FRAUD DETECTION IN COMMUNICATIONS NETWORKS USING NEURAL AND PROBABILISTIC METHODS

Michiaki Taniguchi, Michael Haft, Jaakko Hollmén, Volker Tresp

Siemens AG, Corporate Technology
Department Information and Communications
D-81730 Munich, Germany
e-mail: Michiaki.Taniguchi@mchp.siemens.de

ABSTRACT

Fraud detection refers to the attempt to detect illegitimate usage of a communications network. Three methods to detect fraud are presented. Firstly, a feed-forward neural network based on supervised learning is used to learn a discriminative function to classify subscribers using summary statistics. Secondly, Gaussian mixture model is used to model the probability density of subscribers' past behavior so that the probability of current behavior can be calculated to detect any abnormalities from the past behavior. Lastly, Bayesian networks are used to describe the statistics of a particular user and the statistics of different fraud scenarios. The Bayesian networks can be used to infer the probability of fraud given the subscribers' behavior. The data features are derived from toll tickets. The experiments show that the methods detect over 85 % of the fraudsters in our testing set without causing false alarms.

1. INTRODUCTION

Fraud in communications networks refers to the illegal access to the network and the use of its services. It is estimated that a mobile phone network operator may lose as much as million dollars a day due to fraudulent usage of mobile phones. The development of intelligent data analysis methods for fraud detection can be well motivated from an economic point of view. Additionally, the reputation of a network operator may suffer from an increasing number of fraud cases.

In this paper, we present three approaches to fraud detection. First, feed-forward neural network based on supervised learning is used to learn a non-linear discriminative function between classes fraud and non-fraud. Secondly, density estimation with Gaussian mixture models is applied to modeling the past behavior of each subscriber and detecting any abnormalities from the past behavior. Lastly, Bayesian networks are used to define probabilistic models under the assumptions fraud and non-fraud. The Bayes' rule is used to invert these measures to calculate the probability for fraud given the subscribers' behavior.

The data used in all three approaches are based on toll tickets, which are call records stored for billing purposes. The toll tickets are created for each phone call made and include information like to identification of the caller, starting time of the call, duration of the call, the called party number to mention a few. The suggested solutions, however, are applicable in other types of networks, which store calling information analog to the toll tickets.

To assess the performance of the method, the Receiver Operating Characteristic (ROC) curves are shown for each method. The ROC curves show the detection probability as the function of false alarm probability [10]. The experiments show a high recognition rate taking into account the real-world requirement of low false alarm probability.

2. FRAUD DETECTION

Although there is growing interest in creating fraud detection engines, the articles in the literature are scarce. Barson et al. [1] use feed-forward neural networks based on supervised learning to detect mobile phone fraud in their simulated database of call records. They simulate six types of users ranging from low use local users to high use international business users. They report their neural network classifier to correctly classify 92.5% of the calling data. Their work does not include any comment on the false alarm probability and also is not comparable with our work as it is based on simulated data. Moreau et al. [6] report fraud detection in a real mobile communication networks. Their approach is based on feed-forward neural networks with supervised learning. They use different user-profiles and also consider comparisons between past and present behavior. They conclude that although their work is in a prototype phase, they have demonstrated a great potential with their approach.

2.1. Neural networks with supervised learning

The feed-forward neural networks can be used to represent an arbitrary non-linear mapping, provided that we have data exemplifying mapping as input-output pairs. The problem of supervised learning is to adapt the neural network weights so that the input-mapping corresponds to the input-ouput samples the teacher has provided. The feed-forward mapping of a three-layer neural network is defined by the Equation 1.

$$y = \sum_{j=0}^{M} w_j g(\sum_{i=0}^{d} w_{ji} x_i) \quad (1)$$

As outputs we use a linear output signifying the class membership. The g is a non-linear mapping (e.g. $g(x) = \tanh(x)$), w_j are the weights between the output and the hidden layer and w_{ji} are the weigths going from the ith input to the jth neuron in the hidden layer. The feed-forward network consists of five hidden units and one binary output. The neural network was trained using Quasi-Newton optimization. In order to constraint the complexity of the mapping, weight decay type of regularization was used. In weight decay, the cost function (error between the network output and the target) is augmented with an additional term $\lambda \sum_i {w_i}^2$. This term penalizes the large networks and thus for complex mappings [2] by reducing the variance. The magnitude of the penalty is determined by the coefficient λ. In our experiment, for $\lambda = 1$ the network performed the best result.

The features used in this application were average and the standard deviation of the duration and the number of calls made during the day, maximum duration and number of calls per day during the observed time period. The data included 303 samples from users exhibiting fraudulent behavior and some 2100 users exhibiting legitimate subscribers. The data set was divided into a training set and a testing set. We interpreted the output of the neural network as the posterior probabilities of fraud given the inputs.

The performance of the neural detector trained with $\lambda = 1$ is shown in Figure 1. Classifying the fraudulent users is more sensible to the regularization factor λ while the performance rate of the correctly classified non-fraudster is almost 100 % indepedent on λ. Although this result seems very promising, it must be noted that labeled data is hard and expensive to acquire. Our data set contained time series of call records from both fraudulent and legitimate subscribers, although the time of fraud was not recorded. For the purpose of classifying subscribers as fraudulent and legitimate, the features considered were summary statistics over the whole observed time period. Such a detection system would be useful for the analyzing purpose in an off-line mode. For the on-line detection, the input features would be obtained by using the sliding time windows.

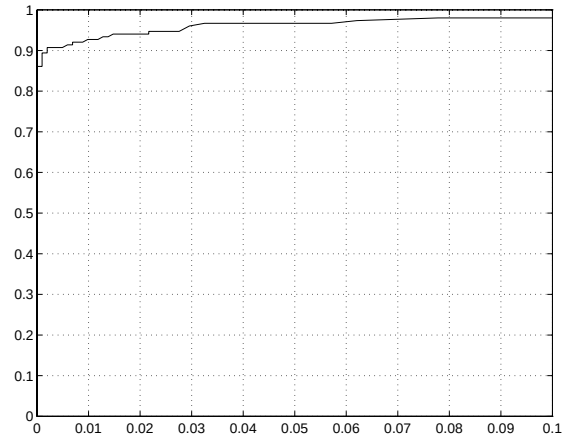

Figure 1: The ROC curve for the feed-forward network is presented. The method detects over 85 % of fraud cases without causing false alarms.

2.2. Probability density estimation methods

The problem of probability density estimation is to model a probability density function $p(\mathbf{x})$, given a finite number of data points drawn from that density [2]. We estimate the probability density function of the mobile phone subscribers' past behavior and then to compute the probability of current usage with the model. To model the probability density function, we use a Gaussian mixture model [9], which is a sum of weighted component densities of Gaussian form. This is shown in Equation 2.

$$p(\mathbf{x}) = \sum_{j=1}^{M} p(\mathbf{x}|j)\, P(j) \quad (2)$$

The $p(\mathbf{x}|j)$ is the jth component density of Gaussian form and the $P(j)$ is its mixing proportion. The parameters of the Gaussian mixture model can be estimated using the EM algorithm. The EM algorithm is a general approach to iterative computation of maximum-likelihood estimates when the observations can be viewed as incomplete data [3].

The features using this approach were the daily number of calls and the length of the calls occurring during the office hours, the evening hours and the night hours. These six features were considered for both national and international calls resulting in twelve features. Using these features as inputs, we estimate the probability density function of the large public. We call this the general model. We specialize the general model by re-estimating the mixing proportions for each subscriber dynamically after each sampling period as new data becomes available. Whereas the means and the variances of the subscriber specific models are common, only the mixing proportions are different between the subscribers' models. This modeling approach is motivated

by its computational feasibility while retaining its expressive power.

In order to estimate the density of past behavior in batch mode, we should retrieve the data from the last k days and adapt the mixing proportions to maximize the likelihood of past behavior. This is done for each subscriber separately. While this approach seems first suited for the job, this requires too much interaction with the billing system to be used in practice. To avoid this burdensome processing of data, we formulate our partial estimation procedure using on-line estimation. The on-line version of the EM algorithm was first introduced by Nowlan [7].

$$P(j)^{new} = \alpha P(j)^{old} + P(j|\mathbf{x}) \tag{3}$$

Remembering that the new maximum likelihood estimate for $P(j)$ is computed as the expected value of $P(j|\mathbf{x})$ over the whole data set with the current parameter fit, we can easily formulate a recursive estimator for this expected value as can be seen in Equation 3. The decay term α determines the efficient length of the exponentially decaying window in the past.

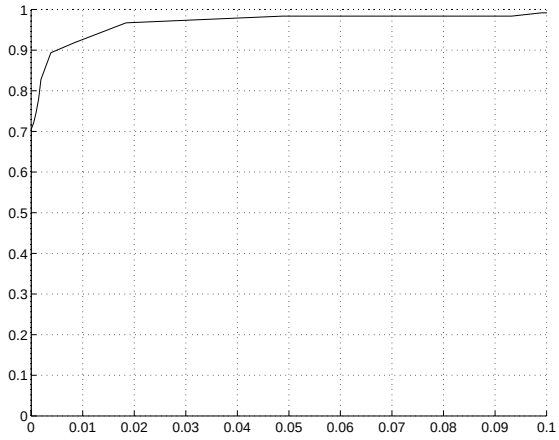

Figure 2: The ROC curve for the Gaussian mixture model operating in on-line mode is shown. The method detects over 70 % fraud cases in our testing set with no false alarms.

The approach performs statistical modeling of past behavior and produces a novelty measure of current usage as a negative log likelihood of current usage [4]. The detection decision is then based on the output of this novelty filter. Subscriber specific modeling by means of adaptive mixing proportions allows different subscriber profiles and also slowly changing behavior. Taking into account the unsupervised learning approach, the results are encouraging.

2.3. Bayesian networks

There are no deterministic rules which allow us to identify a subscriber as a fraudster. We may at best formulate our degree of belief in fraudulent behaviour. Graphical models such as Bayesian networks supply a general framework for dealing with uncertainty in a probabilistic setting [8] and thus are well suited to tackle the problem of fraud detection. Every graph of a Bayesian network codes a class of probability distributions. The nodes of that graph comply to the variables of the problem domain. Arrows between nodes denote allowed (causal) relations between the variables. These dependency are quantified by conditional distributions for every node given its parents.

Bayesian networks can be used as an expert system. This means that an expert of the problem domain draws a graph according to assumed causal impacts between variables. The corresponding conditional distributions can then be injected by the expert as well, who makes judgements about the causal relations or are estimated from data using traditional estimation methods. Once a Bayesian network is set up, we can infer probabilities for unknown variables by inserting evidence in the network and propagating the evidence through the network using propagation rules [5].

For the purpose of fraud detetcion, we construct two Bayesian networks to describe the behavior of mobile phone subscribers. First, a Bayesian network is constructed to model behavior under the assumption that the subscriber is fraudulent (F) and another model under the assumption the subscriber is a legitimate user (NF), see Figure 3. The 'fraud net' is set up by using expert knowledge. The 'user net' is set up by using data from non fraudulent subscribers. During operation user net is adapted to a specific user based on emerging data. By inserting evidence in these networks (the observed user behaviour x derived from his toll tickets) and propagating it through the network, we can get the probability of the measurement x under two abovementoined hypotheses. This means, we obtain judgements to what degree an observed user behaviour meets typical fraudulent or non-fraudulent behaviour. These quantities we call $p(x|NF)$ and $p(x|F)$. By postulating the probability of fraud $P(F)$ and $P(NF) = 1 - P(F)$ in general and by applying Bayes' rule, we get the probability of fraud, given the measurement x,

$$P(F|x) = \frac{P(F)p(x|F)}{p(x)}, \tag{4}$$

where the denominator $p(x)$ can be calculated as

$$P(x) = P(F)p(x|F) + P(NF)p(x|NF) \tag{5}$$

The fraud probability $P(F|x)$ given the observed user behaviour x can be used as an alarm level. The ROC curve

which we obtained varying the alarm level is shown in Figure 4.

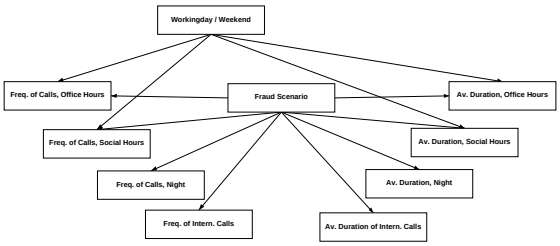

Figure 3: Bayesian network in fraud detection. The nodes in the Bayesian network denote variables and the arrows between the nodes causal relations between the variables.

On the one hand, Bayesian networks allow the integration of expert knowledge, which we used to initially set up the models. On the other hand, the user model is retrained in an unsupervised way using data. Thus our Bayesian approach incorporates both, expert knowledge and learning. The combination of the user and fraud model gives a reliable evaluation of the data.

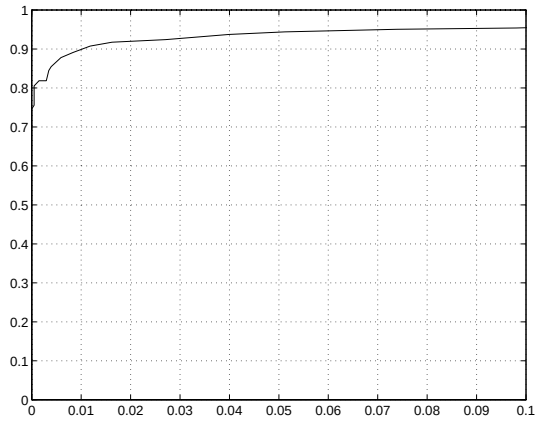

Figure 4: The ROC curve for the Bayesian network is shown. The result is similar to Figure 2 in the relevant parts of the ROC curve.

3. SUMMARY

Three approaches to fraud detection in communications networks were presented. The performance of these methods has been validated with data from a real mobile communications network. The feature vectors used in this application describing the subscribers' behavior were based on toll tickets. For supervised learning approach, the features used were summary statistics over the whole observed time period as no times of fraud were recorded in the data. For the two latter approaches, the features described the daily behavior for every subscriber. The supervised approach needs labeled data for the training where the two latter approach can handle the data without labels.

To improve the fraud detection system, the combination of the three presented methods could be beneficial. Also, the incorporation of rule based systems could show an improvement. Our results encourage us to investigate the performance of our methods in a mobile phone networks of real-world size.

4. REFERENCES

[1] P. Barson, S. Field, N. Davey, G. McAskie, and R. Frank. The detection of fraud in mobile phone networks. *Neural Network World*, 6(4):477–484, 1996.

[2] Chris Bishop. *Neural Networks in Pattern Recognition*. Oxford Press, 1996.

[3] A. P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[4] Jaakko Hollmén. Novelty filter for fraud detection in mobile communications networks. Technical Report A48, Helsinki University of Technology, Laboratory of Computer and Information Science, October 1997.

[5] Finn V. Jensen. *An Introduction to Bayesian Networks*. UCL Press, 1996.

[6] Yves Moreau, Herman Verrelst, and Joos Vandewalle. Detection of mobile phone fraud using supervised neural networks: A first prototype. In *International Conference on Artificial Neural Networks Proceedings (ICANN'97)*, pages 1065–1070, October 1997.

[7] S.J. Nowlan. *Soft Competitive Adaptation: Neural Network Learning Algorithms based on Fitting Statistical Mixtures*. PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, 1991.

[8] J. Pearl. *Probabilistic reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[9] R.A. Redner and H.F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–234, 1984.

[10] Louis L. Scharf. *Statistical Signal Processing- Detection, Estimation and Time Series Analysis*. Addison-Wesley, 1990.