the current network activity is different from the recent history, it need not be indicative of fraud – legitimate behavior might

Rupesh K. Gopal and Saroj K. Meher are with the Applied Research Group, Satyam Computer Services Limited, Entrepreneurship Center, SID Building, Indian Institute of Science campus, Bangalore 560012, India (phone: +91-80-23606830; fax: +91-80-23601011; email: rupesh.gopal@gmail.com, saroj_meher@satyam.com)

have been changed. To avoid such false positives, they also use a database of signatures of fraudsters obtained from fraud investigators. Signatures are updated using decayed functions. On a call-by-call basis, they compute the probability of observing the call assuming that it came from the legitimate account, relative to the probability of observing the call assuming that it came from a generic fraudster. Ferreira *et al* [17] also use signature based method to detect deviations in the behavior. Since the feature components of a signature are of different types, each component is evaluated by different distance functions.

Grosser *et al* [18] suggest the use of self-organizing map for creating resemblance groups for local, national and international calls for users. User profile is created in a method similar to [9] over two time windows and finally Hellinger distance is used to detect anomalous usage of mobile phone.

The aim of the present work is to build a simple rule-based model for anomaly detection. Similar to [9,10,17] we consider customer data over two time windows, which we call as study period and test period. Unlike [9,10,17] we build a generative model to *study* the normal behavior of customers over the study period. This model is *tested* against the data in the test period. Appropriate values for thresholds are learned in the study period. Any deviation in the behavior beyond the thresholds is used to raise an alarm. We explain our probabilistic model in next section. We present empirical results in Section III, and conclude in Section IV.

II. Probabilistic Model

It is rarely practical to access or analyze all call detail records for an account every time it is evaluated for fraud. Hence a common approach is to reduce the call records for an account to several statistics that are computed each period. The summaries that are monitored for fraud may be defined by subject matter experts, and thresholds may be chosen by trial and error. Or, decision trees or machine learning algorithms may be applied to training set of summarized account data to determine good thresholding rules [19]. Thresholding has some disadvantages, although. Thresholds may need to vary with type of account, type of call, and time of the day to be sensitive to fraud without setting off too many false alarms for legitimate accounts. Fawcett and Provost [7] describe a method of setting up account specific thresholds. However, their method is not easily applicable to subscription fraud as there is no period of fraud-free activity.

One approach to reduce false alarms is to *segment* subscribers based on their calling activity. Customer segmentation is useful from telecom customer relationship center (CRM) perspective. For example, in target marketing, subscribers are segmented as domestic, corporate and business accounts. We propose to segment subscribers based on similar calling activity. Subscribers with weekly average usage rate equal to μ with variance σ for different call types are grouped into one segment. Subscribers who belong to one segment have high degree of similarity (i.e., homogeneity); rare events such as sudden deviation from normal behavior can be easily captured. Models involving customer segmentation for lifetime value prediction of customers is popular in telecom industry [20].

For customers in each segment, we have call data coming from two disjoint time periods: study period and test period. Let M_1 be the number of days of call data in study period and M_2 be the number of days of call data in the test period. We assume that numbers of calls follow Poisson distribution and length of the call follows exponential distribution [21]. Under such an assumption, if $D = \{x_1, x_2, x_3, ..., x_M\}$ is a set of M observations sampled from the same distribution $f(x \mid \Theta)$, the likelihood function $L(\Theta \mid x_1, x_2, x_3, ..., x_M)$ is the probability that the data would have arisen fro a given value of Θ, regarded as a function of Θ, that is, $p(D \mid \Theta)$. D could be a *i.i.d* sample denoting number of calls or length of each call. Since both observed data and the model of interest is available, one can use parameter estimation technique like maximum likelihood estimation (MLE) [22].

Let us denote number of calls, length of call and type of call by following random variables:

$T =$ Random variable (r.v.) for type of call

$N =$ Random variable for number of calls of a given type

$L =$ Random variable for call length of a given call

For simplicity, we assume that number of calls a person makes is independent of the type of the call and the lengths of each call are independent of each other. For each call type t, we write down probability density function as follows:

$$P(N = n_t \mid T = t) = \frac{e^{-\mu_t} \mu_t^{n_t}}{n_t!}, \; t = \{1,2,3\} \quad (1)$$

and,

$$f(L_1 = d_1, L_2 = d_2, ..., L_{n_t} = d_{n_t} \mid N = n_t, T = t) = \prod_{j=1}^{n_t} \lambda_t e^{-\lambda_t d_j} \quad (2)$$

then,

$$p(T, N, L) = \sum_{t=1}^{3} P_t \; \frac{e^{-\mu_t} \mu_t^{n_t}}{n_t!} \; \prod_{j=1}^{n_t} \lambda_t \; e^{-\lambda_t d_j} \; = \sum_{t=1}^{3} p_t \quad (3)$$

p_1, p_2 and p_3 represent joint probability density functions for each type of call. Maximum likelihood estimate can now be used to estimate the parameters governing distribution in eqn. (3), as described below:

Let $X = \{x_1, x_2, x_3, ..., x_M\}$ be the r.v. that denotes the number of calls made by a person over a period of M days. Let $Y = \{y_1, y_2, y_3, ..., y_K\}$ be the r.v. that denotes the duration of all calls made till M^{th} day.

$$p_1(X, Y \mid \mu_1, \lambda_1) = \prod_{i=1}^{M} P_1 \; \frac{e^{-\mu_1} \mu_i^{x_i}}{x_i!} \; \prod_{j=1}^{x_i} \lambda_1 \; e^{-\lambda_1 y_j} \quad (4)$$

or,

$$p_1(X,Y \mid \mu_1,\lambda_1) = P_1^M \frac{e^{-\mu_1 M}\mu_{1=1}^{\sum_{i=1}^{M} x_i}}{\prod_{i=1}^{M} x_i!}\lambda_1^{\sum_{i=1}^{M} x_i} e^{-\lambda_1 \sum_{j=1}^{K} y_j} \quad (5)$$

Taking logarithm of both sides of (5) and maximizing the resulting function would yield,

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{M} x_i}{M} \quad (6)$$

and

$$\hat{\lambda}_1 = \frac{\sum_{i=1}^{M} x_i}{\sum_{j=1}^{K} y_j} \quad (7)$$

In general, maximum likelihood estimate for call type t is:

$$\hat{\mu}_t = \frac{\sum_{i=1}^{M} x_i}{M} \quad (8)$$

$$\hat{\lambda}_t = \frac{\sum_{i=1}^{M} x_i}{\sum_{j=1}^{K} y_j} \quad (9)$$

where M and K are the number of data points respectively in random variables X and Y.

MLE is used on estimate averages of number of calls and length of calls from study and test data, for each type of call and for all the customers belonging to that category. Next, rules are created for each type of call by considering the maximum of the averages obtained for all customers in that group. For example, if certain group has k customers, then threshold for a call type t is computed as:

$$\Delta_t^{\mu} = \max(\hat{\mu}_t^1, \hat{\mu}_t^2, \hat{\mu}_t^3, ..., \hat{\mu}_t^k)$$
$$\Delta_t^{\lambda} = \max(\hat{\lambda}_t^1, \hat{\lambda}_t^2, \hat{\lambda}_t^3, ..., \hat{\lambda}_t^k) \quad (10)$$

where $\hat{\mu}_t^j$ and $\hat{\lambda}_t^j$ are the maximum likelihood estimates of number of calls and call lengths of call type t for j^{th} customer of the group over the number of days specified. Thresholds are computed for all call types, and are combined to get a rule-set, which defines limits on all customers belonging to that group. The assumption here is that a customer's calling behavior could be different from others but should be within the threshold determined for the group as a whole. The rule-set hence developed raises alarms whenever the calling pattern crosses the threshold, thus facilitating the fraud analysts to identify which rule was broken and why the alarm was generated.

III. Empirical Results

As said earlier, a sudden change in calling pattern could also be interpreted as having a potential fraud. MLE is applied to data in study period and test period. The lengths of windows chosen for study and test periods are, respectively, M_1 days and M_2 days. For testing the model, customers of one segment are simulated with number of customers, $N = 500$. The study period data is generated with following parameters (Table I):

TABLE I
Average Number of Calls for Local, National and International Calls For Non-Anomalous Behavior Window

Average number of calls	Average length of a call (in seconds)
Local = 13 per week Nat. = 5 per 15 days Intl. = 3 per month	Local = 300 Nat. = 300 Intl. = 300

Test data is generated for all customers by changing parameter values as shown below ($M_1 = 90$ days)

Table II shows the results we have obtained. From case (1), we notice that when users are found to deviate from the general behavior of the group, the rule set is able to detect all of them (Detection rate = 100%), with just 10 days of current data. From case (2), we notice that when change in behavior is small, the method is able to detect change with almost 90% accuracy. From case (4) where we used 90 days of test data and no change is introduced, we notice that false alarm rate is less than 1%.

IV. Conclusion

This report illustrated the use of MLE for parameter estimation from historical (study) and current (test) data and learning of appropriate rules that detects changes in calling pattern of customers. It is to be noted here that, deviation of calling pattern of a particular customer is considered to be safe so long as the change falls within the maximum deviation expected from the category he/she belongs to. Hence, such a deviation could also mean that he/she is migrating to a higher category. Hence all alarms generated by the rule set must be cautiously scrutinized by fraud analysts before coming to any conclusion.

It is assumed that the calling patterns follow Poisson and exponential distributions. It is also assumed that the number of calls and length of each call are independent of the type of call. Also, the samples are assumed to be independent and identically distributed, which means that the averages defined previously remained constant throughout the 90 days period. These limitations will be relaxed in our next work. The method could be further enhanced by making use of use customer information and billing information. For example, payment

TABLE II
RESULTS FOR SMALL AND LARGE DEVIATIONS IN THE TEST PERIOD

Case	M_2 (in days)	Change in parameter values	Average number of calls	Average length of a call (in seconds)	Number of customers deviated	Result
1	10	Large	Local = 25 per week Nat. = 15 per 15 days Intl. = 10 per month	Local = 600 Nat. = 500 Intl. = 600	500	Detection accuracy = 100 %
2	10	Small	Local = 15 per week Nat. = 8 per 15 days Intl. = 5 per month	Local = 420 Nat. = 360 Intl. = 360	457	Detection Accuracy = 91 %
3	90	Small	Local = 15 per week Nat. = 8 per 15 days Intl. = 5 per month	Local = 420 Nat. = 360 Intl. = 360	485	Detection Accuracy = 97 %
4	90	No change	Local = 13 per week Nat. = 5 per 15 days Intl. = 3 per month	Local = 300 Nat. = 300 Intl. = 300	4	False Alarm Rate = 0.8 %

pattern of a customer in past few months and customer demographic information can be used to generate very effective and confident rules.

ACKNOWLEDGMENT

Authors thank Sridhar Gangadharpalli and Sridhar Varadarajan for discussion and helpful comments.

REFERENCES

[1] CFCA press release, http://www.cfca.org/pdf/press/3-28-06PR.pdf, 2006 (last accessed on October 18, 2007).
[2] Shawe-Taylor, J., Howker, K., Gosset, P., Hyland, M., Verrelst, H., Moreau, Y., et al, "Novel techniques for profiling and fraud in mobile telecommunications," In P. J. G. Lisboa, B. Edisbury, A. Vellido (Eds.), *Business Applications of Neural Networks. The State-of-the-art of Real World Applications*, pp. 113-139, Singapore: World Scientific.
[3] Hoath. P, "What's new in telecoms fraud?," *Computer Fraud and Security*, Vol. 1, pp. 10-14, 1998.
[4] Estevez, P. A, Held, C. M., Perez, C. A., "Subscription Fraud Prevention in Telecommunications using Fuzzy Rules and Neural Networks," *Expert Systems with Applications*, Vol. 31, pp. 337-344, 2006.
[5] Fawcett, T., Provost, F., "Combining Data Mining and Machine Learning for User Profiling," In *AI approaches to fraud detection and risk management, workshop technical report WS-97-07*, pp. 14-19, AAAI Press, 1997.
[6] Taniguchi, M., Haft, M., Hollmen, J., Tresp, V., "Fraud Detection in Communication Networks using Neural and Probabilistic Methods," *IEEE International Conference in Acoustics, Speech and Signal Processing*, Vol. 2, pp. 1241-4, 1998.
[7] Fawcett, T., Provost, F., "Adaptive Fraud Detection," *Data Mining and Knowledge Discovery*, Vol. 1, pp. 291-316, 1997.
[8] Hollmen, J., "Novelty Filter for Fraud Detection in Mobile Communications Networks," Technical Report submitted to Department of Computer Science and Engineering, Helsinki University of Technology, 1997.
[9] Burge, P., Shawe-Taylor, J., "Detecting cellular fraud using adaptive prototypes," *AAAI Workshop on AI Approaches to Fraud Detection and Risk Management*. AAAI Press, Menlo Park, CA.
[10] Burge, P., Shawe-Taylor, J., "An Unsupervised Neural Network Approach to Profiling the Behaviour of Mobile Phone Users for Use in Fraud Detection," *Journal of Parallel and Distributed Computing*, Vol. 61, 2001.
[11] Grabec, I, "Modelling of Chaos by a Self-Organizing Neural Network," *Proceedings of International Conference on Artificial Neural Networks*, Vol. 1, pp. 151-156, Elsevier publications, 1989.
[12] Moreau, Y., Vandewalle, J., "Detection of mobile phone fraud using supervised neural networks: A first prototype," *International Conference on Artificial Neural Networks*, 1065- 1070. Springer, Berlin., 1997.
[13] Moreau, Y., "A hybrid system for fraud detection in mobile communication," *European Symposium on Artificial Neural Networks*, pp. 447-454, 1999.
[14] Cortes, C., Prebigon, D., "Signature-Based Methods for Data Streams," Data Mining and Knowledge Discovery, pp. 167-182, 2001.
[15] Cortes, C., Prebigon, D., Volinsky, C., "Communities of Interest," *Intelligent Data Analysis 2001*, pp. 105-114, 2001.
[16] Cortes, C., Prebigon, D., Volinsky, C., "Computational Methods for Dynamic Graphs," *Journal of Computational and Graphical Statistics*, Vol. 12, pp. 950-970, 2003.
[17] Ferreira, et al. "Establishing Fraud Detection Patterns Based on Signatures," *Industrial Conference on Data Mining*, LNAI Springer-Verlag, 2006.
[18] Grosser, et al. "Detecting Fraud in Mobile Telephony Using Neural Networks," Lecture Notes in AI 3533, Springer-Verlag, 2005
[19] Cahill, M. H., Lambert, D., Pinheiro, J. C., Sun, D. X. "Detecting Fraud in Real World", In J. Abello, P. M. Pardalos, M. G. C. Resende (Eds.), *Handbook of Massive Datasets*, pp. 913-930, Kluwer Academic Publishers, 2002.
[20] Rosset, S., Neumann, E., Eick, U., Vatnik, N., "Customer LTV modelling and its use for Customer Retention Planning", *Data Mining and Knowledge Discovery*, Vol. 7, pp. 321-339,2003.
[21] Samfat, D., Molva, R., "IDAMN: an intrusion detection architecture for mobile networks", *IEEE Journal on Selected areas of Communications*, Vol. 15, pp. 1373-1380, 1997.
[22] Duda, R. O., Hart, P. E., Stork, D. G., "Pattern Classification", Second edition, John-Wiley and Sons, 2001.