Neural Data Mining for Credit Card Fraud Detection

R. Brause,[1] T. Langsdorf[1], M. Hepp[2]
[1]*J.W.Goethe-University, Frankfurt a. M.,*
[2]*Gesellschaft f. Zahlungssysteme GZS, Frankfurt a. M., Germany*

Abstract

The prevention of credit card fraud is an important application for prediction techniques. One major obstacle for using neural network training techniques is the high necessary diagnostic quality: Since only one financial transaction of a thousand is invalid no prediction success less than 99.9% is acceptable.

Due to these credit card transaction proportions complete new concepts had to be developed and tested on real credit card data. This paper shows how advanced data mining techniques and neural network algorithm can be combined successfully to obtain a high fraud coverage combined with a low false alarm rate.

1 Introduction

The prediction of user behavior in financial systems can be used in many situations. Predicting client migration, marketing or public relations can save a lot of money and other resources. One of the most interesting fields of prediction is the fraud of credit lines, especially credit card payments. For the high data traffic of 400,000 transactions per day, a reduction of 2.5% of fraud triggers a saving of one million dollars per year.

Certainly, all transactions which deal with accounts of known misuse are not authorized. Nevertheless, there are transactions which are formally valid, but experienced people can tell that these transactions are probably misused, caused by stolen cards or fake merchants. So, the task is to avoid a fraud by a credit card transaction *before* it is known as "illegal".

With an increasing number of transactions people can no longer control all of them. As remedy, one may catch the experience of the experts and put it into an expert system. This traditional approach has the disadvantage that the expert's knowledge, even when it can be extracted explicitly, changes rapidly with new kinds of organized attacks and patterns of credit card fraud. In order to keep track with this, no predefined fraud models as in [5] but automatic learning algorithms are needed.

This paper deals with the problems specific to this special data mining application and tries to solve them by a combined probabilistic and neuro-adaptive approach for a given data base of credit card transactions of the GZS.

1.1 Modeling the data

The transaction data are characterized by some very special proportions:

- The probability of a fraud is very low (0.2%) and has been lowered in a preprocessing step by a conventional fraud detecting system down to 0.1%.
- Most of the 38 data fields (about 26 fields) per transaction contain symbolic data as merchant code, account number, client name etc.
- A symbolic field can contain as low as two values (e.g. the kind of credit card) up to several hundred thousand values (as the merchant code).
- The confidence limit for a transaction abort is very subjective and subject to client policy. Transactions with a confidence for fraud of higher than 10% are accepted to be revised or aborted.

These data proportions have several implications. For the very low fraud occurrence of only 0.1% a constant, "stupid" diagnosis of "transactions is no fraud" will have a success rate of 99.9%. All adaptive fraud diagnosis which has lower success than this 99.9% (e.g. [3] with 92.5% or [7] with 50%) is questionable. In principal, we are aiming for maximizing the correct diagnosis by minimizing both the number of false alarms and the number of fraud transactions not recognized.

2 Mining the symbolic data

One transaction can be seen as a data tuple $\mathbf{x}$ of features x_i : $\mathbf{x} = (x_1,..,x_n)$. For the analysis we distinguish between the categorical, symbolic features and the analog, numerical data. Let us treat the symbolic data first.

Our main concept for mining the symbolic data relays on the idea that all misuse transactions can be seen as a kind of rules: IF all symbolic features are given THEN misuse takes place. Combining several misuse rules together will result in less and shorter, more general rules.

Thus, we have to design a generalization mechanism in order to reduce the dependence of a rule on unimportant features.

2.1 Generalizing and weighting the association rules

In contrast to standard basket prediction association rules [1], [2] our goal does not consist of generating long associating rules but of shortening our raw associations by generalizing them to the most common types of transactions. Although generalizations are common for symbolic AI, there are no standard algorithms in data mining to do this.

How can such a generalization be done? We start with the data base of fraud transactions and compare each transaction with all others in order to find pairs of similar ones. Each pair is then merged into a generalized rule by replacing a non-identical feature by a 'don't-care'-symbol '*'. By doing so, a generalization process evolves, see Fig. 1. Here, the generalization of two transactions with the feature tuples $\mathbf{x}_1$ = (F,D,C,D,A) and $\mathbf{x}_2$ = (F,D,G,D,A) (dotted circle) to the rule (F,D,*,D,A) and further up to (F,*,*,D,A) and to (*,*,*,D,*) is shown. Thus, each generalization provides at least one 'don't-care'-symbol for an unimportant feature, increases the generalization level by one and shortens the rule excluding one feature. All generalizations which have not been generalized themselves are the root of the subgraph, forming a tree.

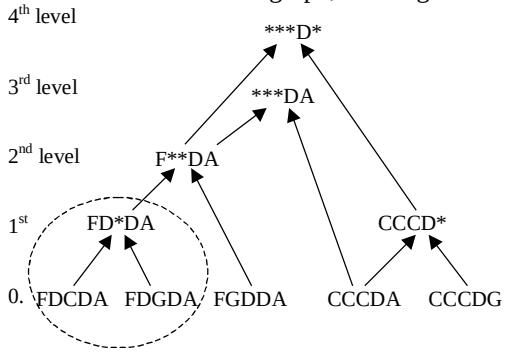

Fig. 1 *The generalization graph*

For the example of 5850 fraud data, there are 4 generalized rules in level 16 shown in Table 1.

Rule	ACCT_NBR	TRN_TYP	CURR_CD	POS_ENT_CD	FAL_SCOR	CRD_TYP	ICA_CD	AID_CD	SIC_CD	ACT_CD	MSG_TYP	MER_ID	MER_CNTY_CD	CTY_1	POST_CD_1	CNTY_CD1	CR_LMT	ATV_IND	ACCT_STAT	CTY_2	POST_CD_2	ADDR_STAT	EMIT_NBR	INST_NBR	ISS_REAS	GEN_CD	CARD_TYP
1	*	EA	840	*	*	EM	2768	8403184	*	0	1100	*	0	*	*	*	*	*	I	*	*	*	0	*	*	N	*	*
2	*	EA	840	[1]	0	EM	*	*	563	0	1100	*	0	*	*	*	*	*	I	*	*	*	0	*	*	*	*	*
3	*	EA	840	*	0	EM	2768	8403184	*	*	1100	*	0	*	*	*	*	*	I	*	*	*	0	*	*	*	002	*
4	*	EA	840	*	995	EM	*	*	*	0	1100	*	0	*	0	*	*	I	R	*	*	0	*	*	*	*	*	*

[1] ZZUTSZ1UZZZ1

Table 1 *Generalized transactions with 16 wildcards*

The feature names are labeled on the top of the columns.

All rules differ from each other. In general, there are many rules in a level. We define the *share* of a fraud rule as the percentage of fraud transactions which is covered by the rule.

Nevertheless, the share does not reflect the fact that there are also legal transactions which may fit a fraud rule leading to a wrong diagnosis. The more transactions with a correct diagnosis we have the more confidence in the diagnostic process we get. We define therefore the *confidence* in a fraud diagnosis as

$$\text{confidence} = \frac{\#\text{of misuse covered by the rule}}{\#\text{of transactions covered by the rule}} \qquad (2.1)$$

We can show that confidence = 1– P(false alarm) ≤ 1– P(false alarm|legal). Thus, when the confidence is maximized, the probability of a false alarm is minimized. For the rest of the paper, our main goal consists of maximizing the confidence of a fraud decision for an acceptable probability of fraud detection when fraud is present.

The mining algorithm is described in more detail in [4].

2.2 Results

For the analysis we used a sample set of 5,850 fraud transactions and 542,858 legal transactions, ordered by their time stamps. It should be noted that the mining algorithm has a high runtime complexity. Therefore, we used only 30,000 of the legal transactions. The resulting values for the confidence were compared to the whole set of transactions.

In the following Fig. 2 the performance of the rule diagnosis is shown as function of the generalization level.

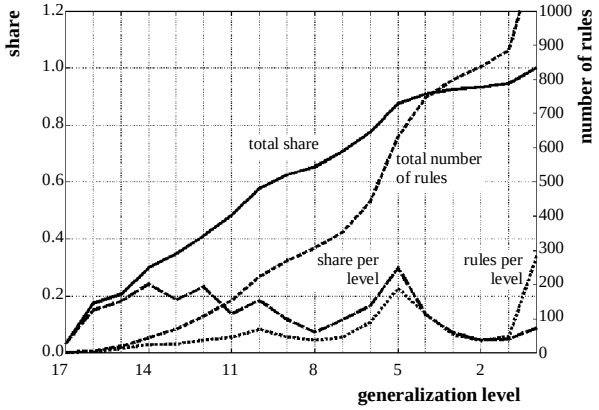

Fig. 2 *The performance of the rule diagnosis*

For each generalization level, i.e. for each number of wildcards, a set of active, non-generalized rules exists. They are denoted as "rules per level". Each set detects a certain part of the fraud, measured as "share per level". We can see that the main part of the share and the rules

are obtained for level 5 and above. Certainly, the more rules we take the better we perform. But, the less general the rules are, the more the performance will depend on statistical variations of the fraud data. If we take all the 747 rules from generalization level 4 up to level 17 we obtain a moderate confidence for the fraud detection on the set of all transactions, see Table 2.

#rules	**% correct diagnosis**			**confidence**
	legal	fraud	total	%
747	99.73	90.91	99.64	25.14
			(99.72)	(25.2)
510	99.97	83.08	99.79	75.17
			(99.953)	(73.5)
0	*99.9*	*0.0*	*99.9*	*0.0*

Table 2 *Fraud detection vs. confidence*

However, when we select only those rules which also preserve their confidence sufficiently on the whole transaction set, we obtain 510 rules. Certainly, with less rules the fraud diagnosis probability decreases slightly, but, as we see in the table, our main goal, the confidence in the diagnosis, is dramatically increased up to 75 % due to the high proportion of legal data which are less misclassified. This is also true when we use the real proportion for legal vs. misuse transactions of 1000:1 which are shown in round brackets in Table 2. Additionally, the diagnosis performance is even better than the constant, "stupid" diagnosis mentioned before and noted in the last table row.

3 Mining the analog data

Each transaction is characterized by symbolic and analog data. So far we have only used the symbolic part of the transactions. Does the analog part containing transaction time, credit amount etc. provide any useful information? Will it be possible to enhance the fraud diagnosis?

The problem of fraud diagnosis can be seen as separating two kinds or classes of events: the good and the bad transactions. Our problem is indeed a classification problem. One major approach for dynamic classification with demand driven classification boundaries is the approach of *learning* the classification parameters, the classification boundaries, by an adaptive process. Learning is the domain of artificial neural networks, and we used a special model of it to perform the task.

3.1.1 The network

There are several possible network approaches for the task. For our model we used one expert net for each feature group (time, money, etc.) and grouped the experts together to form a common vote. In Fig. 3 this architecture is shown.

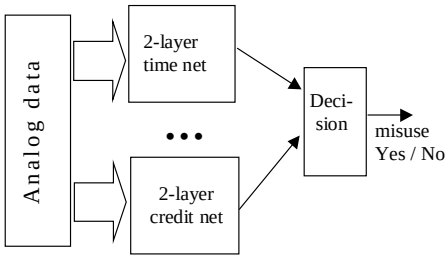

Fig. 3 *The neural network experts for analog data*

We used several networks of the Radial Basis Function (RBF) type [8], each one specialized on one topic.

3.1.2 The results

Because we have a very low fraud occurrence of only 0.1% the simple constant diagnosis "transactions is no fraud" will have a success rate of 99.9%. To compete with this trivial diagnosis, the task of diagnosing a transaction is not easy to do. If we use only the analog data, all transactions patterns characterized by *n* symbolic and *m* analog features are projected from the *n+m*-dimensional space into the *m*-dimensional space. Generally, this results in overlapping classes and therefore in diagnostic success far worse than 99.9%. In Fig. 4 the typical situation is shown for the separation of two classes by one analog variable x.

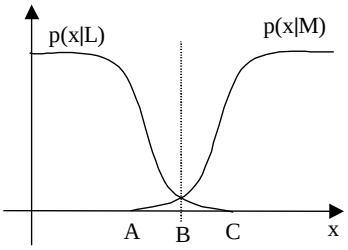

Fig. 4 *Diagnosis for overlapping classes*

Here, the two probability density functions p(x|M) for the fraud data and p(x|L) for the legal data are shown. For the best separation probability of the two clusters, the class boundary is located at point B in Fig. 4 where both densities are equal. But, for our two goals of high fraud detection success and high confidence in the detection we encounter a trade-off: If we choose the boundary at point A we get a high fraud discover probability and a low confidence (high false alarm rate) whereas for a high confidence we have to choose the class decision boundary at point C with a smaller fraud discovery success.

Now, let us diagnose one transaction by the means of the neural network. For that purpose, we used the neural expert system shown in Fig. 3 and trained it with our fraud data. We used 300 transactions for training and analyzed the state of the whole network afterwards by presenting 250 legal and 250 fraud data. The proportion

of legal to fraud data for training was changed, causing different diagnosing behavior. The results are shown in Table 3.

proportion	**correct diagnosis %**			**faulty diagnosis %**		**confidence %**
	total	legal	fraud	legal	fraud	
2:1	78.8	95.2	62.4	4.8	37.6	1.3
4:1	58.2	99.6	16.8	0.4	83.2	4.0
10:1	50.0	100	0	0	100	100

Table 3 *Shifting the class boundary*

As we can see, by augmenting the number of legal transactions in the training the class boundary shifts towards point C in Fig. 4. Here, the confidence is high, but the fraud discovery becomes zero.

4 Combining symbolic and analog information

In the previous sections we encountered the fact that the analog data can not serve as a satisfying criterion for fraud diagnosis. Therefore, we combined the diagnostic information of the rule-based association system of section 2 with the expert information of section 3 in a parallel network including a decision stage. The diagnostic influence of all the experts are initially the same and converge by 1:1 training in the limit to their appropriate value. In all situations, decisions based on the analog data can override the rule based expert. This is shown as "combined parallel approach" in Table 4.

Diagnostic method	**Prob. of correct diagnose**		**Confidence %**	
Data set size	1000	11,700	1000	11,700
Rule based	.901	.915	100.0	100.0
Analog data	.853	.817	1.55	93.1
Comb. par.	.928	.898	100.0	1.05
Comb. seq.	.845	.876	100.0	81.49
		(**.9995527**)		(**79.0**)

Table 4 *Comparing the performance of different diagnostic expert systems on two sets of data*

The parallel approach results in some extra diagnosis errors for legal transactions which decrease heavily the confidence down to 1%. Can we change this?

To do this, we also constructed a sequential system. Here, the decisions for "fraud" by the highly successful rule based expert module are checked additionally by the analog neural expert. Certainly, this does not decrease the probability for the first stage to classify fraud data as "legal", but it increases the probability for the diagnosis "fraud" to be correct and therefore increases the confidence and decreases the number of false alarms, see Table 4.

In summary, by an automatically generated rule system we managed to increase the inherent correct diagnosis of 99.9% to 99.95 %. Including also the analog information we increased this to 99.955%.

As most important topic the fraud decisions are about 80% valid which is quite high for this kind of problem.

5 Discussion

In this contribution we developed concepts for the statistic-based credit card fraud diagnosis. We showed that this task has to be based on the very special diagnostic situation imposed by the very small proportion of fraud data of 1:1000.

Additionally, we showed that, by algorithmically generalizing the transaction data, one may obtain higher levels of diagnostic rules. Combining this rule-based information and adaptive classification methods yield very good results.

Based on these results for a sample data base, additional work is necessary to design an online learning diagnostic system.

References

[1] R. Agrawal, H.Mannila, R. Srikant, H. Toivonen, A.I. Verkamo: *Fast Discovery of Association Rules.* In: U. Fayyad, G. Piatesky-Shapiro, P. Smyth, R. Uthurusamy (eds.): Advances in Knowledge Discovery and Data Mining. Menlo Park, AAAI/MIT Press 1996

[2] R. Agrawal, R. Srikant: *Fast Algorithms for mining association rules*. Proceedings of the VLDB Conference, Santiago, Chile, 1994

[3] P. Barson, S. Field, N. Davey, G. McAskie, R. Frank*: The Detection of Fraud in Mobile Phone Networks;* Neural Network World 6, 4, pp. 477-484 (1996)

[4] R. Brause, T. Langsdorf, M. Hepp: *Credit Card Fraud Detection by Adaptive Neural Data Mining,* J.W.Goethe-University, Comp. Sc. Dep., Report 7/99, Frankfurt, Germany (1999), also by http://www.cs.uni.frankfurt.de/fbreports/07.99.ps.gz

[5] S. Ghosh, D.L. Reilly: *Credit Card Fraud Detection with a Neural-Network*; Proc. 27th Annual Hawaii Int. Conf. on System Science, IEEE Comp. Soc. Press, Vol.3, pp.621-630 (1994)

[6] R.J. Hildermann, C.L. Carter, H.J. Hamilton, N. Cercone: *Mining Association Rules from Market Basket Data using Share Measures and Characterized Itemsets;* Int. J. of AI tools vol.7, No.2, pp.189-220, 1998

[7] Y. Moreau, H. Verrelst, J. Vandwalle: *Detection of Mobile Phone Fraud using Supervised Neural Networks: A First Prototype;* Proc. ICANN ´97, Lecture notes on computer science LNCS 1327, Springer Verlag 1997

[8] S. Haykin: *Neural networks* - a comprehensive foundation, MacMillan, New York 1994

[9] R. Srikant, R. Agrawal : *Mining generalized association rules*. Proc. VLDB Conference, Zurich, Switzerland, 1995