International Journal of Forecasting 28 (2012) 216–223

Contents lists available at SciVerse ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Overcoming selectivity bias in evaluating new fraud detection systems for revolving credit operations

David J. Hand [a,b], Martin J. Crowder [b,*]

[a] Department of Mathematics, Imperial College London, SW7 2AZ, United Kingdom

[b] Institute for Mathematical Sciences, Imperial College London, SW7 2PG, United Kingdom

ARTICLE INFO

Keywords:
Financial fraud
Fraud detection
Fraud estimation

ABSTRACT

When proposed new fraud detection systems are tested in revolving credit operations, a straightforward comparison of the observed fraud detection rates is subject to a selectivity bias that tends to favour the existing system. This bias arises from the fact that accounts are terminated when the existing system, but not the proposed new system, detects a fraudulent transaction. This therefore flatters the estimated detection rate of the existing system. We develop more formal estimators that can be used to compare the existing and proposed new systems without risking this effect. We also assess the magnitude of the bias.

1. Problem setting

In this introductory section we set out the problem, before getting on with the analysis in the following two sections.

Fraud is a perennial problem of revolving credit operations. Fraud in this area comes in many guises, including counterfeit, mail-non-receipt, and card-not-present fraud. Although card operators constantly introduce new detection and prevention strategies, fraudsters likewise are constantly trying to develop new ways of circumventing them. A recent example was provided by the 14th February 2006 roll-out of the chip and PIN system in the UK. As expected, this led to a reduction in face-to-face fraud, mail-non-receipt fraud, and lost and stolen fraud, decreasing by 47%, 62%, and 23% respectively from 2005 to 2006 (APACS figures). However, card-not-present fraud increased by 16% over this period, to the extent that, at £212.6 million, it accounted for half of all losses in 2006 (APACS figures).

This paper is concerned with systems for detecting fraud in such revolving credit operations. Many approaches to developing such systems have been explored, based on a variety of data sources, including the transactions themselves, the behavioural and demographic characteristics of the account holders, and the properties of the merchants from whom purchases are made. Reviews are given by Bolton and Hand (2002), Fawcett and Provost (2002) and Phua, Lee, Smith, and Gayler (2005). These methods include:

- rule-based methods, which detect the occurrence of certain known kinds of suspicious transaction (e.g. Deshmukh & Talluru, 1997; Rosset, Murad, Neumann, Idan, & Pinkas, 1999);
- behavioural profiling methods, which model each individual's behaviour pattern and monitor it for departures from the norm, based on outlier detection and peer-group analysis methods (e.g. Juszczak, Adams, Hand, Whitrow, & Weston, 2008; Weston, Hand, Adams, Juszczak, & Whitrow, 2008); and
- discriminant methods, which model the differences between fraudulent and non-fraudulent transactions, and which may be based on a wide variety of supervised classification tools, including logistic regression, neural networks, random forests, etc (e.g. Whitrow, Hand, Juszczak, Weston, & Adams, 2008).

Examples of such systems include Falcon, DMS-250, Base24, Cardinel Centinel, and components of the SAS

* Corresponding author.
E-mail address: m.crowder@imperial.ac.uk (M.J. Crowder).

0169-2070/$ – see front matter
doi:10.1016/j.ijforecast.2010.10.005

system, as well as various specially-developed in-house tools.

Central to both the development and use of a fraud detection system is being able to measure its performance. One needs to know how effective a system is, whether modifications will improve it, and whether it is better or worse than some alternative. Superficially, this issue is straightforward: one can simply determine how many fraudulent transactions the system detects and how many it misses, or perhaps base similar calculations on the amount of money that would have been fraudulently stolen had the system not been in place. In practice, however, things are much more complicated. For a start, one also needs to know how many legitimate transactions a system flags as suspicious: a system which flagged every transaction for investigation would be a hundred percent successful in detecting fraudulent transactions, but would be useless in practice. Moreover, the timeliness of detection is of the essence: a system which was perfectly accurate, but only detected the fraud three months after the event, would be useless. One would like to know as soon as possible if a series of fraudulent transactions was being made, ideally in time to stop the first such transaction. Or, to take another example, the counterfactual amount of money that would have been stolen if the fraud detection system had not been in place is ill-defined: how long does one imagine that the series of fraudulent transactions would have continued for if it had not been detected? In particular, the standard performance criteria for evaluating classifiers and detectors (Hand, 1997) are typically not appropriate for fraud detection. Such issues have been explored by Hand, Whitrow, Adams, Juszczak, and Weston (2008), who develop a fraud detection system performance criterion.

A further complication is that, as noted above, one characteristic of the banking fraud environment is that it is constantly changing. The introduction of new technologies and financial products generates new opportunities for fraud. Improvements in fraud detection systems make life harder for fraudsters, who therefore adapt their tactics. Such changes mean that one cannot assume that an existing fraud detection system will remain effective; instead, new and updated methods must constantly be introduced. The situation has been described as an arms race between the fraudsters and the financial institutions. All of this means that one needs effective ways of evaluating the performance of detection systems.

When a potential improvement to an existing fraud detection system is proposed, or when an institution is considering installing an alternative system, it is necessary to evaluate the two competing systems, the existing one and the new one, to decide which is superior. This is not straightforward, because of the asymmetry in the ways in which the two systems are being applied. The data showing transaction streams will have been collected using the existing system, not the proposed new one. In particular, transaction sequences may have been terminated on the basis of the existing detector's scores, but will not have been terminated because of the scores of the new detector. This asymmetry, which results in a selection bias in the data available for analysis, is described in more detail in the next section. A failure to take it into account when evaluating fraud detection systems means that the evaluations are biased in favour of the existing system. That is, a proposed new system which is in fact more effective in detecting fraud than the existing system may well not be recognised as being more effective. The aim of this paper is to develop an unbiased estimator of the effectiveness of a fraud detection system, and to assess the size of the biases in practice.

In what follows, we assume that each account generates a sequence of transactions, each of which may be either fraudulent (labelled f) or legitimate (labelled n, for 'non-fraudulent'). Hopefully, most accounts will consist entirely of ns, but some will contain a mix of ns and fs (typically a sequence of ns followed by a mix of ns and fs, or a sequence of ns followed by a sequence of fs). A generic detection system produces a *suspicion score* (Bolton & Hand, 2002) for each transaction. We need not go into details here about how this score is constructed; for more information, see the references cited above. When a transaction's suspicion score exceeds some given threshold (i.e., when the transaction is *flagged*), it is subjected to a closer examination. An example of such a 'closer examination' would be a call to the account holder, to check that they made the transaction. If this closer examination reveals that the flagged transaction is legitimate, then the sequence of transactions continues. If, however, the closer examination reveals that the flagged transaction is indeed fraudulent, then the sequence is terminated. For the sake of simplicity, we assume that the termination occurs *at* the flagged fraudulent transaction, but one could easily relax this assumption. Thus, transaction sequences continue until either the detector identifies a true fraud in this way, or we reach the end of the observation period which is to generate data for evaluating the system's performance. When a flagged transaction is discovered to be fraudulent, all preceding transactions in that account are checked, to determine their true n/f status. If a flagged transaction is discovered to be legitimate, previous transactions are not checked, meaning that their true status remains unknown.

This outline can easily be extended to the case where the basic units of analysis are *activity records* (Whitrow et al., 2008), where detection is based on statistical summaries of short series of transactions, instead of on individual transactions.

From a purely statistical perspective, the problem is clearly one of hypothesis testing: we have one detector, and we want to know whether or not the performance of a proposed new detector is significantly better, based on samples of data from each detector, and subject to the complications outlined above. In practice, however, if one is to have any hope of having the industry adopt a proposed new method, it is necessary to take into account the perspectives and context in which the methods are to be applied. A radically new approach is unlikely to gain acceptance, especially in an intrinsically conservative industry. Instead, a gradual and incremental process of improvement is needed. This explains why this paper concentrates on a comparison of point estimates, rather than a hypothesis testing approach: the work in this paper has been developed in the context of our discussions with