Emergent self-organising feature maps used for prediction and prevention of churn in mobile phone markets

Received (in revised form): 5th September, 2001

Alfred Ultsch

studied computer science at the Technical University of Munich. He holds a Master of Science degree from the Technical University of Munich (Diplom) and a Master of Science from Purdue University, USA. Professor Ultsch did his PhD at ETH Zurich, Switzerland in the field of artificial intelligence. His work at the University of Dortmund was on self-organising neural networks. He is the developer of the widely-known U-matrix, a visualisation tool for high-dimensional dates–space on self-organising neural networks. Since 1993 he has been full professor at Philipps-University of Marburg and head of the NeuroInformatics and Artificial Intelligence research group. His research areas are knowledge discovery and data mining with neural networks.

Abstract Typical approaches to customer relationship management (CRM) construct a black box model for the prediction of churning. The approach presented here aims at new insights into the phenomena of the particular customer-business relationship. The application of a combination of emergent self-organising maps, U-matrix methods and knowledge conversion technique to mobile phone customer data is reported. The aim of this approach is to discover knowledge to be used for the prevention of churn. The output of the approach are rules that lead to a better understanding of who clients are, what profitable customers are and why churning is happening. This knowledge leads to direct business actions for the redesign of products, services and marketing activities to prevent churning. Without prior knowledge it will be possible to identify products bought by persons who will remain as customers only for a short time and thus become unprofitable customers.

INTRODUCTION

Customer relationship management (CRM) means to know and understand the interaction between customers and a business.[1] CRM is a central issue in many businesses. Data mining can be defined as the process of discovering new, understandable and useful knowledge in data sets.[2] Applying data mining to data sources created by the customer/business interaction might therefore be effectively used to acquire valuable knowledge about a customer. Of particular interest is knowledge about which customers will quit and what their motives for quitting are.

Emergent self-organising maps (SOM) together with U-matrix methods (hereafter called Allview technology) allow discovery of structures in large high-dimensional data sets.[3] These structures are formed by an emergent process and represent a higher level of structuring the data set. They provide a non-linear mapping superior to classical clustering algorithms.[4] The chances are

Alfred Ultsch
Philipps-University of Marburg, Department of Computer Science, Hans-Meerwein-Str., 35032 Marburg, Germany.

Tel: +49 6421 282 2185; Fax: +49 6421 282 8902; e-mail: ultsch@Mathematik.Uni-Marburg.de

good that Allview will show new, formerly unknown, structures in the data set.

The structures discovered by Allview might be new, but in general they are not immediately useful for CRM. The structures by themselves do not represent knowledge. Knowledge is understood to be statements about data sets that are understandable by humans, that are also interpretable by a knowledge-based system and, in particular, have meaning for the business, ie knowledge must lead to a non-trivial understanding of important features of the data generating process. Using the knowledge conversion algorithm sig* knowledge in this sense can be extracted from the structures in emergent SOMs.[5]

This paper reports on the application of the Allview knowledge discovery method to data from a Swiss mobile phone company.[6] The main focus was to find knowledge which can be used to predict, and ultimately prevent, customers from discontinuing contracts with mobile phone companies.

CHURNING

Churn means discontinuation of a contract with a business.[7] In many European countries wireless telecommunication markets have changed from being monopolistic to very competitive. To be able to control and reduce the customer churn rate may be a vital survival factor for mobile phone companies.[8] In particular it would be necessary to prevent the churn of profitable customers. If it is possible to predict whether a customer is likely to churn in two months, given current customers' records, appropriate action by the business may prevent the loss of the customer. The acquisition of a new customer in the telecommunication market usually involves a substantial investment. In order to be profitable a new customer should buy services from the company over a long period. The profitability of a business–customer relationship is therefore directly proportional to the period of time a customer buys services from the business. Central issues for churn management are:

— what types of customers are likely to quit
— are there subgroups of potential churners that are so promising that further efforts to keep them are worthwhile?

To know why a profitable customer decides to discontinue a contract is essential better to tailor products to profitable market segments. Such knowledge is also essential for the marketing strategies to acquire new profitable customers. Business objectives for churn predictions are to concentrate marketing activities on those customers that are likely to produce profit by remaining with a company for longer periods of time and to retain profitable customers.

CONVENTIONAL CHURN PREDICTION

A typical approach to the problem of churn prediction is as follows: a sufficiently large data set containing churning and non-churning customers is used to construct a classifier. This classifier is an algorithm that is able to decide, given a customer data set, whether the customer is likely to churn or not. These classifiers may be constructed using, for example, artificial neuronal networks like multilayer perceptrons (backpropagation) or Bayesian statistics. Many data mining programs use decision trees constructed with heuristics like CART or C4.5.[9]

The quality of the output of the classifier is measured in terms of sensitivity, specificity and accuracy.[10] The sensitivity of a classifier to predict a class c is defined as the number of data for which a correct prediction is given divided by the number of all members of class c. The specificity to predict a class c is defined as the number of data sets that are correctly classified to be not in c divided by the number of data not belonging to c. The best possible classifier has a sensitivity and specificity of 100 per cent. In many classifiers there is a trade-off between sensitivity and specificity. An increase in sensitivity will be typically accompanied by a decrease in specificity. A display of sensitivity *vs.* specificity is called the receiver operating characteristic (ROC) curve.[11] Besides sensitivity and specificity the overall accuracy is also used to describe the quality of a prediction. Accuracy is the percentage of correct predictions. These quality measures are typically used to adjust parameters of a classifier in order to find an optimal classifier.[12]

PROBLEMS OF CONVENTIONAL CHURN PREDICTION

In a typical churn prediction approach a classifier is constructed and the measurements mentioned above are made. Typically, the first results of a classifier are not sufficient and either the parameters of the classifier are changed or another classifier model is used. This process is repeated until the quality of the prediction achieved by the classifier is sufficient. If an accuracy of around 90 per cent is reached for a given classifier, the ability of the constructed classifier to predict churn might be considered sufficient.[13]

A big problem is, however, to find out why a customer wants to churn. If it is worthwhile keeping a certain customer it is necessary to decide what business actions should be taken in order to prevent churning (CRM). For this a hypothesis about the reasons for the discontinuation of a contract is necessary. Such knowledge is, however, hardly provided by the constructed classifier.

So while the constructed classifier might be useful for churn prediction it is not very helpful for churn prevention. In the following the construction of a classifier is described using knowledge extracted from emergent SOM that overcomes these problems.

MOBILE PHONE CUSTOMER DATA

The data used for this study consisted of a sample of 300,000 customer records. The data were provided by Swisscom AG.[14] The time span of the data was June, 1999 to February, 2000. Twenty-one variables concerning usage of the Swisscom mobile telecommunication network were used. The variables described the following aspects of customer behaviour:

— money spent, for example, accounting dates for the contract
— usage of services, like SMS
— usage of different networks
— destinations of long-distance calls
— usage times, when the calls took place, for example daytime, night etc.

ESOM AND U-MATRIX FOR THE IDENTIFICATION OF CUSTOMER GROUPS

In this section a brief explanation of ESOM and U-matrix technology is given. Self-organising maps (SOM) belong to a general class of neural network methods, which are non-linear regression techniques that can be applied to find relationships between inputs and

outputs or to organise data so as to disclose previously unknown patterns or structures. This approach has been demonstrated to be highly relevant to many financial, economic and marketing applications.[15] SOM, developed by Teuvo Kohonen in 1982, exhibit the interesting and non-trivial ability of emergence through self-organisation.
'Self-organisation' means the ability of a system to adapt its internal structure to structures sensed in the input of the system. This adaptation should be performed in such a way that first, no intervention from the environment is necessary (unsupervised learning) and secondly, the internal structure of the self-organising system represents features of the input-data that are relevant to the system.

Emergence means the ability of a system to produce a phenomenon on a new, higher level. In physics this change of level is termed 'mode' or 'phase-change'. It is produced by the cooperation of many elementary processes. Emergence happens in natural, technical and human systems. The formation of cumulus cloud streets and lasers are examples of natural and technical emergence. Emergent phenomena can also happen in crowds of human beings. An example is the Mexican wave in sports stadiums. Participating human beings function as the elementary processes when they produce a large wave by rising from their places and throwing their arms up in the air. This wave can be observed on a macroscopic scale and could be described in terms of wavelength, velocity and repetition rate. For emergence to occur it is absolutely necessary that a huge number of elementary processes cooperate. A new, higher-level phenomenon can only be observed when elementary processes are disregarded and only structures formed by the cooperation of many elementary processes are considered. In typical applications of SOM the number of nodes in the maps are too few to show emergence. In typical applications a single node may be regarded as a cluster, ie all data, whose best matches fall on this node, are members of this cluster. This type of application performs clustering in a way that is similar to statistical clustering algorithms like, for example, k-means or single-linkage. Emergence can only be expected to happen with a large number of nodes. Such maps, called emergent self-organising maps (ESOM), have typically at least thousands (if not tens of thousands) of nodes. In particular, the number of nodes may be much bigger than the number of data points in the input data. Consequently, most of the nodes of ESOMs will represent few input points if at all. Clusters are detected on ESOM not by regarding single nodes but by regarding the overall structure of the whole map. The latter can be done using U-matrix methods.

The simplest of these is to sum up the distances between the node weights and those of its immediate neuron neighbours. This sum of the distances to its neighbours is displayed as elevations at the position of each neuron. The elevation values of all nodes produce a three-dimensional landscape, the U-matrix. U-matrices have the following properties:

— neighbouring data in the high-dimensional input data space lie in a common valley
— gaps in the distribution of input points produce hills on the U-matrix
— elevations or hills are proportional to the gap distances in the input-space.

The principal properties of ESOM in conserving the overall topology of the

input space, is inherited by a U-matrix. Data closest in the input-space can also be found at neighbouring places on the U-matrix. Topological relations between the clusters are also represented on the two dimensional layout of the nodes.

With U-matrix methods emergence in ESOM can be observed. The cluster-structure of the input data set is detected in the U-matrix as valleys surrounded by hills with more or less elevation, ie clusters are detected, for example, by raising a virtual water level up to a point, where the water floods a valley on the U-matrix. The user can grasp the high-dimensional structure of the data: nodes that lie in a common valley are subsumed to a cluster; regions of a feature map that have high elevations in a U-matrix are not identified with a cluster; nodes that lie in a valley but are not best matches are interpolations of the input data. This approach has been extensively tested on many different applications. It can be shown that this method gives a good picture of the high-dimensional and otherwise invisible structure of the data. In many applications meanings for clusters could be detected.[16] ESOM can be easily used to construct classifiers. If the U-matrix has been separated into valleys corresponding to clusters and hills corresponding to gaps in the data, then an input data point can be easily classified by looking at the best match of this data point. If the point's best match lies inside a cluster-region on the U-matrix the input data are in that cluster. If the best match lies on a hill in the U-matrix, no classification of this point can be assigned. This is particularly so if the data set possesses new features, ie aspects that were not included in the data learned so far. With this approach, for example, outliers and errors in the data are easily detected.

For the telecommunication data an ESOM of dimension 128×128 was constructed. With U-matrix technologies 47 groups could be found in the data. Figure 1 shows an example of such a U-matrix for customers that are characterised by using only telephone services in Switzerland. As it can be seen seven customer groups can be identified in this class.

KNOWLEDGE CONVERSION

The main aim of the data mining approach discussed here is the discovery of new and useful knowledge in a data set. With the Allview technology described above groups of customers with common mobile phone usage characteristics could be identified. This by itself might be interesting but not so much as a description of what the meaning of the groups is. The algorithm sig* operates on the groups identified by Allview and produces a description of these groups in the form of understandable decision rules.[17] In contrast to other decision rule algorithms, like decision tree inference, the understandability of the knowledge generated is the main aim of this algorithm. With the knowledge conversion algorithm sig* rules for each of these groups could be extracted from the emergent SOMs. Sig* produces a description for each group in the form of characterising and differentiating rules. Such a description for group 13 is given in Figure 2.

A group is described using the most significant variables first. This allows meaning to be given to a group. For example, a group could be characterised as 'customers using SMS primarily at night'. Besides the description of a group in terms of variables sig* returns a measure of how significant the variable for the description of a group is.[18] In the example in Figure 2 the significance

Figure 1 U-matrix of domestic mobile phone users

```
rule__4 : Class  of Customer  is_a '13' if for SwisscomMobileCustomer holds:
          'Var_4'    in [ 0.2, 285.5]          % (Sig=28.9191) and
          'Var_9'    in [ 1, 223]              % (Sig=23.4303) and
          'Var_12'   in [ 0.3, 51.4]           % (Sig=23.1201) and
          'Var_10'   in [ 0, 4]                % (Sig=17.5463) and
          'Var_15'   in [ 0.1, 28.5]           % (Sig=13.1106) and
          'Var_20'   in [ 1, 22]               % (Sig=12.3796) and
          'Var_7'    in [ 3, 11.4]             % (Sig= 9.9734) and
Customer  is_a '13' but_not '23'.

rule__4__1 : SwisscomMobileCustomer is_a   '13', but_not   '23'
 if for Customer holds :
          2 of  [ 'Var_3', geq 0,   'Var_11' geq 6.3, 'Var_17' geq 23.6 ].
```

Figure 2

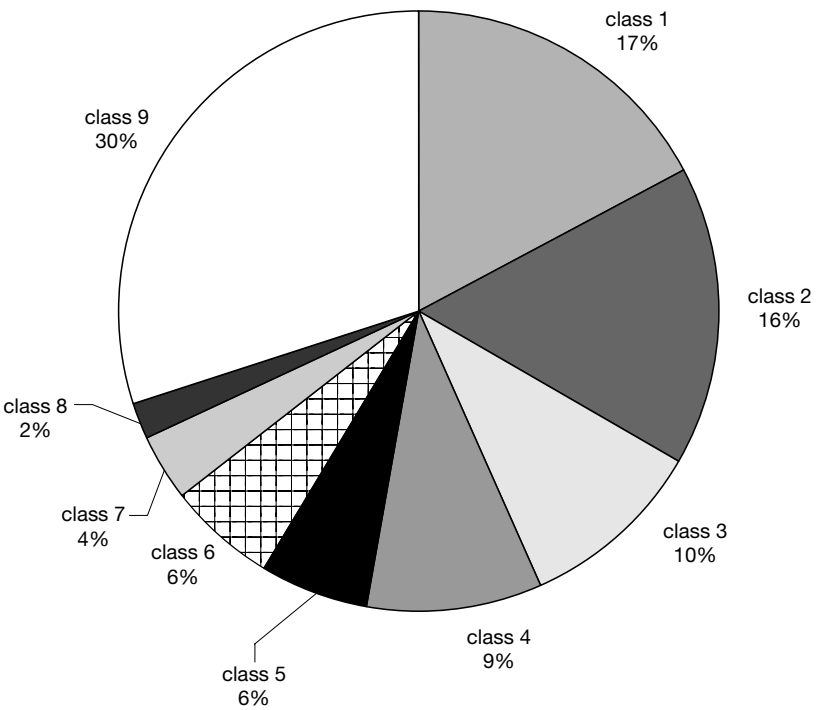

Figure 3 Classes aggregated from 47 groups

value is given at 'Sig = . . .'. The rules for the different groups were analysed and it was found that the same variables were the most important for several groups. So these groups could be aggregated to classes with common features. The 47 groups could be aggregated to nine classes having common usage profiles. The size of the groups is given in Figure 3.

A knowledge-based system can interpret the rules generated by sig*. This results in a classifier for mobile phone customers. A ROC plot of this classifier is given in Figure 4. Sensitivity and specificity of the classifier is close to 100 per cent for all rules.

CHURN PREDICTION

Churners were defined as customers who discontinued at least one contract with Swisscom Mobile in month m. For these customers the data of the second preceding month (m-2) were selected and called churn data. These churn data were also classified using Allview.

With the methods described previously the 47 groups and nine classes detected in the data set could be identified in the churn data. The rules generated by sig* for the churn data turned out to be effective predictors for churning. An overall accuracy for the prediction rules was measured to be 99.8 per cent.

DISCOVERED KNOWLEDGE

The description of the 47 groups and nine classes in the form of rules allows marketers to understand the type of customers who are using the Swisscom mobile phone network. An attempt was made to find a description of these 47 groups using socioeconomic variables. The significance values of such descriptions were too small to allow a

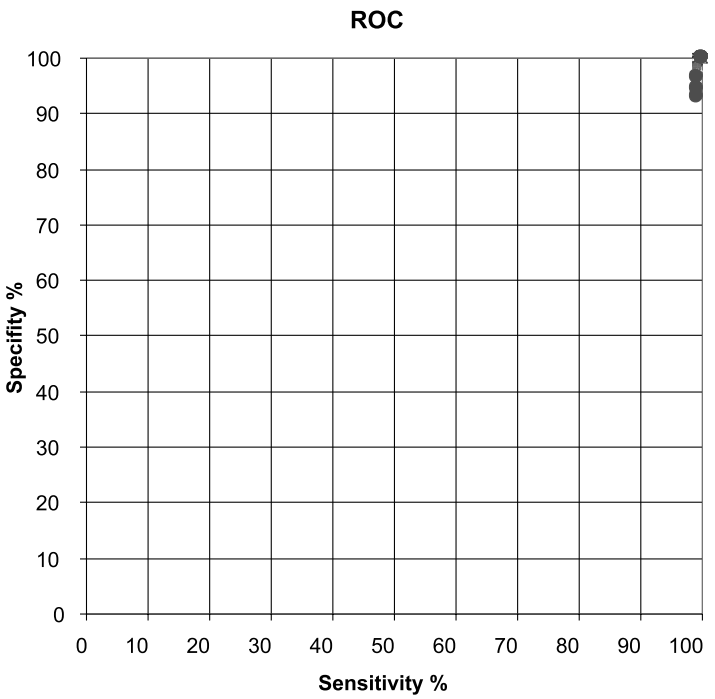

Figure 4 ROC curve of the rule-based classifier for 47 groups

meaningful identification of customer groups. The conclusion might be drawn that customers of Swisscom Mobile are characterised by their usage of the network than by their social or economic background.

A comparison of the distribution of the groups among churners and non-churners reveals, however, interesting properties. Figure 5 shows the distribution of the nine customer classes in the churn data.

The distribution of the classes in the non-churn data may now be compared to the churn data (see Figure 3). Most striking is the increase in size in class 7. The conclusion might be drawn that this type of customer is most likely to churn. Class 8 also more than doubled it size. Classes 1, 2, 4 and 5 showed a significant reduction in size. Customers of these groups might be less interested in churning. Since all groups have a meaningful description in terms of network usage the reasons for a churning decision can be understood.

VALUABLE CUSTOMERS

It is particularly important to prevent the churn of customers who generate substantial revenues. Therefore, the revenues of each of the 47 identified customer groups were calculated. Figure 6 shows the total revenue of the groups in relation to the size of the group.

The zero line in Figure 6 means that the members of the group contribute just as much as expected by the size of the group to the total revenue. The positive and negative numbers are percentages of additional or missing revenue in the group. Relating revenues to the groups allows a very fine-grained identification of valuable customers.

It can be seen that some groups contribute substantially more revenue than would be expected by the size of

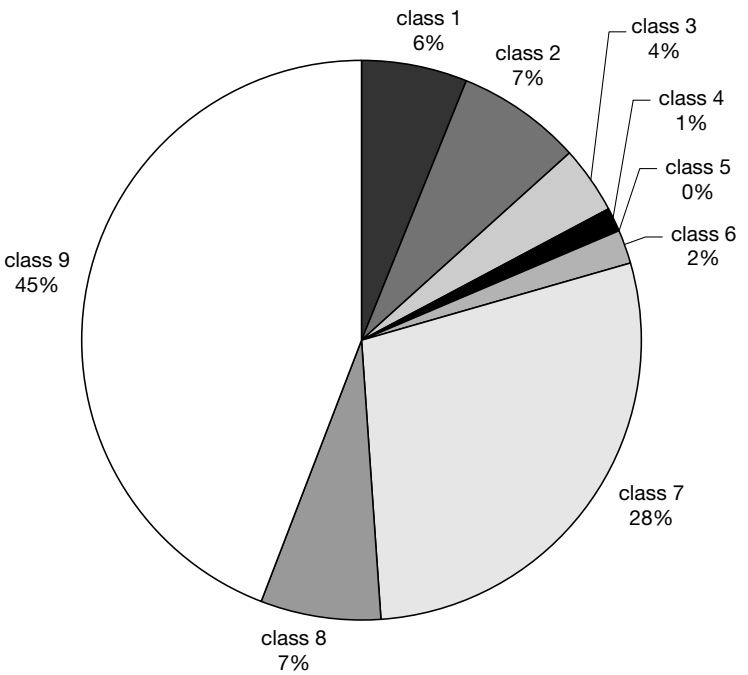

Figure 5 Distribution of classes in churn data

the group. Other groups do not meet expectations. Combined with the knowledge of what the groups mean, ie what services the customers use, this allows efficient customer relationship management.

DISCUSSION

Most approaches to churn prediction produce a black box model that is able to predict churn. Quality criteria for such an approach are numbers such as sensitivity, specificity, ROC curves, accuracy, positive predictive value etc. Standard scientific procedure for such an approach is to clearly separate training and test data sets and measure the quality of the prediction on both of these sets.

The approach presented here aims at a different target: the understanding of the phenomenon of churn in a particular business, ie a characterisation of which customers are likely to quit a contract and what the likely reasons for doing so are. The product of the method presented here using ESOM, U-matrix and knowledge conversion (sig*) is a set of rules that lead to a better characterisation of the customer groups that are likely to churn or not to churn. Furthermore, segments that are more profitable than others can be identified among these groups. The ROC plots given above are therefore only an indication of how well the rules describe the general properties of the detected subgroups in the data. As can be seen, the rules give a rather accurate description of the groups. The quality of the output of the approach can therefore only be measured in terms of the following questions:

— do the rules lead to a better understanding of who the clients are and why churning is happening

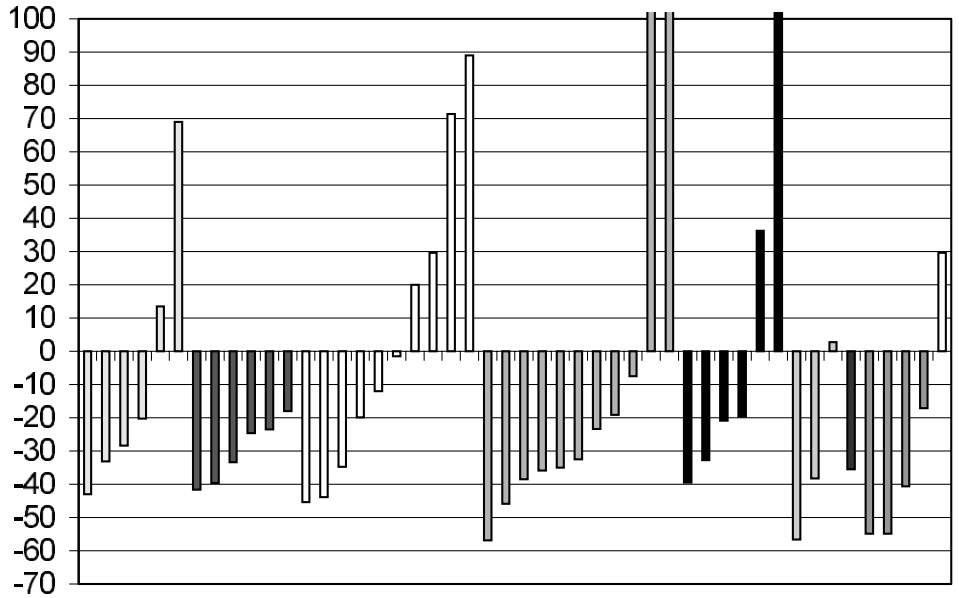

Figure 6 Revenues generated by customer groups

— can the rules be used as indications of what to do in order to prevent churn?

In the case presented here concrete characterisations of a large numbers of churners could be derived by the interpretation of the rules. Furthermore, product groups could be identified that were likely to attract new customers but were not attractive for building up a longer-lasting business–customer relationship. In particular, customers who used (or hardly used) certain services and/or made phone calls from/to certain countries could be identified to be churners. Special offers could be identified that attracted only short-term customers who became, therefore, unprofitable.

The method presented is, therefore, data mining in the sense of producing new (formerly unknown) knowledge about customers out of the data generated by customer–business interaction.

CONCLUSION

ESOM is a non-linear mapping technique that captures structures in high dimensional data not detectable by other clustering algorithms.[19] SOM together with U-matrix visualisation use the phenomena of emergence to detect formerly unknown structures in data sets. In an experimental setting with real-world data it could be shown that ESOM with U-matrix technologies (Allview) can be effectively used for the identification of groups of customers of mobile phone services. The algorithm sig* operates on the groups identified by Allview and produces a description of the groups in the form of understandable decision rules. In contrast to other decision rule algorithms the understandability of the generated rules is the main aim of this algorithm. The generated rules allowed understanding of the important characteristics of the identified groups. The rules are shown to be effective predictors of churn. A comparison of the distribution of the groups of churning and

non-churning customers allows inferences to be made about why customers churn. If the groups are analysed with respect to their total revenues, customer groups for which retention actions should be taken can be identified.

In many approaches to CRM a black box model for the prediction of churning is constructed. It is recommended that the effective construction of a performing classifier is not the central issue of effective CRM. The approach presented here aims at new insights into customer–business relationships. The output of this approach is rules that lead to a better understanding of who clients are and why churn is happening. These rules lead to direct business actions in re-designing products, services and marketing activities in order to prevent churn. This approach is data mining in the sense of producing new (formerly unknown) knowledge about a business's customers out of the data generated by customer–business interaction.

Acknowledgment

The project was supported by a research grant from Swisscom AG, Bern, Switzerland. The author wishes to thank Dr Manfred Schmidt of Swisscom Corporate Technology for initiating the project and for his cooperation. Susanne Schwenke and Ulrich Penndorf did substantial work for the project at the NeuroInformatics group at the University of Marburg.

References

1 Brown, S. A. (2000) 'Customer relationship management: A strategic imperative in the world of e-business', John Wiley & Sons.
2 Ultsch, A. (1999) 'Data mining and knowledge discovery with emergent self-organizing feature maps for multivariate time series', in Oja, E. and Kaski, S. (eds) 'Kohonen Maps', pp. 33–46.
3 Kohonen, T. (1982) 'Self-organized formation of topologically correct featuremaps', *Biological Cybernetics*, Vol. 43, pp. 59–69.
4 Ultsch, A. (1995) 'Self-organizing neural networks perform different from statistical k-means clustering', Gesellschaft f. Klassifikation, Basel, 8th–10th March.
5 Ultsch, A. (1994) 'The integration of neural networks with symbolic knowledge processing', in Diday *et al*. 'New approaches in classification and data analysis', pp. 445–454, Springer Verlag.
6 Ultsch, A. (1998) 'The integration of connectionist models with knowledge-based systems: Hybrid systems', Proceedings of the 11th IEEE SMC 98 International Conference on Systems, Men and Cybernetics, 11–14th October, San Diego.
7 Brown (2000) *op. cit*.
8 Edmonds, D. (1997) 'Fair trading in the mobile telephone market', Office of Telecommunications, http://www.oftel.gov.uk/fairtrade/mobser.htm, OFTEL, London, April.
9 Woods, E. and Kyral, E. (1997) 'Data mining, ovum evaluates', Catalunya, Spain.
10 Griner, P. F., Mayewski, R. J., Mushlin, A. I. and Greenland, P. (1981) 'Selection and interpretation of diagnostic tests and procedures', *Annals of Internal Medicine*, Vol. 94, pp. 555–600.
11 Erkel, A. R. v. and Pattynama, P. M. Th. (1998) 'Receiver operating characteristic (ROC) analysis: Basic principles and applications in radiology', *European Journal of Radiology*, Vol. 27, pp. 88–94.
12 Domingos, P. (1999) 'The role of Occam's Razor in knowledge discovery', in Fayyad, U. (ed.) 'Data mining and knowledge discovery', Vol. 4, Kluwer Academic Publishers, pp. 409–425.
13 Domingos (1999) *op. cit*.
14 Bern, Switzerland (http://www.swisscom.com).
15 Deboeck G. and Kohonen, T. (1998) 'Visual explorations in finance with self-organizing maps', Springer-Verlag.
16 Ultsch (1999) *op. cit*.
17 Ultsch (1994) *op. cit*.
18 *Ibid*.
19 Ultsch (1995) *op. cit*.