Unsupervised Profiling Methods for Fraud Detection

Richard J. Bolton and David J. Hand
Department of Mathematics
Imperial College
London
{r.bolton, d.j.hand}@ic.ac.uk

Abstract

Credit card fraud falls broadly into two categories: behavioural fraud and application fraud. Application fraud occurs when individuals obtain new credit cards from issuing companies using false personal information and then spend as much as possible in a short space of time. However, most credit card fraud is behavioural and occurs when details of legitimate cards have been obtained fraudulently and sales are made on a 'Cardholder Not Present' basis. These sales include telephone sales and e-commerce transactions where only the card details are required.

In this paper, we are concerned with detecting behavioural fraud through the analysis of longitudinal data. These data usually consist of credit card transactions over time, but can include other variables, both static and longitudinal. Statistical methods for fraud detection are often classification (supervised) methods that discriminate between known fraudulent and non-fraudulent transactions; however, these methods rely on accurate identification of fraudulent transactions in historical databases – information that is often in short supply or non-existent. We are particularly interested in unsupervised methods that do not use this information but instead detect changes in behaviour or unusual transactions. We discuss two methods for unsupervised fraud detection in credit data in this paper and apply them to some real data sets.

Peer group analysis is a new tool for monitoring behaviour over time in data mining situations. In particular, the tool detects individual accounts that begin to behave in a

way distinct from accounts to which they had previously been similar. Each account is selected as a target account and is compared with all other accounts in the database, using either external comparison criteria or internal criteria summarizing earlier behaviour patterns of each account. Based on this comparison, a peer group of accounts most similar to the target account is chosen. The behaviour of the peer group is then summarized at each subsequent time point, and the behaviour of the target account compared with the summary of its peer group. Those target accounts exhibiting behaviour most different from their peer group summary behaviour are flagged as meriting closer investigation.

Break point analysis is a tool that identifies changes in spending behaviour based on the transaction information in a single account. Recent transactions are compared with previous spending behaviour to detect features such as rapid spending and an increase in the level of spending, features that would not necessarily be captured by outlier detection.

Introduction

In the fight against fraud, actions fall under two broad categories: fraud prevention and fraud detection. Fraud prevention describes measures to stop fraud occurring in the first place. These include PINs for bankcards, Internet security systems for credit card transactions and passwords on telephone bank accounts. In contrast, fraud detection involves identifying fraud as quickly as possible once it has been perpetrated. We apply fraud detection once fraud prevention has failed, using detection methods continuously, as we will usually be unaware that fraud prevention has failed. In this article we are concerned solely with fraud detection.

Fraud detection must evolve continuously. Once criminals realise that a certain mode of fraudulent behaviour can be detected, they will adapt their strategies and try others. Of course, new criminals are also attempting to commit fraud and many of these will not be aware of the fraud detection methods that have been successful in the past, and will adopt strategies that lead to identifiable frauds. This means that the earlier detection tools need to be applied as well as the latest developments.

Statistical fraud detection methods may be ‘supervised’ or ‘unsupervised’. In supervised methods, models are trained to discriminate between fraudulent and non-fraudulent behaviour, so that new observations can be assigned to classes so as to optimise some measure of classification performance. Of course, this requires one to be confident about the true classes of the original data used to build the models; uncertainty is introduced when legitimate transactions are mistakenly reported as fraud or when fraudulent observations are not identified as such. Supervised methods require that we have examples of both classes, and they can only be used to detect frauds of a type that have previously occurred. These methods also suffer from the problem of unbalanced class sizes: in fraud detection problems, the legitimate transactions generally far outnumber the fraudulent ones and this imbalance can cause misspecification of models. Brause et al (1999) say that, in their database of credit card transactions, ‘the probability of fraud is very low (0.2%) and has been lowered in a preprocessing step by a conventional fraud detecting system down to 0.1%.’ Hassibi (2000) remarks ‘Out of some 12 billion transactions made annually, approximately 10 million – or one out of every 1200 transactions – turn out to be fraudulent.’

In contrast, unsupervised methods simply seek those accounts, customers, etc. whose behaviour is ‘unusual’. We model a baseline distribution that represents normal behaviour and then attempt to detect observations that show greatest departure from this norm. These can then be examined more closely. Outliers are a basic form of non-standard observation that can be used for fraud detection.

This leads us to note the fundamental point that we can seldom be certain, by statistical analysis alone, that a fraud has been perpetrated. Rather, the analysis should be regarded as alerting us to the fact that an observation is anomalous, or more likely to be fraudulent than others – so that it can then be investigated in more detail. One can think of the objective of the statistical analysis as being to return a suspicion score (where we will regard a higher score as more suspicious than a lower one). The higher the score is, then the more unusual is the observation, or the more like previously fraudulent values it is. The fact that there are many different ways in which fraud can be perpetrated, and many different scenarios in which it can occur, means that there are many different ways of computing suspicion scores.

We can compute suspicion scores for each account in the database, and these scores can be updated as time progresses. By ordering accounts according to their suspicion score, we can focus attention on those with the highest scores, or on those that exhibit a sudden increase in suspicion score. If we have a limited budget, so that we can only afford to investigate a certain number of accounts or records, we can concentrate investigation on those thought to be most likely to be fraudulent.

Credit Card Fraud

Credit card fraud is perpetrated in various ways but can be broadly categorised as application, ‘missing in post’, stolen/lost card, counterfeit card and ‘cardholder not present’ fraud. Application fraud arises when individuals obtain new credit cards from issuing companies using false personal information; application fraud totalled £10.2 million in 2000 (Source: APACS) and is the only type of fraud that actually declined between 1999 and 2000. ‘Missing in post’ (£17.3m in 2000) describes the interception of credit cards in the post by fraudsters before they reach the cardholder. Stolen or lost cards accounted for £98.9 million in fraud in 2000, but the greatest percentage increases between 1999 and 2000 were in counterfeit card fraud (£50.3m to £102.8m) and ‘cardholder not present’ (i.e. postal, phone, internet transactions) fraud (£29.3m to £56.8m). To commit these last two types of fraud it is necessary to obtain the details of the card without the cardholder’s knowledge. This is done in various ways, including employees using an unauthorised ‘swiper’ that downloads the encoded information onto a laptop computer and hackers obtaining credit card details by intrusion into companies’ computer networks. A counterfeit card is then made, or the card details simply used for phone, postal or Internet transactions.

Supervised methods to detect fraudulent transactions can be used to discriminate between those accounts or transactions known to be fraudulent and those known (or at least presumed) to be legitimate. For example, traditional credit scorecards (Hand and Henley,

1997) are used to detect customers who are likely to default, and the reasons for this may include fraud. Such scorecards are based on the details given on the application forms, and perhaps also on other details, such as bureau information. Classification techniques, such as statistical discriminant analysis and neural networks, can be used to discriminate between fraudulent and non-fraudulent transactions to give transactions a suspicion score.

However, information about fraudulent transactions may not be available and in these cases we apply unsupervised methods to attempt to detect fraud. These methods are scarce in the literature and are less popular than supervised methods in practice as suspicion scores reflect a propensity to act anomalously when compared with previous behaviour. This is different to suspicion scores obtained using supervised techniques, which are guided to reflect a propensity to commit fraud in a manner already previously discovered. The idea behind suspicion scores from unsupervised methods is that unusual behaviour or transactions can often be indicators of fraud. An advantage of using unsupervised methods over supervised methods is that previously undiscovered types of fraud may be detected. Supervised methods are only trained to discriminate between legitimate transactions and previously known fraud.

Unsupervised methods and their application to fraud detection

As we mentioned above, the emphasis on fraud detection methodology is with supervised techniques. In particular, neural networks have proved popular – predictably, perhaps, given the attention they have received. Researchers who have used neural networks for supervised credit card fraud detection include Ghosh and Reilly (1994), Aleskerov et al. (1997), Dorronsoro et al. (1997), and Brause et al (1999). However, unsupervised credit card fraud detection has not received attention in the literature.

Unsupervised fraud detection methods *have* been researched in the detection of computer intrusion (hacking). Here profiles are trained on the combinations of commands that a user uses most frequently in their account. If a hacker gains illegal access to the account then their intrusion is detected by the presence of sequences of commands that are not in the profile of commands typed by the legitimate user. Qu, Vetter et al. (1998) use probabilities of events to define the profile, Lane and Brodley (1998), Forrest et al (1996) and Kosoresow and Hofmeyr (1997) use similarity of sequences that can be interpreted in a probabilistic framework.

Unsupervised methods are useful in applications where there is no prior knowledge as to the particular class of observations in a data set. For example, we may not be able to know for sure which transactions in a database are fraudulent and which are legitimate. In these situations, unsupervised methods can be used to find groups or find outliers in the data. Essentially, we collect data to provide a summary of the system that we are studying. Once we have a summary of the behaviour of the system, we can identify those observations that do not fit in with this behaviour, i.e. anomalous observations. This is our aim in using unsupervised statistical techniques for fraud detection.

The most popular unsupervised method used in data mining is clustering. This technique is used to find natural groupings of observations in the data and is especially useful in market segmentation. However, cluster analysis can suffer from a bad choice of metric (the way we scale, transform and combine variables to measure the 'distance' between observations); for example, it can be difficult to combine categorical and continuous variables in a good clustering metric. Observations may cluster differently on some subsets of variables than they do on others so that we may have more than one valid clustering in a data set.

We can use unsupervised methods such as clustering to help us form local models from which we can find local outliers in the data. In the context of fraud detection, a global outlier is a transaction anomalous to the entire data set; for example, a purchase of several thousand pounds would be a global outlier if all other transactions in the database were considerably less than that amount. Local outliers describe transactions that are anomalous when compared to subgroups of the data. Local outlier detection is effective in situations where the population is heterogeneous; this is true of credit card transaction data where spending behaviour between accounts can vary according to amounts spent and the purchases that are made. If we can identify the spending behaviour of a particular account, then a transaction is a local outlier if it is anomalous to spending in that account (or accounts similar to it), but not necessarily anomalous to the entire population of transactions. For example, a transaction of a thousand pounds in an account where, historically, all transactions have been under a hundred pounds might be considered as a local outlier; however, such a transaction may not have been considered unusual if it had occurred in a high spending account, and thus would not be a global outlier.

The fundamental challenge is in the formation of the local model, which can be achieved in a variety of ways. One way is through cluster analysis. Here, legitimate transactions from all accounts are clustered into groups with similar characteristics. The local model, or profile, of a particular account is then determined by the clusters to which its transactions are allocated. If a future transaction from the account is then allocated to a cluster not in the account profile, then an alarm is raised for that transaction. Care must be exercised in choosing variables and metrics on which to cluster.

Nearest-neighbour methods can be employed to combine transaction information from accounts that exhibit similar behaviour. We have developed Peer Group Analysis as a tool that uses local models of spending behaviour over time to detect changes in spending within accounts; we describe an application of Peer Group Analysis to fraud detection below. We follow this with a description of Break Point Analysis. Here, a local model is created and updated by drawing information from transactions within the same account. Sequences of transactions within that account are compared with this local model to indicate changes in spending behaviour.

Peer Group Analysis

We propose Peer Group Analysis (Bolton and Hand, 2001) as a candidate method for unsupervised fraud detection. Peer group analysis is a new tool for monitoring behaviour over time in data mining situations. In particular, the tool detects individual objects that begin to behave in a way distinct from objects to which they had previously been similar. Each object is selected as a target object and is compared with all other objects in the database, using either external comparison criteria or internal criteria summarizing earlier behaviour patterns of each object. Based on this comparison, a peer group of objects most similar to the target object is chosen. The behaviour of the peer group is then summarized at each subsequent time point, and the behaviour of the target object compared with the summary of its peer group. Those target objects exhibiting behaviour most different from their peer group summary behaviour are flagged as meriting closer investigation. The tool is intended to be part of the data mining process, involving cycling between the detection of objects that behave in anomalous ways and the detailed examination of those objects. Several aspects of peer group analysis can be tuned to the particular application, including the size of the peer group, the width of the moving behaviour window being used, the way the peer group is summarised, and the measures of difference between the target object and its peer group summary. The distinguishing feature of Peer Group Analysis (PGA) lies in its focus on local patterns rather than global models (Hand *et al*, 2000; Hand, Mannila, and Smyth, 2001): a sequence may not evolve unusually when compared with the whole population of sequences but may display unusual properties when compared with its peer group. That is, it may begin to deviate in behaviour from objects to which it has previously been similar.

Let us suppose that we have observations on N objects, where each observation is a sequence of d values, represented by a vector, $\mathbf{x}_i$, of length d. The jth value of the ith observation, x_{ij}, occurs at a fixed time point t_j. Let $PG_i(t_j)$ = {Some subset of observations ($\neq \mathbf{x}_i$) which show behaviour similar to that of $\mathbf{x}_i$ at time t_j}. Then $PG_i(t_j)$ is the peer group of object i, at time j. The parameter *npeer* describes the number of objects in the peer group and effectively controls the sensitivity of the peer group analysis. The size of *npeer* reflects how local a model we require. Of course, if *npeer* is chosen to be too small then the behaviour of the peer group may be too sensitive to random errors and thus inaccurate.

Let S_{ij} be a statistic summarizing the behaviour of the ith observation at time j. We will define similarity between objects in terms of their measures, S_{ij}. This measure could be a sequence of observations preceding time point j or it could be some statistical summary of these observations, such as a moving average or a trend. We define a (dis)similarity metric $D(S_{i1}, S_{j1})$, $\forall j \neq i$, on the S_{i1} to order objects according to how similar their behaviour at t_1 is to that of the target object, $\mathbf{x}_i$. The *npeer* most similar objects to the target object comprise the peer group, PG_i. Choice of a suitable metric depends on the data to be analyzed; we have used a two-stage variant of the Euclidean distance metric in this paper since the example data sets here contain continuous variables, but different metrics will be more suitable for categorical data or for data with variables on greatly differing scales of measurement. Different metrics may well yield different results (as with cluster analysis), so are worth exploring.

Once we have found the peer group for the target observation $\mathbf{x}_i$ we can calculate peer group statistics, P_{ij}. These will generally be summaries of the values of S_{ij} for the members of the peer group. The principle here is that the peer group initially provides a local model, P_{i1}, for S_{i1}, thus characterizing the local behavior of $\mathbf{x}_i$ at time t_1, and will subsequently provide models, P_{ij}, for S_{ij}, at time $t_j, j>1$. If our target observation, S_{ik}, deviates 'significantly' from its peer group model P_{ik} at time t_k, then we conclude that our target is no longer behaving like its peers at time t_k. If the departure is large enough, then the target observation will be flagged as worthy of investigation.

To measure the departure of the target observation from its peer group we calculate its standardized distance from the peer group model; the example we use here is a standardized distance from the centroid of the peer group based on a t-statistic. The centroid value of the peer group is given by the equation:

$$P_{ij} = \frac{1}{npeer}\left(\sum_{p \in PG_i(t_1)} S_{pj} \right); \qquad j \geq 1,\ p \neq i.$$

where $P_i(t_1)$ is the peer group calculated at time t_1. The variance of the peer group is then

$$V_{ij} = \frac{1}{(npeer - 1)} \sum_{p \in PG_i(t_1)} \left(S_{pj} - P_{ij}\right)\left(S_{pj} - P_{ij}\right)'; \ j \geq 1,\ p \neq i.$$

The square root of this can be used to standardize the difference between the target S_{ij} and the peer group summary P_{ij}, yielding

$$T_{ij} = \left(S_{ij} - P_{ij}\right) \Big/ \sqrt{V_{ij}}$$

Targets behaving anomalously when compared to their peer group will produce large values of the T_{ij} measures. Issues of multiple testing make a direct probability interpretation difficult; we recommend instead using the T_{ij} measures as scores (Hand, 2000; Hand, Mannila, and Smyth, 2000), simply flagging those objects with scores that deviate most substantially as worthy of closer investigation.

More details of the PGA method can be found in Bolton and Hand (2001). This paper describes an implementation of PGA to detect changes in credit card account spending behaviour and illustrates its propensity to detect outliers through a simulation study. Here, the initial peer group for a target is based on similarity of statistics for a time window containing the first 15 time periods – this number was chosen arbitrarily as a period long enough to measure the spending profile of an account. Once the peer group has been identified, the statistics for the target account are calculated for future time windows and compared to summary statistics for its peer group. Any target account showing spending behaviour anomalous to that of its peer group is flagged for further inspection. In this application a time window the same length as that used for the peer group determination is used to calculate statistics for future time periods.

We adapt PGA method for use in detecting credit card fraud by changing the length of the time windows subsequent to that used initially to determine the peer group. This is because we are interested in short-term changes in spending behaviour and shortening the time window allows these changes to be detected; a longer time window is more likely to have a smoothing effect and hide these changes. We keep the initial time window for the determination of the peer group long enough to give a good estimate of the spending behaviour for each account. Our example data set contains the total credit card spending in 858 accounts over a 52-week period, with the total spending recorded per week. In our example, we set the initial time window for calculating the peer group to be equal to thirteen weeks (a quarter of a year) and future time windows to have length equal to four weeks. Of course, we could reduce the length of the future time windows to just one week; however, there are advantages in using a longer time window here. A card user may spend a relatively large amount in one week but compensate by spending less in the weeks preceding or succeeding this purchase. A time window of one week will flag this purchase as an outlier, whereas the longer time window uses the history of spending in the account to adjust for such a purchase and is thus robust to such practices. Outliers are still flagged when we use a time window of four weeks, but only if they are inconsistent with spending trends for the peer group in the last four weeks and not just the week in which they occur. In this way, we can reduce the number of accounts that we flag unnecessarily.

The statistic that we use to compare spending between accounts is the mean amount spent over the time window. An example of the patterns exhibited by individual customers (Figure 16 of Hand and Blunt (2000)), shows how the slopes of cumulative credit card spend over time are remarkably linear; this suggests that a linear statistic such as the mean will be a suitable measure of credit card spending over time. Sudden jumps in these curves, or sudden changes of slope, merit investigation. A large increase in the mean spending of an account is an indicator of unusual behaviour and perhaps fraud.

Plots illustrate the power of PGA to detect local anomalies in the data. The vertical axis shows cumulative credit card spend as weeks pass on the horizontal axis. The spending of the target observation is represented by a thick black line and the spending of the peer group by orange (greyscale: dark grey) lines; spending from a sample of the remaining accounts is represented by pale blue (greyscale: light grey) lines. Figure 1 shows an account flagged as having the highest suspicion score at week 17. The spending in this account shows spending in this week that is large when compared to the spending from accounts in its peer group. Figure 2 displays another example of unusually large spending for a particular account at week 23. Neither of the weekly spends for the target account appear large when compared with data from accounts outside their peer group; however, PGA can detect that the spending for these weeks is unusual amongst accounts that have similar spending trends.

We are currently investigating the possibility of adapting PGA so that suspicion scores can be calculated for each transaction, rather than for spending over weekly periods.

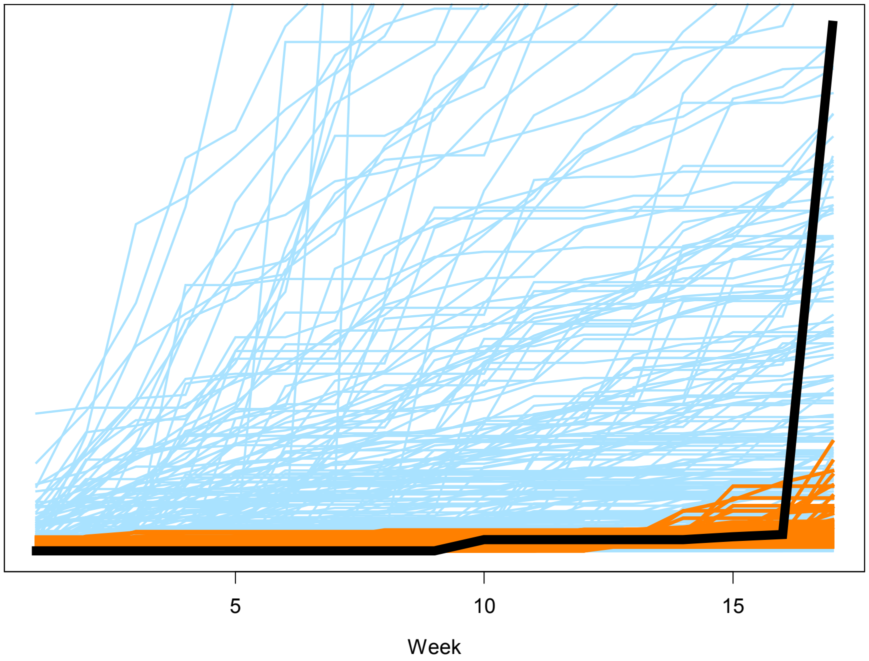

Figure 1. Acct #320

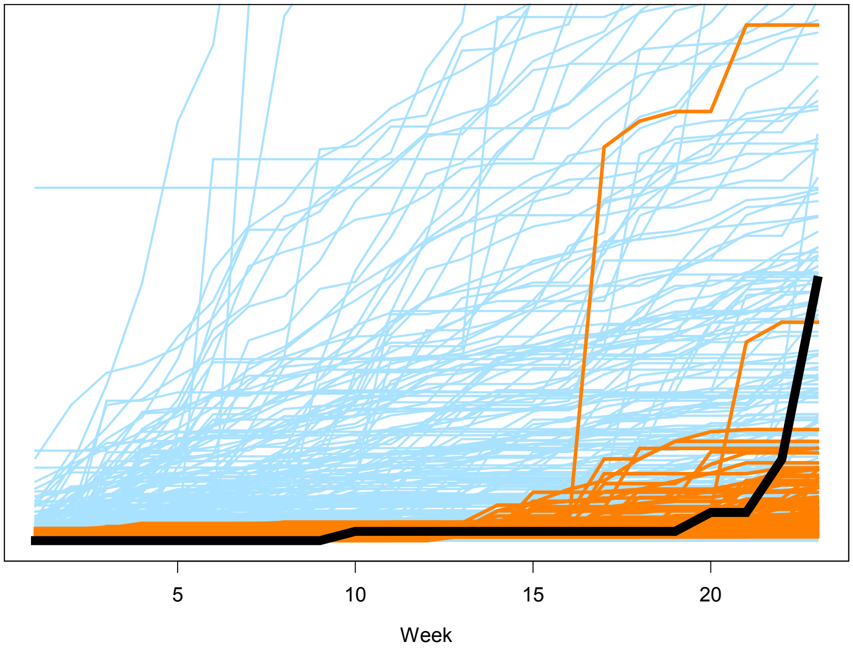

Figure 2. Acct #591

Break Point Analysis

Break Point Analysis is another unsupervised outlier detection tool that we are developing for behavioural fraud detection. A break point is an observation or time where anomalous behaviour is detected; Senator (2000) mentions break detection in a supervised context for detecting money laundering. Break point analysis operates on the account level, comparing sequences of transactions (their amount or frequency) to detect a change in behaviour for a particular account. In break point analysis, we have a fixed-length moving window of transactions: as a transaction occurs so it enters the window and the oldest transaction in the window is removed. Transactions in the most recent part of the window are then compared with those in the early part of the window to see if a change in spending behaviour has occurred. We must set parameters such as the length of the window and the proportion of 'old' to 'new' transactions in the comparison. Statistical tests are employed to see if the recent transactions follow a different pattern of behaviour to older transactions. Sudden increases in frequency of transactions or amount of transactions can be indicators of fraudulent behaviour. An advantage of break point analysis is that we do not require 'balanced' data (i.e. data summarised at fixed time points, e.g. weekly), as we are not comparing transactions between different accounts; we can also identify anomalous sequences of events that may indicate fraudulent behaviour. However, break point analysis does not draw on transactions from similar accounts to form a profile so the profile tests are not as powerful as those in peer group analysis.

We applied break point analysis on spending in some credit card accounts, choosing the window of transactions arbitrarily to contain 24 transactions - 20 transactions that form the local model or profile, and the next 4 transactions to test for an increase in spending. We compared mean values of amounts spent in each window using a simple *t*-test for computational efficiency. Accounts were ranked by the size of the *t*-statistic (which we can use as a suspicion score), using a window of 20 transactions for the local model and a window of 4 transactions to compare against the local model.

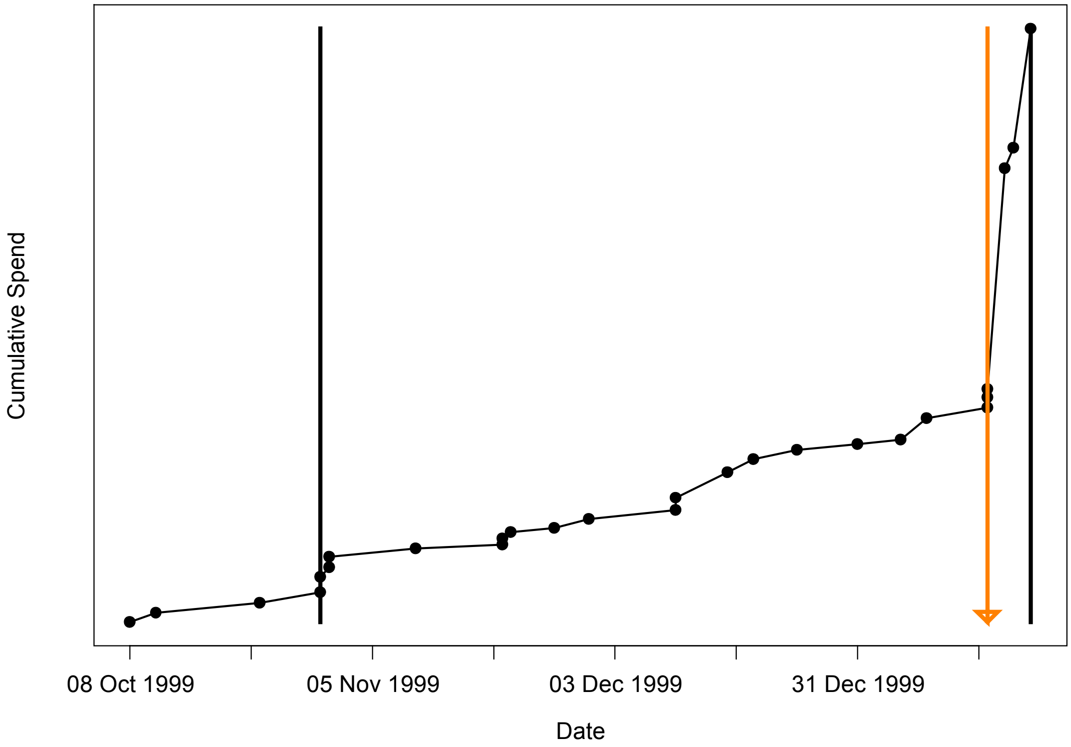

Figure 3. Breakplot by amount.

The spending behaviour for the account with the largest *t*-statistic from a sample of 200 accounts is shown in Figure 3. The vertical black lines define the window of observations that produced the large *t*-statistic. Transactions before the arrow form the local model and transactions after the arrow display spending behaviour anomalous to the local model. Spending clearly increases substantially immediately after the arrow and may merit investigation.

Although we have only used a small sample here, break point analysis translates easily to very large data sets as it scales linearly with the number of transactions.

Conclusions

We have discussed some possible approaches to unsupervised credit card fraud detection through behavioural outlier detection techniques. The methods in this article describe early stages of research to produce some frameworks for unsupervised fraud detection and elementary examples are shown for illustrative purposes. We aim to proceed by incorporating other information, other than simply the amount spent, into the anomaly detection process and identifying the most useful and practical methods to develop for fraud detection.

References

Aleskerov, E., Freisleben B., and Rao B.(1997). CARDWATCH: A Neural Network Based Database Mining System for Credit Card Fraud Detection. *Computational Intelligence for Financial Engineering, Proceedings of the IEEE/IAFE*, 220-226.

Blunt G. and Hand D.J. (2000) *The UK credit card market*. Technical Report, Department of Mathematics, Imperial College, London.

Bolton R.J. and Hand D.J. (2001) *Peer Group Analysis.* Technical Report, Department of Mathematics, Imperial College, London.

Brause, R., Langsdorf T. and Hepp M. (1999). Neural Data Mining for Credit Card Fraud Detection. *Proceedings. 11th IEEE International Conference on Tools with Artificial Intelligence.*

Dorronsoro, J. R., Ginel F., Sanchez C. and Cruz C. S. (1997). Neural Fraud Detection in Credit Card Operations. *IEEE Transactions on Neural Networks* **8**(4), 827-834.

Forrest, S., Hofmeyr S., Somayaji A. and Longstaff T. (1996). A sense of self for unix processes. *Proceedings of the 1996 IEEE Symposium on Security and Privacy*, Los Alamitos, CA.

Ghosh, S. and Reilly D. L. (1994). Credit Card Fraud Detection with a Neural-Network. *Proceedings of the 27th Annual Hawaii International Conference on System Science. Volume 3 : Information Systems: DSS/Knowledge-Based Systems*, J. F. Nunamaker and R. H. Sprague, Eds., Los Alamitos, CA, USA

Hand D.J. and Blunt G. (2000) Prospecting for gems in credit card data. *Proceedings of the Workshop on Statistical Modelling for Data Mining*, University of Pavia

Hand D.J., Blunt G., Kelly M.G., and Adams N.M. (2000) Data mining for fun and profit. *Statistical Science*, **15**, 111-131.

Hand D.J. and Henley W.E. (1997) Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society, Series A*, **160**, 523-541.

Hand D.J., Mannila H., and Smyth P. (In press) Principles of Data Mining, MIT Press.

Hassibi, K. (2000). Detecting Payment Card Fraud with Neural Networks. *Business Applications of Neural Networks*. P.J.G. Lisboa, A.Vellido, B.Edisbury Eds. Singapore: World Scientific.

Kosoresow, A. P. and Hofmeyr S. A. (1997). Intrusion Detection via System Call Traces. *IEEE Software* **14**(5), 24-42.

Lane, T. and Brodley C. E. (1998). Temporal Sequence Learning and Data Reduction for Anomaly Detection. *Proceedings of the 5th ACM Conference on Computer and Communications Security (CCS-98)*. New York, 150-158.

Leonard, K. J. (1993). Detecting Credit Card Fraud Using Expert Systems. Computers and Industrial Engineering 25(1-4), 103-1

Qu, D., Vetter B. M., Wang F., Narayan R., Wu S. F., Hou Y. F., Gong F. and Sargor C. (1998). Statistical Anomaly Detection for Link-State Routing Protocols. *Proceedings. Sixth International Conference on Network Protocols*, 62-70.

Senator, T. E. (2000). Ongoing Management and Application of Discovered Knowledge in a Large Regulatory Organization: A Case Study of the Use and Impact of NASD Regulation's Advanced Detection System (ADS). *Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* N.Y., 44-53.