Jaakko Hollmén

Probabilistic Approaches to Fraud Detection

Licentiate's Thesis

Helsinki University of Technology
Department of Computer Science and Engineering

HELSINKI UNIVERSITY OF TECHNOLOGY **ABSTRACT OF THE LICENTIATE'S THESIS**

Author and name of the thesis:

Jaakko Hollmén

Probabilistic Approaches to Fraud Detection

Date: 15.12.1999 **Number of pages:** 37

Department:	Department of Computer Science and Engineering
Professorship:	Tik-115 Information Sciences
Supervisor:	Professor Olli Simula
Instructor:	Professor Olli Simula

In telecommunication, a network operator may loose several percent of its revenue due to fraud. Fraud may be defined as dishonest or illegal use of services, with the intention to avoid or to reduce service charges. Fraud detection attempts to discover fraudulent activity in a telecommunication network.

In this thesis, the problem of fraud detection is treated as a pattern recognition problem. Detection is based on the call data of mobile phone subscribers, which are used for describing calling behavior. The goal is to develop learning methods that detect fraud accurately based on call data.

Fraud detection can be based on two hypotheses. On the one hand, models of both fraudulent behavior and normal behavior may be formulated, on the other hand, usage profiles may be used to detect abrupt changes in calling behavior. Both of these approaches are used in this thesis. Models are realized using probabilistic models and neural networks.

This thesis consists of an introduction and three publications.

Keywords: fraud detection, mobile networks, probabilistic model, neural network, Self-Organizing Map, ROC analysis.

TEKNILLINEN KORKEAKOULU **LISENSIAATINTUTKIMUKSEN TIIVISTELMÄ**

Tekijä ja työn nimi:

Jaakko Hollmén

Probabilistiset menetelmät väärinkäytön detektoinnissa

Päivämäärä: 15.12.1999 **Sivumäärä:** 37

Osasto:	Tietotekniikan osasto
Professuuri:	Tik-115 Informaatiotekniikka
Valvoja:	Professori Olli Simula
Ohjaaja:	Professori Olli Simula

Telekommunikaation alalla, verkko-operaattori voi menettää useita prosentteja tulostaan väärinkäytön vuoksi. Väärinkäyttö (engl. fraud) voidaan määritellä epärehelliseksi tai laittomaksi palveluiden käytöksi tarkoituksena välttää tai pienentää palvelumaksuja. Väärinkäytön detektoinnilla yritetään paljastaa epärehellistä toimintaa telekommunikaatioverkossa.

Tässä tutkimuksessa käsitellään väärinkäytön detektointia hahmontunnistusongelmana. Detektointi perustuu matkapuhelimien tilaajien yhteystietoihin, joita käytetään puhelukäyttäytymisen kuvaukseen. Tavoitteena on kehittää oppivia menetelmiä, jotka detektoivat väärinkäytön tarkasti yhteystietoihin perustuen.

Väärinkäytön detektointi voi perustua kahteen hypoteesiin. Toisaalta, sekä epärehellistä että normaalia käytöstä voidaan mallittaa ja toisaalta, käyttöprofiileja voidaan käyttää puhelukäyttäytymisessä tapahtuvaan, äkillisen muutoksen detektointiin. Tässä työssä käytetään molempia lähestymistapoja. Malleina käytetään probabilistisia malleja sekä hermoverkkoja.

Tämä tutkimus koostuu johdannosta sekä kolmesta julkaisusta.

Avainsanat: väärinkäytön detektointi, mobiiliverkot, probabilistinen malli, hermoverkko, itseorganisoiva kartta, ROC-analyysi.

Contents

List of Publications

This thesis consists of an introduction and the following publications:

Publication 1 Taniguchi, M., M. Haft, J. Hollmén, and V. Tresp (1998, May). Fraud detection in communications networks using neural and probabilistic methods. In *Proceedings of the 1998 IEEE International Conference in Acoustics, Speech and Signal Processing (ICASSP'98)*, Volume 2, pp. 1241–1244.

Publication 2 Hollmén, J. and V. Tresp (1999). Call-based fraud detection in mobile communications networks using a hierarchical regime-switching model. In M. Kearns, S. Solla, and D. Cone (Eds.), *Advances in Neural Information Processing Systems: Proceedings of the 1998 Conference (NIPS'11)*, pp. 889 – 895. MIT Press.

Publication 3 Hollmén, J., V. Tresp, and O. Simula (1999, September). A self-organizing map algorithm for clustering probabilistic models. In *Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN'99)*, Volume 2, pp. 946–951. IEE.

Preface

The work presented in this thesis was carried out at the Laboratory of Computer and Information Science in the Department of Computer Science and Engineering of the Helsinki University of Technology and in the group of Neural Computation in the Department of Information and Communications at Siemens Corporate Technology in Munich.

I wish to thank my supervisor, professor Olli Simula for guidance and encouragement during the thesis work. The group of Neural Computation at Siemens, headed by professor Bernd Schürmann, provided an interesting subject for research. Also, I wish to thank my co-authors and colleagues for collaboration. Without the support and encouragement from my family, Suvi and Risto, this thesis work would not have been possible.

This work was funded by the Laboratory of Computer and Information Science at the Helsinki University of Technology and the Department of Information and Communications at Siemens Corporate Technology. A travel grant was received from the NIPS Foundation. Additional financial support was received from the Finnish Foundation for the Promotion of Technology (Tekniikan Edistämissäätiö) and the NEuroNet — the European Network of Excellence in neural networks, set up under European Commission's ESPRIT III Program.

Glossary of Terms and Abbreviations

BMU	Best-Matching Unit (also winner unit)
DAG	Directed Acyclic Graph
EM	Expectation Maximization Algorithm
GSM	Global System for Mobile Communications
HMM	Hidden Markov Model
IMSI	International Mobile Subscriber Identity
KL	Kullback-Leibler distance
ROC	Receiver Operating Characteristic curve
SOM	Self-Organizing Map
$\alpha(t)$	adaptation gain value, also learning rate
c	index of the winner unit
$\delta(x)$	Dirac impulse function at x
i, k	unit index
$h^c(t, k)$	neighborhood kernel function
$m^i(t), m^i$	weight vector of the unit i
$P(S)$	probability of hidden state vector $s_1, \ldots, s_T$
$P(s_t)$	probability of a hidden variable s at time t
$P(s_t \vert s_{t-1})$	conditional probability of s_t given s_{t-1}
$p(x)$	probability density of x
$q(x;\ \theta)$	probability density of x (parameterized by θ)
r^k, r^c	location vector inside the array of neurons
$\sigma(t)$	neighborhood kernel width function
t	time variable
$x(t), x_i$	measurement vector
$x \sim p(x)$	x is distributed according to $p(x)$
Y	observed variable $y_1, \ldots, y_T$
y_t	observed variable y at time t
$\vert\vert.\vert\vert$	Euclidean distance

Chapter 1

Fraud Detection

1.1 Introduction

Telecommunication industry suffers losses in the order of billions of US dollars annually due to fraud in its networks (Davis and Goyal 1993; Johnson 1996; Parker 1996; O'Shea 1997; Pequeno 1997; Hoath 1998). In addition to financial losses, fraud may cause distress, loss of service, and loss of customer confidence (Hoath 1998). The financial losses account for about 2 percent to 6 percent of the total revenue of network operators, thus playing a significant role in total earnings. Keeping in mind that operators are facing increasing competition and that losses have been on the rise (Parker 1996), fraud has gone from being a problem carriers were willing to tolerate to being one that dominates the front pages of both trade and general press (O'Shea 1997). Johnson (1996) also affirms that network operators see call selling as a growing concern.

Johnson (1996) defines fraud as any transmission of voice or data across a telecommunication network where the intent of the sender is to avoid or reduce legitimate call charges. In similar vein, Davis and Goyal (1993) define fraud as obtaining unbillable services and undeserved fees. According to Johnson (1996), the serious fraudster sees himself as an entrepreneur, admittedly utilizing illegal methods, but motivated and directed by essentially the same issues of cost, marketing, pricing, network design and operations as any legitimate network operator. Hoath (1998) considers fraud as attractive from the fraudsters' point of view, since detection risk is low, no special equipment is needed, and the product in question is easily converted to cash. In all, fraud detection should be seen as an important countermeasure in a combat against fraud.

Historically, earlier types of fraud used technological means to acquire free access. Cloning of mobile phones by creating copies of mobile terminals with identification numbers from legitimate subscribers was used as a means of gaining free access (Davis and Goyal 1993). In the era of analog mobile terminals, identification numbers could be easily captured by eavesdropping with suitable receiver equipment in public places, where mobile phones were evidently used. One specific type of fraud, tumbling was quite prevalent in the United States. It exploited deficiencies in the validation of subscriber identity, when a mobile phone subscription was used outside of the subscriber's home area. The fraudster kept tumbling (switching between) captured identification numbers to gain access. Davis and Goyal (1993) state that the tumbling and cloning fraud have been serious threats to operators' revenues. As a countermeasure, first fraud detection systems examined whether two instances of one subscription were used at the same time (overlapping calls detection mechanism) or at locations far apart in temporal proximity (velocity trap). Both of these fraud types were later invalidated by technological improvements. However, new forms of fraud came to existence. A few years later, O'Shea (1997) reports the so-called subscription fraud to be the trendiest and the fastest-growing type of fraud. In similar spirit, Hoath (1998) characterizes subscription fraud as being probably the most significant and prevalent worldwide telecommunication fraud type. In subscription fraud, a fraudster gets a subscription (possibly with false identification) and starts a fraudulent activity with no intention to pay the bill. It is indeed non-technical in nature and by call selling, the entrepreneur-minded fraudster can generate significant revenues for a minimal investment in a very short period of time (Johnson 1996). From the above explanation it is evident that the detection mechanisms of the first generation soon became obsolete. Fawcett and Provost (1997) also report poor performance with these methods. The more advanced detection mechanisms must be based on the calling activity itself, which is also the subject of this thesis. Basing detection on the calling activity, there are two problems to be solved: the representation problem and the modeling problem. The representation problem involves describing domain-specific behavior with a representation that is suitable for the modeling approach used. For the methods used in this thesis, call data are mapped to numeric feature variables, which summarize the relevant about the domain. Naturally, the choice of representation is tightly coupled with the choice of models. The subject of this thesis is to model the quantified behavior with probabilistic models and neural networks. These types of models need numeric random variables as the input data.

The goal in developing a fraud detection system in the context of this thesis is to devise a machinery to automatically detect fraudulent activity accurately, based on the call data. Using the absolute analysis approach, this requires modeling fraudulent activity and normal activity, whereas in the differential approach, it is sufficient to formulate usage profiles and to detect abrupt changes. Both of the approaches are used in this thesis. Models are realized using neural networks and probabilistic models. Call data are from mobile communications networks and exhibit both normal and fraudulent calling behavior.

Although the ability to distinguish between fraud and normal behavior is the basic building block for any operational fraud detection system, there are further requirements in building a functional system. More specifically, the actions on the alarms given by the detection system must be specified and the consequences of these actions carefully weighted. When opted with different actions, a cost model may help in choosing the most suitable action in the light of possible mistakes made. Although these are important issues, they are clearly outside of the scope of this thesis. This thesis concentrates on the mere classification of calling behavior and the proper assessment of the detection accuracy. In further work, we may pursue these issues more closely.

The Chapter 1 is introductory in nature and presents a review of the published works in fraud detection in telecommunication. Also, related fields such as credit card fraud detection and intrusion detection in computer systems are reviewed. In Chapter 2, we describe the call data used in this thesis. Then we proceed to Chapter 3, where methods, in particular probabilistic networks and Self-Organizing Maps are reviewed. In Chapter 4, we discuss the assessment issues. Chapter 5 lists the contents of the publications and contributions of the author. In the end, the work is summarized and the broader applicability of the methods is discussed.

1.2 Previous Work

In this section, we attempt to summarize published works with relevance to fraud detection in telecommunication networks. Section 1.3 presents fraud detection methods in related fields, such as credit card fraud detection, intrusion detection in computer systems and other fields, such as health care fraud detection. Although the term fraud has a particular meaning in legislation, we use this established term broadly to mean misuse, dishonest intention or improper conduct without implying any legal consequences.

Fraud in telecommunication networks can be characterized by fraud scenarios, which essentially describe how the fraudster gained the illegitimate access to the network. Detection methodologies designed for one specific scenario are likely to miss plenty of the others. For example, velocity trap and overlapping calls detection methodologies are solely aimed at detecting cloned instances of mobile phones and do not catch any of the subscription fraud cases. As stated in the Introduction, the nature of fraud has changed from cloning fraud to subscription fraud, which makes specialized detection methodologies obsolete. Instead, we focus on the detection methodologies based on the calling activity (a stream of transactions), which in turn can be roughly divided into two categories. In *absolute analysis*, detection is based on the calling activity models of fraudulent behavior and normal behavior. In *differential analysis*, on the contrary, the hypothesis is that fraud is characterized by a sudden change in behavior. Using differential analysis, methods typically model behavior in a longer time period and a shorter one and compare these models with each other with a proper distance measure. When current behavior differs from longer-term behavior, alarm is raised. In both cases, the analysis methods are usually realized by using probabilistic models, neural networks or rule-based systems.

Davis and Goyal (1993) report on the use of a knowledge-based approach to analyze call records delivered from cellular switches in real time. In addition to knowledge about the general fraudulent behavior, user profiles are used to distinguish likely fraudulent activity in users' accounts on an individual basis.

In (Barson et al. 1995; Barson et al. 1996), the authors report their first experiments of detecting fraud in a simulated database of calls. They use a supervised feed-forward neural network to detect anomalous use. Six different user types are simulated stochastically according to their calling patterns. Two types of features are derived from this data, one set describing the recent use and the other set describing the longer-term behavior. They report the total error, but give no estimates of false-alarm probabilities or detection probabilities.

Burge and Shawe-Taylor (1996, 1997) focus on unsupervised learning techniques in computing user profiles over sequences of call records. They apply their adaptive prototyping methods in creating models of recent and long-term behavior and calculate a distance measure between the two profiles. A large change in behavior is reported as an alarm. Moreau, Verrelst, and Vandewalle (1997) report work on fraud detection based on supervised feed-forward neural network techniques. They train their neural network on features on fraudulent and normal user data and use the neural network as classifier. The joint research effort of two groups is reported in (Moreau et al. 1996) and is culminated in a hybrid detection system (Burge, Shawe-Taylor, Moreau, Verrelst, Störmann, and Gosset 1997), which combines adaptive prototyping, neural networks and rule-based systems. They report the performance of each component in the system, but not the overall performance of the hybrid system.

A geographical point of view on fraud is promoted by some authors (Yuhas 1993; Shortland and Scarfe 1994; Connor et al. 1995; Cox et al. 1997; Field and Hobson 1997; Samfat and Molva 1997). Indeed, the fraudsters tend to make calls to various destinations, especially to distant countries. Yuhas (1993) clusters geography-based feature data with hierarchical clustering in order to model fraud with feature prototypes. Connor et al. (1995) apply neural networks to detect calling card fraud by correlating fraud with geographic quantities such as distances and entropy measures of calling card patterns. A neural network is used to prioritize a list of suspicious users. In similar fashion, Field and Hobson (1997) present a neural network based fraud management technique based on profiling techniques. Another geography-based method is presented in Cox et al. (1997), who discuss human pattern recognition capabilities in fraud detection by visualizing domain-specific information for interpretation by a domain expert. The calling activity of a user may be browsed and is represented as circles whose size is proportional to the call volumes. The operator is able to look for unusual calling activity by examining the subscribed-specific behavior. Shortland and Scarfe (1994) also present a method to visualize the telephone connections for human interpretation. Samfat and Molva (1997) present an intrusion detection architecture for mobile networks. Their approach is based on combining models of calling behavior and migration patterns.

Fawcett and Provost (1996, 1997) present rule-based methods for fraud detection. The authors use adaptive rule sets to uncover indicators of fraudulent behavior from a database of cellular calls. These indicators are used to create profiles, which then serve as features to a system that combines evidence from multiple profilers to generate alarms. They use rule selection to select a set of rules that span larger sets of fraudulent cases. Furthermore, these rules are used to formulate monitors, which are in turn pruned by feature selection methodology. The output of these monitors is weighted together by a learning unit, which is a linear threshold unit (Fawcett and Provost 1996) and in later their work a neural network (Fawcett and Provost 1997). They assess the results with a cost model in which misclassification cost is proportional to time. In (Provost and Fawcett 1997), the authors present a ROC analysis in the case of non-uniform class and cost distributions. In (Fawcett and Provost 1997), the authors hint at the use of Hidden Markov Models in fraud detection, but doubt its usefulness in fraud domain.

Fraud and uncollectible debt detection with Bayesian networks has been presented in (Ezawa 1995; Ezawa, Singh, and Norton 1996; Ezawa and Norton 1996). The authors use a Bayesian network as a normative expert system. They focus on the unbalanced ratio of fraudsters to non-fraudsters and the unequal misclassification costs. They present a goal-directed algorithm for constructing Bayesian networks for predicting uncollectible debt in telecommunication risk-management datasets.

Our recent research on fraud detection in mobile communications networks is reported in (Taniguchi, Haft, Hollmén, and Tresp 1998; Hollmén and Tresp 1999; Hollmén, Tresp, and Simula 1999). In (Taniguchi, Haft, Hollmén, and Tresp 1998), supervised feed-forward neural networks, unsupervised density estimation, and Bayesian networks are used. Feed-forward neural networks are used to classify the users based on summary statistics over an observation period. The density estimation is used in the novelty detection fashion, in which a Gaussian mixture density is used to model the recent behavior of subscribers and is adapted to track slowly changing behavior (Hollmén 1997). Deviations from the model are alarmed as fraud. In the Bayesian network approach, two Bayesian networks are built by the expert, one describing the subscriber behavior and one describing the fraudulent behavior. These are combined with a Bayes's rule to give a probability of fraud given calling data. Hollmén and Tresp (1999) present a hierarchical regime-switching model for fraud detection. Dynamic modeling of calling behavior is achieved through hierarchical layers of variables, which obey Markov transitions in time. In (Hollmén, Tresp, and Simula 1999), clustering with the Self-Organizing Map is presented, where the cluster models are formulated as probabilistic models.

Comparisons between the approaches are difficult to make, since there are no common basis for evaluation. The work of Fawcett and Provost (1997) is certainly most sophisticated, as far as the cost issues are concerned, but the cost issues still remain an open research issue. Meaningful evaluation must be based on ROC curves or cost considerations. Further divisions of work may be done according to the type of models used. Rule-based approach to fraud detection allows a domain expert to formulate the knowledge using expertise. The model is fully understandable, and any alarm provided by such a system may be understood by a set of rules that triggered the alarm. However, as noted by Sternberg and Reynolds (1997), management of a complex ruled base becomes a difficult task. They attempt to solve the maintainability problem of the rule-based system by introducing a cultural algorithm, which modifies the rule base as needed. They also present a case study around a fraud detection system. In changing environments, such as in fraud detection, this may become an important issue. Learning systems, in turn, once devised for the task, may adapt to new environments. This is our motivation of using learning systems in development of a fraud detection system.

1.3 Related Areas

1.3.1 Credit Card Fraud Detection

There are two problems of credit card fraud detection, namely credit scoring and the real-time credit card fraud detection problem. In credit scoring, the credit worthiness of a customer is evaluated once, whereas the real-time credit card fraud detection is closely related to the problem of fraud detection in communications networks, since the data is a series of transactions in time and problem is to detect any fraudulent use of a credit card as soon as it occurs. Rosenberg and Gleit (1994) present a survey of quantitative methods in a broader context of credit management.

Kauderer and Nakhaeizadeh (1997) consider different input variable transformations for supervised learning and present a survey in credit scoring. Haimowitz and Schwarz (1997) present a framework for credit customer optimization based on clustering and prediction. Customers are first clustered using past credit performance data and thereafter, census data are used to predict credit performance for each cluster.

Hanagandi et al. (1996) use radial basis function networks to create credit card scores from historical credit card transactions. Aleskerov et al. (1997) present a neural network based database mining system for credit card detection and test it on synthetically generated data. Ghosh and Reilly (1994) present a case study in credit card fraud detection with neural networks. Dorronsoro et al. (1997) present an operational system for fraud detection of credit card operations based on a neural classifier.

Stolfo et al. (1997) present a meta-learning approach in credit card fraud detection to combine results from multiple classifiers. Chan and Stolfo (1998) present a meta-learning approach to scalable learning with non-uniform class and cost distribution in credit card fraud detection. Bax (1998) presents validation of voting committees, with application to credit scoring.

1.3.2 Intrusion Detection on Computer Systems

Intrusion detection methods attempt to find unauthorized use of computer systems. Dixon (1991) outlines the history of computer related fraud, describes the automated auditing methods in use and reports on a survey on perceptions on the risks of fraud. He concludes that computer fraud is a serious problem, but is not taken seriously enough. Surveys of intrusion detection methodologies can be found in (Lunt 1988; Lunt 1993; Frank 1994; Mukherjee, Heberlein, and Levitt 1994; Kumar 1995).

In (Lee, Stolfo, and Chan 1997; Lee, Stolfo, and Mok 1998), two approaches to intrusion detection are presented. Firstly, they detect against known scenarios of intrusion and secondly, they detect anomalous deviations from normal behavior. This division is similar to absolute and differential analysis discussed earlier. The audit trails of computer systems are used to characterize normal behavior patterns and to distinguish it from the intruders.

Neural networks have been widely used to learn patterns of usage for intrusion detection (Fox, Henning, Reed, and Simonian 1990). Tan (1995) presents an application of neural networks to computer security by learning behavioral patterns. Ryan et al. (1997) present intrusion detection with neural networks by learning how frequently commands are used.

Denning (1987) presents a model of a real-time intrusion detection expert system capable of detecting break-ins, penetrations and other forms of computer abuse. Garner and Chen (1994a, 1994b) present hypothesis generation based anomaly detection.

Goldberg and Senator (1997) discuss break detection systems, breaks being indication of instances where some violation of proper conduct has occurred. Break detection systems attempt to detect violations in which actors use their own resources to conduct fraudulent activity. They further state that fraud detection differs from fault detection where the normal operation is well defined and where the anomalies are readily apparent.

Lunt (1988, 1990, 1993) presents the IDES system for detecting intruders in computer systems. White, Fisch, and Pooch (1996) present intrusion detection based on peer-based, co-operating security managers. Lane and Brodley (1997) present matching functions to compare current behavioral sequence to a historical profile to be used in intrusion detection. Crowder (1997) discusses fraud detection techniques and computer assisted fraud detection.

1.3.3 Other Work on Fraud Detection

There are numerous fields where one is interested to find anomalous or illegitimate behavior based on the transactions we can measure about the process. Similar work may be found in diverse fields, such as in insurance industry, health care, finance, and management.

Glasgow (1997) discusses the risk in the insurance industry and divides it to two parts: risk as an essential element of the related underwriting task and the fraud risk. In health care fraud detection, knowledge-based systems are used in health care fraud detection (Sokol 1998) and in detection and preinvestigation of health care fraud (Major and Riedinger 1992). He, Wang, Graco, and Hawkins (1997) present medical fraud detection by grouping practice profiles of medical doctors to normal and abnormal profiles with the aid of neural networks.

An assessment of AI technologies for detection of money laundering can be found in (Jensen 1997). Schuerman (1997) discusses risk management in the financial industry, and Barney (1995) about closely related trading fraud. Allen et al. (1996) transform financial transaction data to be visualized for further inspection by a domain expert.

Menkus (1998) defines management directed fraud as any fraud committed by senior executives or external auditors. Curet, Jackson, and Tarar (1996) discuss detection of top management fraud and the development, implementation and evaluation of a case-based learning and reasoning tool. Fanning, Cogger, and Srivastava (1995) use neural networks in detecting management fraud.

In addition to previously mentioned work, fraud detection methods may be found in various application areas such as automatic toll collection on motor ways (Zimmermann and Neumayer 1995) or detecting copied or cloned parts of a software (Barson et al. 1995). Ramani et al. (1997) discuss the application of several neuro-fuzzy paradigms in check authorization from incomplete information.

Chapter 2

Materials

2.1 Call Data

In this thesis, fraud detection is based on calling activity of mobile phone subscribers. This calling activity is recorded for the purpose of billing in call records, which store attributes of calls, like the identity of the subscriber (IMSI), time of the call, duration of the call to mention a few. In the context of GSM networks, the standard about administration of subscriber related event and call data in a digital cellular telecommunications system can be found in (European Telecommunications Standards Institute 1998). In order to develop learning systems to discriminate between fraud and normal behavior, data representing both kinds of behavior is necessary. However, a data collection procedure may be costly and subject to restrictions due to legislation on privacy of data. We now describe two ways of collecting fraud data for development of a fraud detection system. The first approach is based on a customer complaint or uncollected fees, the second on a scan in the call database.

After each billing period, telephone bills are created as reports from the subscriber specific call data using appropriate tariffs (pricing) for each service. If a fraudster exploited an account during the billing period, it will become evident to the victimized subscriber at the time of the billing, or in the case of subscription fraud when the operator realizes the fraudster has no intention to pay the bill. Fawcett and Provost (1997) describe the process of block crediting, where a representative of the carrier and the defrauded customer examine the telephone bill call by call and label the data to fraud and non-fraud segments. Such labeling is naturally expensive, they also admit that such a process is likely to contain errors. As an advantage, data is segmented to fraud and non-fraud segments more precisely, which enables using more efficient learning mechanisms.

It would be beneficial if fraud detection system could be designed using data from a fraudulent account without extensive labeling work and human intervention. Fraud data used in this thesis were filtered from a database of call data using a velocity trap detection mechanism. Velocity trap alarms if calls are made from locations geographically far apart in temporal proximity. In essence, this sets a limit on the velocity a mobile phone subscriber may travel, hence the name. No geographical information about the calls were made available in the call data nor when the velocity trap gave an alarm. An important consequence of this is to understand that the data does not contain information on which calls were fraudulent or which periods contained fraudulent activity. Rather, fraudulent behavior is superimposed on the normal calling activity. Data labeled as fraudulent is a sample from a *mixture* of normal and fraudulent data, the mixing coefficients being unknown and changing in time. The nature of the data is evidently reflected by the developed methods for learning to detect fraud. The database of fraudulent behavior contained call data of 304 subscribers. Each call data spanned a period of 92 days. The data describing normal behavior was extracted from a large database of calls spanning 49 days and was assumed to contain no fraudulent activity. The use of data may vary in the publications, consult individual publications for details.

2.2 Representation of Call Data

As described earlier, the behavior of a mobile phone subscriber must be represented in a way that fits the modeling paradigm used. Calls are transactions in time and are not suitable as such for modeling approaches such as neural networks or probabilistic models. Therefore, a more suitable representation must be derived by appropriate pre-processing, feature extraction and selection steps. In feature extraction, variables believed to summarize the essential are calculated from raw data and in feature selection, a relevant subset of possible variables is chosen.

Two representations of data are used in this thesis. In the first representation, the detection is based on feature variables derived from call data during a given time period, typically one day (Taniguchi, Haft, Hollmén, and Tresp 1998). Such a mapping transforms the transaction data ordered in time to static variables residing in feature space. The features used typically reflect the usage of an account. We have used number of calls and summed length of calls to describe the daily usage of a mobile phone. National and international calls were regarded as different categories, also calls made during business hours, evening hours and night hours were separated to sub categories. In the second approach, we describe the calling data as a time-series of calling activity. In (Hollmén and Tresp 1999), the calling data was represented as a time-series of zeros and ones indicating whether a mobile phone was used during a particular minute. For example, a call for three minutes would be represented as . . . 0011100 . . . . This representation enables the dynamic modeling of the calling behavior expressed with transitions from one time step to another. This representation was also the basis of modeling in (Hollmén, Tresp, and Simula 1999).

Chapter 3

Methods

3.1 Probabilistic Networks

Probabilistic networks allow an efficient description of multivariate probability densities and can also be used as probabilistic expert systems (Cowell et al. 1999). Probabilistic formulations allow uncertainty both in the formulation of the solution as well as in the statements made about the problem. This makes the framework of probabilistic networks appealing for real-world problems. Of particular interest here are the Bayesian networks (Pearl 1988; Jensen 1996), which can be represented as directed acyclic graphs (DAG). A Bayesian network may be thought of as a graph $\mathcal{G} = (V, E)$, where V is the set of vertices or nodes and E is the set of edges or links, which is defined as an ordered set of vertices $E \subset V \times V$. The nodes of the graph correspond to the domain variables and an edge to the qualitative dependency between two variables (See Figure 3.1).

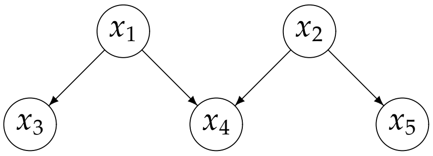

Figure 3.1: A simple Bayesian network is shown. Variables are marked with graph nodes, the dependency relationships as edges. The joint probability density can be factorized as $P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2)P(x_3|x_1)P(x_4|x_2, x_1)P(x_5|x_2)$.

Graphical representation makes it easy to understand and manipulate networks. The term graphical model refers to this dual representation of probabilistic models as graphs. In the next sections, we review the concept of conditional independence, which is used in defining qualitative relationships between the variables, and distributional assumptions, which in turn define the quantitative aspect of the probabilistic networks. Learning from data is then briefly described within the framework of maximum likelihood using the EM algorithm (Dempster, Laird, and Rubin 1977).

3.1.1 Conditional Independence

A problem domain consists of a set of random variables. Random variable is an unknown quantity that can take on one of a set of mutually exclusive and exhaustive outcomes (Cowell et al. 1999). The joint probability density $P(x_1, \dots, x_n)$ of the random variables $x_1, \dots, x_n$ can be decomposed according to the chain rule of probability as

$$P(x_1, \dots, x_n) = \prod_{i=1}^{n} P(x_i | x_{i-1}, \dots, x_1).$$

Each term in this factorization is a probability of a variable given all lower numbered variables. In real life, however, not all factors influence the others in a given domain, thus this kind of qualitative knowledge can be formulated by assuming conditional independence of the form

$$P(x_3 | x_2, x_1) = P(x_3 | x_1).$$

This is to say that for all values of x_1 knowing about x_2 does not bring any additional information about x_3. This leads to an alternative formulation of the form

$$P(x_3, x_2 | x_1) = P(x_3 | x_2, x_1) P(x_2 | x_1) = P(x_3 | x_1) P(x_2 | x_1).$$

In both examples, x_3 and x_2 are conditionally independent given x_1. The use of conditional independence assumptions allow us to construct global joint distribution from a set of local conditional probability distributions. Defining $\pi_i \subseteq \{x_1, \dots, x_{i-1}\}$ as the parent set of x_i or the set of variables that renders x_i and $\{x_1, \dots, x_{i-1}\}$ conditionally independent, we can rewrite the joint probability density as

$$P(x_1, \dots, x_n) = \prod_{i=1}^{n} P(x_i | \pi_i)$$

A Bayesian network defines this joint probability density as the product of local, conditional densities. The main contribution of the conditional independence assumptions is that the expression is far simpler as the trivial decomposition achieved by the application of chain rule of probability.

According to the conditional independence assumptions, we can identify some interesting model classes. In finite mixture models (Everitt and Hand 1981; Redner and Walker 1984; Titterington et al. 1985), we assume an observed variable Y that is conditioned on a discrete hidden variable S. The observed variable may be either discrete or continuous. The joint probability density is then

$$P(S,Y) = P(S)P(Y|S).$$

Integration (summation) over the hidden variable S gives us an equation for calculating the likelihood of observed data. In this model, there are no assumptions of conditional independence. This is shown in Figure 3.2.

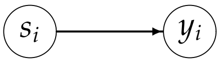

Figure 3.2: A mixture model is shown. The observed variable y_i is conditioned on a discrete hidden variable s_i. Observed samples are assumed to be independent.

A more complicated model that takes time dependencies into account is the Hidden Markov Model (HMM), which is widely used in sequence processing and speech recognition (Baum 1972; Juang and Rabiner 1991; Bengio 1999). Smyth et al. (1997) consider HMMs in a general framework of probabilistic independence networks and show that algorithms for inference and learning are special cases of more general class of algorithms. For a review on HMM, see (Levinson et al. 1983; Poritz 1988). These models assume a discrete, hidden state s_t, observations y_t that are conditioned on the hidden state as $P(y_t|s_t)$ and the state transitions as $P(s_t|s_{t-1})$. The joint probability density is then

$$P(Y,S) = P(y_0,s_0)\prod_{t=1}^{T} P(s_t|s_{t-1};\theta_1)\prod_{t=1}^{T} P(y_t|s_t;\theta_2)$$

in which the current state is conditionally independent of the whole history given the previous state $P(s_t|s_{t-1},s_{t-2},\ldots,s_1) = P(s_t|s_{t-1})$. This is called the Markov property, which is prevalent in many kinds of time-series models. Moreover, the current observation is conditionally independent of the whole history given the current hidden state. In essence, the state information summarizes the whole history. The graphical presentation of the HMM is shown in Figure 3.3. In (Hollmén and Tresp 1999),

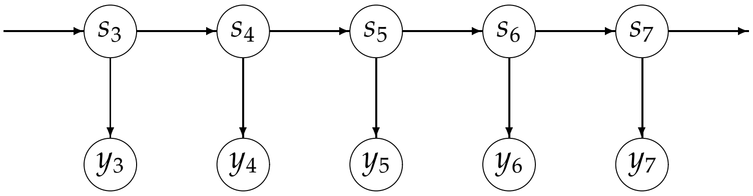

Figure 3.3: In a Hidden Markov model, we assume hidden variables s_t that obey transitions in time defined by $P(s_t|s_{t-1})$, the observations are conditioned on the hidden variable as $P(y_t|s_t)$.

a more complicated structure was used, which differs from HMM in two aspects. First, the hidden variable that develops in time has a hierarchical structure and second, the probability density for the observations is dependent on past observations. The hierarchical organization involves two layers of states, each of which develops in time according to a Markov chain and the upper layer is conditioned of the layer below. In all, the joint probability for observations and the hidden states (V in the upper layer and S in the layer below, see Figure 3.4) is

$$P(Y,S,V) = \\ P(y_0,s_0,v_0)\prod_{t=1}^{T}P(v_t|v_{t-1};\theta_1)\prod_{t=1}^{T}P(s_t|v_t,s_{t-1};\theta_2)\prod_{t=1}^{T}P(y_t|s_t,y_{t-1};\theta_3)$$

The idea in regime-switching models is to model a problem domain with multiple models allowing the generating mechanism to switch from one mode of operation to another in an indeterministic fashion (Quandt 1958; Quandt and Ramsey 1972; Shumway and Stoffer 1991; Hamilton 1990; Hamilton 1994).

3.1.2 Distributional Assumptions

Conditional independence assertions provide qualitative assumptions between variables in the probabilistic network. To further quantify these established relationships, we need to define for every variable in the network the conditional probability distribution of the variable given its parents. Variables may be discrete or continuous, but variables with children may only be discrete.

If the observations in the finite mixture model are distributed according to a Gaussian (normal) distribution, it is called the Gaussian Mixture

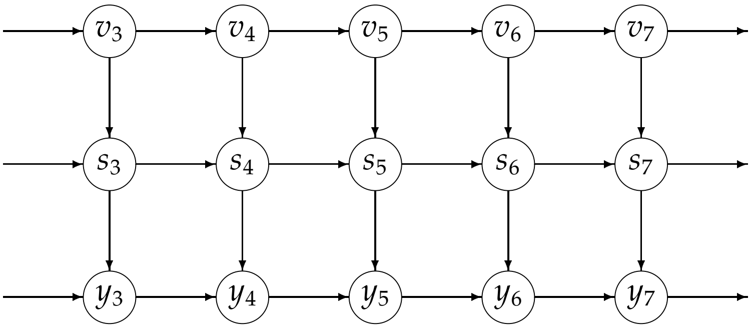

Figure 3.4: Hierarchical regime-switching model is shown. Hidden variables v and s have a hierarchical structure and upper layer is conditioned on the lower layer. Furthermore, observations y_t are conditioned on a previous observation y_{t-1} and the current state s_t.

Model (Redner and Walker 1984; Bishop 1996). In our work, this kind of model was used in (Taniguchi, Haft, Hollmén, and Tresp 1998) to model the probability density of call data to be used in novelty detection to detect changes in behavioral patterns. Discrete states in the models described above are best modeled with the assumption of multinomial distributions, in which the variable can be in one of many states of the variable.

3.1.3 Learning by EM Algorithm

Learning is the process of estimating the parameters of a model from available set of data. In the context of probabilistic models, it is natural to consider the principle of maximum likelihood, according to which the maximum likelihood estimate for our parameter maximizes the probability of our data. This is relatively straightforward if the variables in our model are observed, but becomes somewhat complicated, since the models of interest here have hidden variables. This problem may be overcome by the application of the EM algorithm. The EM algorithm (Dempster, Laird, and Rubin 1977; McLahlan 1996) is an iterative algorithm for estimating maximum likelihood parameters in incomplete-data problems. Incomplete data means that there is a many-to-one mapping between the hidden state space and the observed measurements. Since it is impossible to recover the hidden variable, EM algorithm works with its expectation instead by making use of the measurements and the implied form of the mapping in the model. EM algorithm is guaranteed to increase to likelihood after each iteration, the parameter value it converges to is indeed the maximum likelihood estimate (Dempster, Laird, and Rubin 1977; Wu 1983; Xu and Jordan 1996).

For the purpose of the EM algorithm, we introduce the expected log likelihood of the complete data (Dempster, Laird, and Rubin 1977) as

$$\begin{aligned} Q(\phi|\phi^{(old)}) &= E(logP(Y, S|\phi)|Y, \phi^{(old)}) \\ &= \int_S logP(Y, S|\phi)P(S|Y, \phi^{(old)})ds \end{aligned}$$

where the log-likelihood of the complete data is parameterized by the free parameter value ϕ and the expectation is taken with respect to the second distribution parameterized by the current parameters $\phi^{(old)}$. In the E-step, the Q function is computed. In Bayesian networks, this is achieved through inserting observed evidence in the network and applying propagation rules (Jensen 1996) to form the joint probability distribution of all variables or any marginalization of it. The first account that used inference techniques in the E-step appeared in (Lauritzen 1995). In M-step, we update the parameter values to be

$$\phi^{(new)} = \arg\max_{\phi} Q(\phi|\phi^{(old)})$$

Solution to this maximization problem is usually found by setting the derivatives of the maximized function to zero and solving for ϕ. The application of the EM algorithm in the case of mixture models can be found in the literature (Redner and Walker 1984; Bishop 1996), interestingly the learning technique used in HMM (Baum 1972) turns out to be an instance of the EM algorithm. Learning in regime-switching models within the framework of maximum likelihood was formulated by Hamilton (1990, 1994). He used a regime-switching model to identify recession periods in the US economy. In our work (Hollmén and Tresp 1999), exact inference rules for the hierarchical regime-switching model are derived from the junction tree algorithm of the Bayesian networks (Jensen 1996). A recent account on learning from data with graphical models can be found in (Heckerman 1999).

3.2 Self-Organizing Map (SOM)

3.2.1 SOM Algorithm

The Self-Organizing Map (SOM) is a neural network model for the analysis and visualization of high-dimensional data. It was invented by professor Kohonen (1990, 1995) and is the most popular network model based on unsupervised, competitive learning. It has also been successfully applied for the analysis of industrial processes (Kohonen, Oja, Simula, Visa, and Kangas 1996; Alhoniemi, Hollmén, Simula, and Vesanto 1999). Bibliography on published papers may be found in (Kaski et al. 1998).

The Self-Organizing Map is a collection of prototype vectors, between which a neighborhood relation is defined. This neighborhood relation defines a structured lattice, usually a two-dimensional, rectangular or hexagonal lattice of map units. Training a Self-Organizing Map from data is divided to two steps, which are applied alternately. First, a best-matching unit (BMU) or a winner unit m^c is searched, which minimizes the Euclidean distance between a data sample x and the map units m^k

$$c = \arg\min_k \|x - m^k\|.$$

Then, the map units are updated in the *topological* neighborhood of the winner unit. The topological neighborhood is defined in terms of the lattice structure, not according to the distances between data samples and map units. The update step can be performed by applying

$$m^k(t+1) := m^k(t) + \alpha(t)h^c(t,k)[x(t) - m^k(t)]$$

where the last term in the square brackets is proportional to the gradient of the squared Euclidean distance $d(x, m^k) = \|x - m^k\|^2$. The learning rate $\alpha(t) \in [0, 1]$ must be a decreasing function of time and the neighborhood function $h^c(t, k)$ is non-increasing function around the winner unit defined in the topological lattice of map units. A good candidate is a Gaussian around the winner unit defined in terms of the coordinates r in the lattice of neurons

$$h^c(t,k) = \exp\left(-\frac{\|r^k - r^c\|^2}{2\sigma(t)^2}\right).$$

During learning, the learning rate and the width of the neighborhood function are decreased, typically in a linear fashion. The map then tends to converge to a stationary distribution, which approximates the probability density of data.

The Self-Organizing Map may be visualized by using a unified distance matrix representation (Ultsch and Siemon 1990), where the clustering of the SOM is visualized by calculating distances between the map units locally and representing these visually with gray levels. Another choice for visualization is the Sammon's mapping (Sammon Jr. 1969), which projects the high-dimensional map units on a plane by minimizing the global distortion of inter point distances when applying the mapping.

3.2.2 SOM for Clustering Probabilistic Models

Hollmén, Tresp, and Simula (1999) presented a Self-Organizing Map algorithm, which enables using probabilistic models as the cluster models. In this approach the map unit indexed by k stores the empirically estimated parameter vector θ^k with an associated probabilistic model $q(x\ ;\theta^k)$. For implementing a Self-Organizing Map algorithm, we need to define a distance between the map units (i.e. the θ^k) and data. The distance between θ and a data point itself cannot be defined in a Euclidean space since they may have different dimensionality. The most common distance measure between probability distributions is the Kullback-Leibler distance (Bishop 1996; Ripley 1996), which relates two probability distributions. Let us think of a data point x_i as distributed according to an unknown probability distribution $x_i \sim p(x)$ and then approximate $p(x) \approx \delta(x_i)$. If we substitute this expression to the Kullback-Leibler distance, we get

$$KL(p \parallel q) = -\int p(x) \log \frac{q(x\ ;\theta^k)}{p(x)} dx = -\log q(x_i\ ;\theta^k) \qquad (3.1)$$

which is the negative log probability of data for our empirical model. Thus, minimizing the Kullback-Leibler distance between the unknown true distribution that generated the data point at hand and our empirical model leads to minimizing the negative logarithm of the probability of the data with our empirical model. This justifies the use of this probability measure as a distance measure between models and data. In light of this derivation, we can derive a Self-Organizing Map algorithm for parametric probabilistic models. Winner unit indexed by c is defined by minimizing the negative log-likelihood of the empirical models for a given data point or equivalently, by searching for the maximum likelihood unit as in

$$c = \arg\min_k [-\log q(x_i\ ;\theta^k)] = \arg\max_k q(x_i\ ;\theta^k).$$

The update rules are based on the gradients of this likelihood in the topological neighborhood of the winner unit c as

$$\theta^k(t+1) := \theta^k(t) + \alpha(t) h^c(t,k) \frac{\partial \log q(x(t);\theta^k)}{\partial \theta^k}.$$

To illustrate the idea, we derived an algorithm for a specific case of user profiling in mobile phone networks (Hollmén, Tresp, and Simula 1999). However, the approach presented in generally applicable to user profiling problems often encountered in marketing, for example.

Chapter 4

Assessment

4.1 Receiver Operating Characteristic Curves

In the fundamental detection problem (Green and Swets 1966; Egan 1975), the task of the observer is to decide on the basis of an uncertain evidence whether the stimulus consisted of a signal embedded in noise or noise alone. Observations are either *accepted* as signals in noise or *rejected* as noise alone according to a decision rule. Rephrasing this terminology from the field of psychophysics, we have a detection system (or a classifier) that on the basis of measurements, in our case call data, decides whether the calling behavior is normal or fraudulent. In fraud detection domain, we are interested how accurately these statements can be made. The evaluation must be made for each class separately, since by classifying all the cases trivially as normal a small (misleading) error rate would be achieved. This is based on the observation that fraud is indeed rare and normal behavior is dominating. Also, incorrect classifications may have different consequences. In such domains, it is natural to consider class specific assessment of the detection capability, which leads to Receiver Operating Characteristic analysis (Green and Swets 1966; Egan 1975; Metz 1978; Swets 1988).

$$ROC = \{(u,v)|u = \int_r^\infty p(x|\omega_1)dx\ ; v = \int_r^\infty p(x|\omega_2)dx\}$$

Receiver Operating Characteristic (ROC) curve is a function that summarizes the possible performances of a detector. It does so by varying the cut-off point of decision (threshold) along the chosen decision variable. It can be presented as a graphical plot, where the probability of detection is plotted as a function of the probability of false alarm. Formally, a ROC curve is a curve of points (u, v), where $p(x|\omega_1)$ and $p(x|\omega_2)$ are the probability densities for the decision variable. This is shown in the Equation above. In our applications, the decision variable is based on the likelihood ratio or the posterior class probabilities. ROC visualizes the trade-off between false alarms and detection, thus facilitating the choice of a decision function.

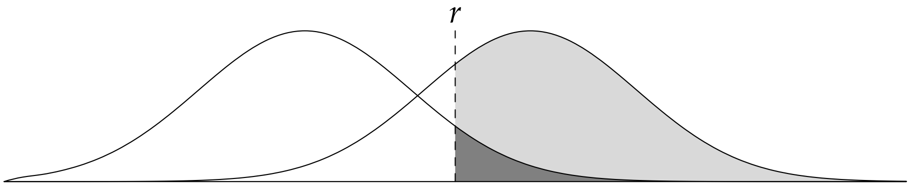

Figure 4.1: The class densities for the decision variable are shown. The dashed vertical line at r is the cut-off point for decision. Probability of detection is marked with light gray and the probability of false alarm with dark gray. ROC curve visualizes the effect of r on these probabilities.

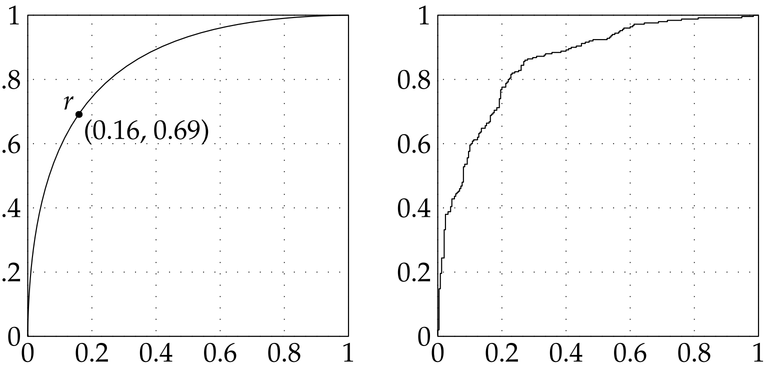

Figure 4.2: In the left panel, ROC curve for the distributions in Figure 4.1 is shown. The cut-off point r is marked in the figure, corresponding to a false alarm probability of 0.16 and a detection probability of 0.69. In the right panel, empirically estimated ROC curve for the same distributions is shown. Samples from each class (n = 250) were generated from the Gaussian distributions $(\mu_1 = 0, \mu_2 = 1.5, \sigma_1 = \sigma_2 = 1)$.

Hanley and McNeil (1982) show that the area under the ROC curve corresponds to the probability that a randomly chosen subject from class ω_2 is correctly rated with greater suspicion than a randomly chosen subject form ω_1. Hilgers (1991) presents a method to estimate the distribution-free confidence bounds of ROC curves for finite samples.

4.2 Cost Issues

In the previous section, the accuracy of a detection system was assessed with ROC curves. Presenting a ROC curve as such is ignorant of the cost issues, but recognizes the importance of class-specific evaluation. The final goal in fraud detection is to minimize costs incurred through fraud. The decision goal determines the decision rule to be used, for example, the goal of minimum probability of misclassification leads to the maximum posterior classification rule and the goal of minimum cost of misclassification leads to minimizing the expected risk of decision (Duda and Hart 1973; Schalkoff 1992). In fraud domain, formulating a cost model presents an area for further work.

Ezawa and Norton (1996) state that the cost issues are handled surprisingly little in the literature. Fawcett and Provost (1997) present cost models in fraud domain. Based on their fraud estimate, they state a fixed cost for every minute of fraudulent activity. They give cost estimates with different decision schemes with their rule-based system and compare them with decision schemes such as "classify all as fraudulent" and "classify all as normal". Other work also discuss cost issues (Pazzani et al. 1994; Provost and Fawcett 1997).

Chapter 5

Summary

5.1 Contents of the Publications

Publication 1 (Taniguchi, Haft, Hollmén, and Tresp 1998) presents three methods for fraud detection. Firstly, a feed-forward neural network is used in classification of users to normal and fraudulent users based on summary statistics over a time period. Secondly, user behavior is modeled with an adaptive Gaussian Mixture model, which is used in a novelty detection fashion to detect sudden changes from the past behavior. This constitutes the contribution of the present author. Thirdly, two Bayesian networks are formulated by expert to reflect domain knowledge about fraud and normal behavior, the outputs from these network are combined together with a Bayes's rule. The two latter methods are based on features calculated over a period of one day. For the methods presented in this paper, a patent (Taniguchi, Haft, Hollmén, and Tresp 1997) has been granted.

Publication 2 (Hollmén and Tresp 1999) uses a hierarchical regime-switching model in detection of fraud. Learning is based on the EM algorithm; inference rules are derived from the junction tree algorithm (Jensen 1996). In addition to unsupervised learning, the models are fine-tuned with supervised learning to improve the discriminative performance of the model. The calling data is represented as a truth valued time-series, which has a high-sampling rate. This work is a step towards real-time detection of fraud. The learning procedure does not require fully labeled accounts, but works with partially labeled data as described in Chapter 2.

Publication 3 (Hollmén, Tresp, and Simula 1999) develops methods to cluster probabilistic models with the Self-Organizing Map algorithm. The standard Self-Organizing Map algorithm is not suitable for the task, since it uses Euclidean distance as an error measure, which cannot sensibly be defined between time-series and probabilistic models. On the contrary, parameters of probabilistic models are stored in map units and a likelihood based distance measure is defined between data and map units. Update equations are derived from the gradients of likelihood; additional parameterization is introduced to handle the constraints on the parameters. A softmax layer is used to map the unconstrained parameters to the constrained parameter space. In experiments, the approach is used to model calling behavior in mobile communications networks with dynamic models.

5.2 Contributions of the Author

In Publication 1, the author of this thesis was responsible for the section on novelty detection with Gaussian Mixtures. The ideas were invented by the author and experiments were made by the author. Also, writing the paper was coordinated by the author. In Publication 2, the author was responsible for the representation of the problem and the experiments. The inference rules were developed jointly with the second author, with whom the paper was also jointly written. In Publication 3, the author was responsible for ideas and the experiments. The paper was written by the author and edited by the co-authors.

5.3 Summary and Conclusions

In this thesis, the problem of fraud detection was treated as a pattern recognition problem. The calling data of mobile phone subscribers constituted the basis of behavior description. It was hypothesized that fraud can either be detected through abrupt changes in calling behavior or by designating models to each type of behavior. These are called differential and absolute approaches to fraud detection, respectively. The models were realized using probabilistic models or neural networks. The data used in learning originated from real mobile communications networks.

The use of probabilistic models to model uncertainty in fraud domain is further motivated by the view by Davis and Goyal (1993), who state that there is nothing about any one call itself that proves incontrovertibly that is it fraudulent. Therefore, any call must be put in a wider context of behavior. In our models, this context is carried forward with the inclusion of time-dependent hidden variables, which put any single call to a context of near history. In the novelty detection approach, we considered feature variables, which summarize the behavior of one day, similarly forming a contextual description of behavior.

The methods presented are based on the concept of learning or forming a general model of a phenomenon from a given set of data samples. As discussed in Chapter 2, collecting accurately labeled fraud data is expensive. The methods developed in this thesis learn to detect fraud from partially labeled data, that is, it is known that an account is defrauded but not exactly when. The data is thus a mixture of normal and fraudulent data with an unknown mixing mechanism. To our knowledge, no other work solves the problem of learning fraud models from data that is partially labeled in this fashion. This approach provides an economic aspect to learning to detect fraud.

The results in this thesis in terms of detection performance are comparable to or better than other works published in the field. As a rough measure of state-of-the-art performance, the detection system should detect most of the fraudsters, but more importantly, false alarm probabilities should be below 2 or 3 percent. Otherwise, as the population of mobile phone users may be large, the absolute amount of alarms is beyond control. As noted in the review part, comprehensive comparisons are difficult to make since there are no common basis for evaluation. In any case, an evaluation should be based on ROC analysis and possibly cost evaluation, if a cost model can readily be formulated. This is one of the issues for further work. However, formulation of the cost model may be complex, since many factors affect the actual loss through fraud. Another dimension of further work is the use of direct discriminative methods to fraud detection. Also, transforming the problem of unsupervised learning to a problem, where supervised learning could be used, would be beneficial.

The research presented in this thesis addresses an important problem of user profiling. More than ever before, companies need to be aware of the existing customer segments, habits, and lifestyles in order to provide suitable products and services to the public. In this endeavor, user profiling helps in defining customized products and services through quantitative analysis of market data. In the era of popularized Internet, vast masses of people have access to a great variety of product offerings. The profound difference to marketing in the past is that instead of companies pushing the products to the customer, the customer himself is seeking for a suitable product or a service. This sets higher requirements on the suitability of a service to a customer. Therefore, companies need to formulate target groups and analyze the market to guarantee satisfaction of the target group of customers. The analysis should have consequences on conceptual design of future products. This cycle results ideally to improved products and services.

Taking this wider view on user profiling, fraud detection may be seen as a specific case of user profiling, where the purpose of the model is to distinguish intentional misuse of mobile phones from legitimate calling behavior.

References

Aleskerov, E., B. Freisleben, and B. Rao (1997). CARDWATCH: A neural network based database mining system for credit card fraud detection. In *Proceedings of the IEEE/IAFE 1997 Conference on Computational Intelligence for Financial Engineering (CIFEr)*, pp. 220–226. IEEE Press.

Alhoniemi, E., J. Hollmén, O. Simula, and J. Vesanto (1999). Process monitoring and modeling using the self-organizing map. *Integrated Computer Aided Engineering 6*(1), 3–14.

Allen, P., R. McKendrick, C. Scott, M. Buonanno, P. Mostacci, C. Naldini, V. Scuderi, and P. Stofella (1996, April). Interactive anomaly detection in large transaction history databases. In *High-Performance Computing and Networking. International Conference and Exhibition HPCN 1996 Proceedings*, pp. 143–149.

Barney, L. (1995). Detecting trading fraud. *Wall Street & Technology 12*(11), 40.

Barson, P., N. Davey, S. Field, R. Frank, and D. S. W. Tansley (1995). Dynamic competitive learning applied to the clone detection problem. In J. Alspector, R. Goodman, and T. X. Brown (Eds.), *Proc. Int. Workshop on Applications of Neural Networks to Telecommunications 2*, Hillsdale, NJ, pp. 234–241. Lawrence Erlbaum.

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

Baum, L. E. (1972). An inequality and associated maximation technique in statistical estimation for probabilistic functions of markov processes. *Inequalities 3*, 1–8.

Bax, E. (1998, May). Validation of voting committees. *Neural Computation 10*(4), 975–986.

Bengio, Y. (1999). Markovian models for sequential data. *Neural Computing Surveys 2*, 129–162.

Bishop, C. (1996). *Neural Networks in Pattern Recognition*. Oxford Press.

Burge, P. and J. Shawe-Taylor (1996). Frameworks for fraud detection in mobile telecommunications networks. In *Proceedings of the Fourth Annual Mobile and Personal Communications Seminar, University of Limerick*.

Burge, P. and J. Shawe-Taylor (1997, July). Detecting cellular fraud using adaptive prototypes. In *Proceedings of AAAI-97 Workshop on AI Approaches to Fraud Detection & Risk Management*, pp. 9–13. AAAI Press.

Burge, P., J. Shawe-Taylor, Y. Moreau, H. Verrelst, C. Störmann, and P. Gosset (1997, October). BRUTUS - a hybrid detection tool. In *Proceedings of ACTS Mobile Telecommunications Summit, Aalborg, Denmark*.

Chan, P. K. and S. J. Stolfo (1998, August). Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. In R. Agrawal, P. Stolorz, and G. Piatetsky-Shapiro (Eds.), *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98)*, pp. 164–168.

Connor, J. T., L. R. Brothers, and J. Alspector (1995). Neural network detection of fraudulent calling card patterns. In *Proceedings of the International Workshop on Applications of Neural Networks to Telecommunications, 2*, pp. 363–370. Laurence Erlbaum Associates.

Cowell, R., A. Dawid, S. Lauritzen, and D. Spiegelhalter (1999). *Probabilistic Networks and Expert Systems*. Statistics for Engineering and Information Science. Springer-Verlag.

Cox, K. C., S. G. Eick, G. J. Wills, and R. J. Brachman (1997). Visual data mining: recognizing telephone calling fraud. *Data mining and Knowledge Discovery 1*(2), 225–231.

Crowder, N. (1997, April). Fraud detection techniques. *Internal auditor 54*(2), 17–20.

Curet, O., M. Jackson, and A. Tarar (1996, October). Designing and evaluating a case-based learning and reasoning agent in unstructured decision making. In *1996 IEEE International Conference on Systems, Man and Cybernetics. Information Intelligence and Systems.*, Volume 4, pp. 2487–2492.

Davis, A. B. and S. K. Goyal (1993). Management of cellular fraud: Knowledge-based detection, classification and prevention. In *Proceedings of the 13th International Conference on Artificial Intelligence,* *Expert Systems and Natural Language, Avignon, France*, Volume 2, pp. 155–164.

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

Denning, D. E. (1987, February). An intrusion-detection model. *IEEE Transactions on Software Engineering SE-13*, 222–232.

Dixon, R. (1991). Audit developments and the detection of computer fraud. *International Journal of Computer Applications in Technology 4*(4), 207–216.

Dorronsoro, J. R., F. Ginel, C. Sánchez, and C. S. Cruz (1997). Neural fraud detection in credit card operations. *IEEE Transactions on Neural Networks 8*(4), 827–834.

Duda, R. O. and P. E. Hart (1973). *Pattern Recognition and Scene Analysis*. John Wiley & Sons.

Egan, J. (1975). *Signal Detection Theory and ROC Analysis*. New York: Academic Press.

European Telecommunications Standards Institute (1998, February). Digital cellular telecommunications system (Phase 2); Event and call data (GSM 12.05 version 4.3.1). European Telecommunication Standard ETS 300 616.

Everitt, B. and D. Hand (1981). *Finite Mixture Distributions*. Monographs on Applied Probability and Statistics. Chapman and Hall.

Ezawa, K., M. Singh, and S. Norton (1996). Learning goal oriented bayesian networks for telecommunications risk management. In *Proceedings of the 13th International Conference on Machine Learning*, pp. 139–147. Morgan Kaufmann.

Ezawa, K. J. (1995). Fraud/uncollectible debt detection using a bayesian network based learning system: A rare binary outcome with mixed data structures. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pp. 157–166. Morgan Kaufmann.

Ezawa, K. J. and S. W. Norton (1996, October). Constructing bayesian networks to predict uncollectible telecommunications accounts. *IEEE Expert 11*(5), 45–51.

Fanning, K., K. O. Cogger, and R. Srivastava (1995). Detection of management fraud: a neural network approach. *International Journal of*

Intelligent Systems in Accounting, Finance and Management 4(2), 113–126.

Fawcett, T. and F. Provost (1996, July). Combining data mining and machine learning for effective user profiling. In E. Simoudis, J. Han, and U. Fayyad (Eds.), *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp. 8–13. AAAI Press.

Fawcett, T. and F. Provost (1997). Adaptive fraud detection. *Journal of Data Mining and Knowledge Discovery 1*(3), 291–316.

Field, S. and P. Hobson (1997). Techniques for telecommunications fraud management. In J. Alspector, R. Goodman, and T. X. Brown (Eds.), *Proc. Int. Workshop on Applications of Neural Networks to Telecommunications 3*, Hillsdale, NJ, pp. 107–115. Lawrence Erlbaum.

Fox, K. L., R. R. Henning, J. H. Reed, and R. P. Simonian (1990). A neural network approach towards intrusion detection. In *Proc. 13th National Computer Security Conference. Information Systems Security. Standards - the Key to the Future*, Volume I, Gaithersburg, MD, pp. 125–134. NIST.

Frank, J. (1994, October). Artififcial intelligence and intrusion detection: Current and future directions. In *National Computer Security Conference*, Volume 1, pp. 22–33.

Garner, B. and F. Chen (1994a, May/June). Anomaly detection modeling. In *Industrial and Engineering Applications of Artificial Intelligence and Expert Systems. Proceedings of the Seventh International Conference*, pp. 509–514.

Garner, B. and F. Chen (1994b, August). Hypothesis generation paradigm for fraud detection. In *Proceedings of the 1994 IEEE Region 10's Ninth Annual International Conference*, Volume 1, pp. 197–201. IEEE Press.

Ghosh, S. and D. L. Reilly (1994, January). Credit card fraud detection with a neural network. In *Proc. of the Twenty-Seventh Hawaii Int. Conf. on System Sciences*, pp. 621–630. IEEE Computer Society Press.

Glasgow, B. (1997, July). Risk and fraud in the insurance industry. In *Proceedings of AAAI-97 Workshop on AI Approaches to Fraud Detection & Risk Management*, pp. 20–21. AAAI Press.

Goldberg, H. G. and T. E. Senator (1997, July). Break detection systems. In *Proceedings of AAAI-97 Workshop on AI Approaches to Fraud Detection & Risk Management*, pp. 22–28. AAAI Press.

Green, D. and J. Swets (1966). *Signal Detection Theory and Psychophysics*. Wiley, New York.

Haimowitz, I. J. and H. Schwarz (1997, July). Clustering and prediction for credit line optimization. In *Proceedings of AAAI-97 Workshop on AI Approaches to Fraud Detection & Risk Management*, pp. 29–33. AAAI Press.

Hamilton, J. D. (1990). Analysis of time series subject to changes in regime. *Journal of Econometrics 45*, 39–70.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

Hanagandi, V., A. Dhar, and K. Buescher (1996, March). Density-based clustering and radial basis function modeling to generate credit card fraud scores. In *Proceedings of the IEEE/IAFE 1996 Conference on Computational Intelligence fo Financial Engineering (CIFEr)*, pp. 247–251. IEEE Press.

Hanley, J. A. and B. J. McNeil (1982, April). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology 143*(1), 29–36.

He, H., J. Wang, W. Graco, and S. Hawkins (1997). Application of neural networks to detection of medical fraud. *Expert Systems with Applications 13*(4), 329–336.

Heckerman, D. (1999). A tutorial on learning with bayesian networks. In M. I. Jordan (Ed.), *Learning in Graphical Models*, pp. 301–354. MIT Press.

Hilgers, R. (1991). Distribution-free confidence bounds for ROC curves. *Methods of Information on Medicine 30*(2), 96–101.

Hoath, P. (1998, January). Telecoms fraud, the gory details. *Computer Fraud & Security 20*(1), 10–14.

Hollmén, J. (1997, October). Novelty filter for fraud detection in mobile communications networks. Technical Report A48, Helsinki University of Technology, Laboratory of Computer and Information Science.

Hollmén, J. and V. Tresp (1999). Call-based fraud detection in mobile communications networks using a hierarchical regime-switching model. In M. Kearns, S. Solla, and D. Cone (Eds.), *Advances in Neural Information Processing Systems: Proceedings of the 1998 Conference (NIPS'11)*, pp. 889–895. MIT Press.

Hollmén, J., V. Tresp, and O. Simula (1999, September). A self-organizing map algorithm for clustering probabilistic models. In *Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN'99)*, Volume 2, pp. 946–951. IEE.

Jensen, D. (1997, July). Prospective assessment of AI technologies for fraud detection: A case study. In *Proceedings of AAAI-97 Workshop on AI Approaches to Fraud Detection & Risk Management*, pp. 34–38. AAAI Press.

Jensen, F. V. (1996). *An Introduction to Bayesian Networks*. UCL Press.

Johnson, M. (1996, December). Cause and effect of telecoms fraud. *Telecommunication (International Edition) 30*(12), 80–84.

Juang, B. and L. Rabiner (1991, August). Hidden markov models for speech recognition. *Technometrics 33*(3), 251–272.

Kaski, S., J. Kangas, and T. Kohonen (1998). Bibliography of self-organizing map (SOM) papers: 1981-1997. *Neural Computing Surveys 1*, 102–350.

Kauderer, H. and G. Nakhaeizadeh (1997, July). The effect of alternate scaling approaches on the performances of different supervised learning algorithms. an empirical study in the case of credit scoring. In *Proceedings of AAAI-97 Workshop on AI Approaches to Fraud Detection & Risk Management*, pp. 39–42. AAAI Press.

Kohonen, T. (1990, September). The self-organizing map. *Proceedings of the IEEE 78*(9), 1464–1480.

Kohonen, T. (1995). *Self-Organizing Maps*. Springer-Verlag.

Kohonen, T., E. Oja, O. Simula, A. Visa, and J. Kangas (1996). Engineering applications of the self-organizing map. *Proceedings of the IEEE 84*(10), 1358–84.

Kumar, S. (1995, August). *Classification and detection of computer intrusions*. Ph. D. thesis, Purdue University.

Lane, T. and C. E. Brodley (1997, July). Sequence matching and learning in anomaly detection for computer security. In *Proceedings of AAAI-97 Workshop on AI Approaches to Fraud Detection & Risk Management*, pp. 43–49. AAAI Press.

Lauritzen, S. L. (1995). EM algorithm for graphical association models with missing data. *Computational Statistics & Data Analysis 19*, 191–201.

Lee, W., S. J. Stolfo, and P. K. Chan (1997). Learning patterns from UNIX process execution traces for intrusion detection. In *Proceedings of AAAI-97 Workshop on AI Approaches to Fraud Detection & Risk Management*, pp. 50–56. AAAI Press.

Lee, W., S. J. Stolfo, and K. W. Mok (1998, August). Mining audit data to build intrusion detection models. In R. Agrawal, P. Stolorz, and G. Piatetsky-Shapiro (Eds.), *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98)*, pp. 66–72.

Levinson, S., L. Rabiner, and M. Sondhi (1983, April). An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition. *The Bell System Technical Journal 62*(4), 1035–1074.

Lunt, T. (1988, October). Automated audit trail analysis and intrusion detection: A survey. In *Proceedings of the 11th National Computer Security Conference*, pp. 65–73.

Lunt, T. F. (1990, November). IDES: An intelligent system for detecting intruders. In *Proceedings of the Symposium on Computer Security (CS'90), Rome, Italy*, pp. 110–121.

Lunt, T. F. (1993). A survey of intrusion detection techniques. *Computers & Security 12*(4), 405–418.

Major, J. A. and D. R. Riedinger (1992). EFD: A hybrid knowledge/statistical based system for the detection of fraud. *International Journal of Intelligent Systems 7*(7), 687–703.

McLahlan, G. J. (1996). *The EM Algorithm and Extensions*. Wiley & Sons.

Menkus, B. (1998, April). Some management-directed fraud incidents. *EDPACS 25*(10), 14–16.

Metz, C. E. (1978, October). Basic principles of ROC analysis. *Seminars in Nuclear Medicine VIII*(4), 283–298.

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

Moreau, Y., H. Verrelst, and J. Vandewalle (1997, October). Detection of mobile phone fraud using supervised neural networks: A first prototype. In *International Conference on Artificial Neural Networks Proceedings (ICANN'97)*, pp. 1065–1070.

Mukherjee, B., L. T. Heberlein, and K. N. Levitt (1994, May/June). Network intrusion detection. *IEEE Network 8*(3), 26–41.

O'Shea, D. (1997, January). Beating the bugs: Telecom fraud. *Telephony 232*(3), 24.

Parker, T. (1996, November). The twists and turns of fraud. *Telephony 231*(supplement issue), 18–21.

Pazzani, M., C. Merz, P. Murphy, K. Ali, T. Hume, and C. Brunk (1994). Reducing the misclassification costs. In *Proceedings of the 11th International Conference on Machine Learning*, pp. 217–225. Morgan Kaufmann.

Pearl, J. (1988). *Probabilistic reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pequeno, K. A. (1997, May). Real-time fraud detection: Telecom's next big step. *Telecommunications (Americas Edition) 31*(5), 59–60.

Poritz, A. B. (1988). Hidden markov models: A guided tour. In *Proceedings of the IEEE International conference of Acoustics, Speech and Signal Processing (ICASSP'88)*, pp. 7–13.

Provost, F. and T. Fawcett (1997, July). Analysis and visualization of classifier performance with nonuniform class and cost distributions. In *Proceedings of AAAI-97 Workshop on AI Approaches to Fraud Detection & Risk Management*. AAAI Press.

Quandt, R. (1958, December). The estimation of parameters of linear regression system obeying two separate regimes. *J. Am. Stat.Assoc. 53*, 873–880.

Quandt, R. and J. Ramsey (1972, June). A new approach to estimating switching regression. *Journal of American Statistical Society 67*(338), 306–310.

Ramani, V., J. Echuaz, G. Vachtsevanos, and S. Kim (1997, July). Neurofuzzy approaches to decision-making: An application to check authorization from incomplete information. In *Proceedings of AAAI-97 Workshop on AI Approaches to Fraud Detection & Risk Management*, pp. 64–71. AAAI Press.

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

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press.

Rosenberg, E. and A. Gleit (1994, July–August). Quantitative methods in credit management: A survey. *Operations Reserach 42*(4), 589–613.

Ryan, J., M.-J. Ling, and R. Miikkulainen (1997, July). Intrusion detection with neural networks. In *Proceedings of AAAI-97 Workshop on AI Approaches to Fraud Detection & Risk Management*, pp. 72–77. AAAI Press.

Samfat, D. and R. Molva (1997, September). IDAMN: An intrusion detection architecture for mobile networks. *IEEE Journal on Selected Areas in Communications 15*, 1373–1380.

Sammon Jr., J. W. (1969, May). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers C-18*(5), 401–409.

Schalkoff, R. J. (1992). *Pattern Recognition: Statistical, Structural and Neural approahes*. John Wiley & Sons.

Schuerman, T. (1997, July). Risk management in the financial services industry: Through a statistical lens. In *Proceedings of AAAI-97 Workshop on AI Approaches to Fraud Detection & Risk Management*, pp. 78–82. AAAI Press.

Shortland, R. and R. Scarfe (1994, October). Data mining applications in BT. *BT Technology Journal 12*(4), 17–22.

Shumway, R. and D. Stoffer (1991, Spetember). Dynamic linear models with switching. *Journal of the American Statistical Association 86*(415), 763–769.

Smyth, P., D. Heckerman, and M. I. Jordan (1997, February). Probabilistic independence networks for hidden markov probability models. *Neural Computation 9*(2), 227–269.

Sokol, L. (1998, March). Using data mining to support health care fraud detection. In *PADD98. Procedings of the Second International Conference on the Practical Application of Knowledge Discovey and Data Mining*, pp. 75–82.

Sternberg, M. and R. Reynolds (1997). Using cultural algorithms to support re-engineering of the rule-based expert systems in dynamic performance environments: a case study in fraud detection. *IEEE Transactions on Evolutionary Computation 1*(4), 225–243.

Stolfo, S. J., D. W. Fan, W. Lee, and A. L. Prodromidis (1997, July). Credit card fraud detection using meta-learning: Issues and initial results. In *Proceedings of AAAI-97 Workshop on AI Approaches to Fraud Detection & Risk Management*, pp. 83–90. AAAI Press.

Swets, J. A. (1988, June). Measuring the accuracy of diagnostic systems. *Science 240*, 1285–1293.

Tan, K. (1995). The application of neural networks to UNIX computer security. In *1995 IEEE International Conference on Neural Networks*, pp. 476–481. IEEE Press.

Taniguchi, M., M. Haft, J. Hollmén, and V. Tresp (1997, July). Erkennung eines betrügerischen anrufs mittels eines neuronalen netzes. Patent DE 197 29 630 A1.

Taniguchi, M., M. Haft, J. Hollmén, and V. Tresp (1998, May). Fraud detection in communications networks using neural and probabilistic methods. In *Proceedings of the 1998 IEEE International Conference in Acoustics, Speech and Signal Processing (ICASSP'98)*, Volume 2, pp. 1241–1244.

Titterington, D., A. Smith, and U. Makov (1985). *Statistical analysis of finite mixture distributions*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons.

Ultsch, A. and H. Siemon (1990). Kohonen's self-organizing maps for exploratory data analysis. In *Proceedings of the International Neural network Conference (INNC'90)*, pp. 305–308. Kluwer.

White, G. B., E. A. Fisch, and U. W. Pooch (1996, January/February). Cooperating security managers: A peer-based intrusion detection system. *IEEE Network 10*(1), 20–23.

Wu, C. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics 11*(1), 95–103.

Xu, L. and M. Jordan (1996). On convergence properties for EM algorithm for gaussian mixtures. *Neural Computation 8*, 129–151.

Yuhas, B. (1993). Toll-fraud detection. In *Proceedings of the International Workshop on Applications of Neural Networks to Telecommunications (IWANNT'93)*, pp. 239–244.

Zimmermann, R. and B. Neumayer (1995). Telematics in traffic: new motorway technologies on trial. *NT 48*(8), 26–29. (in German).