Utility-Based Fraud Detection

Luis Torgo and **Elsa Lopes**
Fac. of Sciences / LIAAD-INESC Porto LA
University of Porto, Portugal
ltorgo@fc.up.pt and elsalopes@gmail.com

Abstract

Fraud detection is a key activity with serious socio-economical impact. Inspection activities associated with this task are usually constrained by limited available resources. Data analysis methods can provide help in the task of deciding where to allocate these limited resources in order to optimise the outcome of the inspection activities. This paper presents a multi-strategy learning method to address the question of which cases to inspect first. The proposed methodology is based on the utility theory and provides a ranking ordered by decreasing expected outcome of inspecting the candidate cases. This outcome is a function not only of the probability of the case being fraudulent but also of the inspection costs and expected payoff if the case is confirmed as a fraud. The proposed methodology is general and can be useful on fraud detection activities with limited inspection resources. We experimentally evaluate our proposal on both an artificial domain and on a real world task.

1 Introduction

Fraud detection is a very important application domain that has been addressed in several research areas (e.g. [Hand, 2002; Phua *et al.*, 2005; Fawcett and Provost, 1997]). Due to the intrinsic characteristics of fraudulent events it is highly related to other research topics like outlier detection, anomaly detection or change detection. The connecting feature among these topics is the interest on deviations from "normal" behaviour that is more frequently observed. Depending on the characteristics of the available fraud data different data mining methodologies can be applied. Namely, when the available data includes information regarding each observation being (or not) fraudulent, supervised classification techniques are typically used (e.g. [Ghosh and Reilly, 1994]). On the contrary, in several domains such classifications do not exist and thus unsupervised techniques are required (e.g. [Bolton and Hand, 2001]). Finally, we may have a mix of both types of data with a few labelled observations (e.g. resulting from past inspection activities) and a large set of unlabelled cases. These situations are often handled with semi-supervised approaches (e.g. [Nigam *et al.*, 2000]).

Fraud detection usually involves two main steps: i) decide which cases to inspect, and ii) the inspection activity in itself. The latter of these steps is frequently constrained by the available resources, e.g. time, man power or financial. An information system that is designed to support these activities should take this important aspect into account. These systems are typically used in the first step - help in deciding which observations should be inspected due to high suspicion of fraudulent behaviour. It is in this decision process that data mining can help by providing some guidance regarding the priorities for the posterior inspection activities. Given a set of observations data mining systems can signal which ones to inspect using data analysis techniques that detect deviations from normality. Due to the low frequency of fraudulent cases, this can be cast as an outlier detection task. Most outlier detection methods produce a "yes/no" answer for each observation. This type of methods is not adequate for most fraud detection tasks due to the above-mentioned limited inspection resources. In effect, such methods can lead to solutions where there are too many inspection signals for the available resources. If that occurs the user is left alone in choosing which ones to address. In this context, we claim that the methods should produce an outlier ranking instead. With such result users can easily adjust the solution to their available inspection resources, with the guarantee that the most "promising" cases are addressed first.

The way the ranking of the candidates is obtained may have a strong impact on the optimisation of the inspection results. Most outlier ranking methods produce ranks according to the probability of being an outlier of the observations. Cases with higher probability of being outliers appear on top positions of the ranks. In this paper we claim that this approach may lead to sub-optimal results in terms of optimising the inspection resources. In typical real world fraud detection applications the costs of inspecting a certain entity may vary from case to case, and moreover, the payoffs of detecting a fraudulent case may also vary. In other words the utility of the fraudulent cases is different. Outlier ranking methods that solely look at the probability of a certain case being or not an outlier (i.e. being suspiciously different from normality) do not take these costs and benefits into consideration. This may lead to situations where case A is ranked above case B, because it diverges more from normality, and yet case B if detected would be more rewarding in the sense that it has a

more favourable balance between inspection costs and detection benefits. The main goal of this paper is to propose a fraud detection methodology that takes the utility of detection into account when producing inspection ranks.

The paper main contribution is the proposal of a general approach to fraud detection in a context of limited inspection resources. This approach can be seen as a form of multi-strategy learning as it integrates several learning steps to meet the application requirements.

2 Utility-based Rankings

The basic assumption behind the ideas described in this paper is that fraudulent cases have different importance, different inspection costs and different payoffs. Another obvious assumption of our work is that inspection resources are limited. In this context, we claim that the support for the inspection decisions should be in the form of an utility ranking, i.e. a ranking where top positions are occupied by cases that if confirmed fraudulent will bring a higher reward.

Utility theory [von Neumann and Morgenstern, 1944] describes means for characterising the risk preferences and the actions taken by a rational decision maker given a certain probability model. This theory assumes we have knowledge about the wealth associated with each possible decision. Given this estimated wealth and a probability associated with each alternative action to be taken, the decision maker uses an utility function to map the wealth into an utility score. The action/decision with higher utility is then selected by the decision maker. Different utility functions exist in the literature that map a wealth value into an utility score (e.g. [Friedman and Sandow, 2011]).

Our proposal casts the decision of which cases to inspect in the context of fraud detection into this utility theory framework. In general terms our method is the following. Given a set of potential candidates for inspection we start by estimating their probability of being a fraud. This is a "classical" outlier detection task. Additionally, for each of these candidates we calculate their potential wealth as a function of estimates of their inspection cost and payoff if confirmed fraudulent. For each case there are two possible outcomes of inspection: either the case is confirmed as fraudulent or not. We calculate the wealth of each of these possible outcomes. Using an utility function we can then calculate the utility of both possible outcomes for each case. Together with the estimated probability associated to each outcome we finally reach an estimated utility score of the case. Our proposal is to rank the cases according to the decreasing value of these estimated utilities and use this ranking to guide the inspection phase. We thus propose to calculate the expected utility value of a case as follows:

$$E[U_i] = \hat{P}_i \cdot u(\hat{B}_i - \hat{C}_i) + (1 - \hat{P}_i) \cdot u(-\hat{C}_i) \qquad (1)$$

where $\hat{P}_i$ is the estimated probability of i being a fraud, $\hat{B}_i$ is the estimated benefit (payoff) of case i if confirmed fraudulent, $\hat{C}_i$ is the estimated inspection cost of case i, and $u(.)$ is an utility function.

In order to adapt this general formulation to our target applications we need to consider how to obtain the several estimates that are mentioned in Equation 1. The estimate of the probability that a case is a fraud can be obtained by any existing outlier detection method that is able to attach a probability of being outlier to its results. In Section 3 we will describe concrete systems that can be used to obtain these scores for any set of cases. With respect to the estimated inspection costs and payoffs ($\hat{C}_i$ and $\hat{B}_i$) the problem is slightly different. These values are clearly application dependent. In this context, we need some user-supervision to obtain these estimates. This can take two forms: either by providing domain knowledge that allows us to obtain these values, or by providing examples that we can use to learn a predictive model that is able to forecast them. The second form is probably more realistic in the sense that in most real-world fraud detection tasks we will have access to a historical record of past inspection activities. These past data can be used as a supervised data set, where for each inspected case we have information on both the inspection costs and also on the outcome of the inspection (either the payoff or the conclusion that it was not a fraud). Using this training set we can obtain two models: one that predicts the estimated inspection cost, and the other that forecasts the expected payoff. Both are supervised regression tasks. In summary, our proposal includes three learning tasks that together provide the necessary information to obtain the utility-based outlier rankings using Equation 1.

Algorithm 1 describes our proposed methodology in general terms. Its inputs are (1) the historical data set ($HistD$) with information on past inspection activities; and (2) the set of candidate cases for inspection ($InspCand$). The practical implementation of the ideas in Algorithm 1 requires some extra decisions. Namely, we need to decide which learning tools are to be included in Steps 1 and 2 of the algorithm, and also on the utility function to use. Note that our proposal is independent of this choice, only requiring that the outlier detection tool is able to attach a score in the interval $[0, 1]$ to each member of the set of candidate cases for inspection.

3 An Implementation of the Proposal

3.1 Predicting Costs and Benefits

The two tasks of predicting the inspection costs and the potential benefits of an inspection are "standard" supervised regression problems. In both cases we are trying to forecast a continuous variable as a function of a set of predictors that describe the case under consideration. This means that any standard regression tool would be capable of handling this task provided we have a training sample with both the values of the target variable and of the predictors. Still, the problems have some characteristics that may bias our choice of models to use. Namely, the data concerning the benefits will have a proportion of the learning sample with the value of zero on the target. These are the past cases that after inspection were tagged as non-fraudulent. This problem should not occur on the task of predicting the inspection costs as every inspection has some cost. Still, in both problems we can expect that we will have a rather diverse range of values, and this is actually one of the motivations for the work presented in this paper.

Algorithm 1 High-level description of the methodology.

procedure UOR($HistD, InspCand$)
  ▷ where $HistD = \{\langle x_1, \cdots, x_p, C, B\rangle\}$,
    $InspCand = \{\langle x_1, \cdots, x_p\rangle\}$,
    and $x_1, \cdots, x_p$ are variables describing each case

  ▷ *Step 1 - Train Cost and Benefit Prediction Models*
  $DS_C \leftarrow \{\langle x_1, \cdots, x_p, C\rangle \in HistD\}$
  $DS_B \leftarrow \{\langle x_1, \cdots, x_p, B\rangle \in HistD\}$
  $C_{model} \leftarrow RegressionTool(DS_C)$
  $B_{model} \leftarrow RegressionTool(DS_B)$

  ▷ *Step 2 - Obtain Outlier Probabilities for InspCand*
  $P \leftarrow OutlierProbEstimator(HistD, InspCand)$

  ▷ *Step 3 - Estimate Utilities*
  for all $i \in InspCand$ **do**
    $C_i \leftarrow Predict(C_{model}, i)$
    $B_i \leftarrow Predict(B_{model}, i)$
    $EU_i = P_i \cdot u(B_i - C_i) + (1 - P_i) \cdot u(-C_i)$
  end for

  ▷ *Step 4 - Obtain utility ranking (solution)*
  return $InspCand$ ranked by decreasing EU_i
end procedure

In our experiments with the proposed method we will use both regression trees and neural networks for illustrative purposes. These are two methodologies with a rather different approach and thus we expect this to be a good test of the robustness of the methodology to this choice.

3.2 Outlier Ranking

Our methodology also requires probabilities of being outlier to be obtained. There are many existing outlier detection tools that can be used to estimate these probabilities. Once again we have selected two particular representatives to illustrate our methodology in the experimental section.

The first approach we have selected is the method OR_h [Torgo, 2010]. This method obtains outlier rankings using a hierarchical agglomerative clustering algorithm. The overall motivation/idea of the method is that outliers should offer more resistance to being merged with large groups of "normal" cases and this should be reflected in the information of the merging process of this clustering algorithm. The OR_h method calculates the outlier score (a number in the interval $[0, 1]$) of each case as follows. For each merging step i involving two groups of cases ($g_{x,i}$ and $g_{y,i}$) the following value is calculated,

$$of_i(x) = \max\left(0, \frac{|g_{y,i}| - |g_{x,i}|}{|g_{y,i}| + |g_{x,i}|}\right) \quad (2)$$

where $g_{x,i}$ is the group to which x belongs, and $|g_{x,i}|$ is that group cardinality.

Each observation may be involved in several merges throughout the iterative process of the hierarchical clustering algorithm. Sometimes as members of the larger group, others as members of the smaller group. The final outlier score of each case is given by,

$$OR_h(x) = \max_i of_i(x) \quad (3)$$

In our experiments we use an implementation of this method available in the R package **DMwR** [Torgo, 2010].

The second method we have used is the LOF outlier ranking algorithm [Breunig *et al.*, 2000]. LOF obtains an outlier score for each case by estimating its degree of isolation with respect to its local neighbourhood. This method is based on the notion of local density of the observations. Cases in regions of very low density are considered outliers. The estimates of the densities are obtained with the distance between cases. The authors define a few concepts that drive the algorithm used to calculate the outlier score of each point. These are: the (1) concept of *core distance* of a point p that is defined as its distance to its k^{th} nearest neighbour; the (2) concept of *reachability distance* between the case p_1 and p_2 that is given by the maximum of the core distance of p_1 and the distance between both cases; and the (3) *local reachability distance* of a point that is inversely proportional to the average reachability distance of its k neighbours. The local outlier factor (LOF) of a case is calculated as a function of its local reachability distance.

In our experiments with LOF we have used a R implementation of this method available in package **DMwR** [Torgo, 2010]. Please note that the scores produced by LOF are not in the interval $[0, 1]$ so we have used a soft max scaling function to cast these values into this interval.

We should remark that neither the scores obtained by OR_h and LOF can be considered as "proper" probabilities. Still, they lead to values in a $[0, 1]$ interval and we can look at them as badly calibrated probabilities of being outlier. We should expect better results with more reliable estimates of these probabilities.

4 Experimental Evaluation

4.1 Artificially Generated Data

In this section we present a set of experiments[1] with artificially created data using an implementation of Algorithm 1. The goal of these experiments is to test the validity of some of our assumptions and also to observe the performance of the proposal on a controlled experiment.

We have generated a data set with the following properties. For easier visualisation we have described each case by two variables, x_1 and x_2. Two well-separated clusters of data points were generated, each containing some local outliers, though on one of them these outliers are more marked. The idea here is that any outlier ranking method would easily be able to rank these outliers on top positions, but the ones that are more marked should appear on higher ranking positions. For all data points we have artificially generated costs and benefits of their inspection. We have done so in a way that the cluster containing the less evident outliers brings a higher utility by having significantly higher payoffs for fraudulent

[1] All code and data available at `http://www.dcc.fc.up.pt/~ltorgo/IJCAI11`

cases, though with a higher inspection cost. Figure 1 shows the resulting data formed by 2200 randomly generated data points following the general guidelines just described.

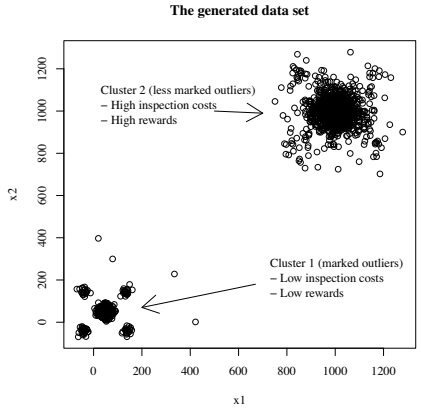

Figure 1: The Artificial Data Set.

The payoffs were set to zero for cases that we have considered non-fraudulent. We have chosen these cases to be the ones near the centre of the two clusters. More specifically, cases for which $(0 < x_1 < 100 \wedge 0 < x_2 < 100) \vee (900 < x_1 < 1100 \wedge 900 < x_2 < 1100)$ is true. Still, all cases, even non-fraudulent, have an inspection cost.

We have randomly selected 20% of these data to be the set for which we want to obtain an inspection priority ranking (i.e. a kind of test set). For this test set the information on the cost, payoff and utility was hidden from our algorithm. They are to be estimated by the regression models included in our methodology (lines 8-9 in Algorithm 1).

Looking at the characteristics of the above artificial problem we would want our algorithm to include the outliers near Cluster 2 as the top priority cases for inspection as these will bring the higher rewards, i.e. are more useful in terms of optimising the inspection resources. Moreover, our hypothesis is that most standard outlier ranking methods will fail to achieve this goal and will put the outliers nearer Cluster 1 on top positions as these are more obviously deviating from the main (local) bulk of data points.

Although this is an artificially created data set we have tried to include properties of real world applications. In effect, if we think for instance of the scenario of detecting tax frauds, we can see that within this high impact application there are also marked clusters of tax payers with rather different characteristics (e.g. due to different economic activities). It is also easy to imagine that for these different groups of tax payers the inspection costs and eventual payoffs, if confirmed fraudulent, can be rather diverse. For instance, on low income professionals we could imagine a smaller inspection cost due to the their simpler economic activity, and also a lower potential reward caused by the small monetary values involved in their activity (these are properties similar to points in Cluster 1). On the other hand, tax payers with a very high income (e.g. large companies) would probably involve a more complex (and thus costly) inspection phase, but if found fraudulent would also potentially bring a much higher payoff (Cluster 2 situation). In summary, we think this artificial scenario is a simplified version of some highly relevant real world applications of fraud detection constrained by limited inspection resources.

We have applied an implementation of our algorithm to this problem. We have tested it using different setups regarding the learning components. Namely, we have used both a regression tree and a neural network as supervised regression learners, and both the OR_h and LOF algorithms as outlier probability estimators. Figure 2 shows the results of two of the tested combinations: (1) regression trees with OR_h and linear utility ($u(w) = w$); and (2) neural nets with LOF and also linear utility. We have also tested the methodology with other combinations of these components, like for instance a power utility function ($u(w) = \frac{w^{1-k}-1}{1-k}$, with $k = 0.2$). The results with the other tested combinations were qualitatively similar to these so we omit them for space reasons.

Figure 2 includes information on the top 20 cases according to two different rankings. The first is obtained by looking only at the probabilities of being outlier produced by the outlier ranking method (OR_h in one case and LOF on the other). The second is the ranking obtained with our utility-based methodology. As we can observe in both setups the results provide clear evidence that our proposal is able to produce a ranking more adequate to the application goals, i.e. better optimise the inspection resources. In effect, while most of the top positions of the standard outlier rankers belong to Cluster 1, our method ranks at top positions mostly outliers of Cluster 2. These are the cases that provide a larger payoff. We have confirmed this by calculating the total balance of both rankings if we inspect their respective top 20 suggestions. The scores are shown in Table 1.

	Regr.Tree+OR_h+linearU		NNet+LOF+linearU	
	$UOR(OR_h)$	$\hat{P}(OR_h)$	$UOR(LOF)$	$\hat{P}(LOF)$
Costs	190 687.8	4 719.9	118 089.6	15 037.5
Benefits	423 566.6	5 140.7	260 907.3	26 172.1
Utility	232 878.8	420.7	142 817.7	11 134.7
% Util. Gain	55 243.9%		1 182.6%	

Table 1: Net Inspection Results.

The results shown in Table 1 reveal an overwhelming different in net results of the inspection activity if we follow either the advice of a standard outlier ranking method ($\hat{P}$) or our proposed utility-based outlier ranker (UOR).

4.2 Foreign Trade Transactions

In this section we describe an application of our methodology to a real world domain. This domain consists on trying to detect errors and/or frauds in foreign trade transaction reports. The data was provided by a national institute of statistics and consists of information regarding these reports that describe transactions of companies with foreign countries. Due to the impact that errors and/or frauds on this database can have on official statistics it is of high relevance to detect them. At the end of each month the institute faces the task of trying to detect these problems on the reported transactions that cover

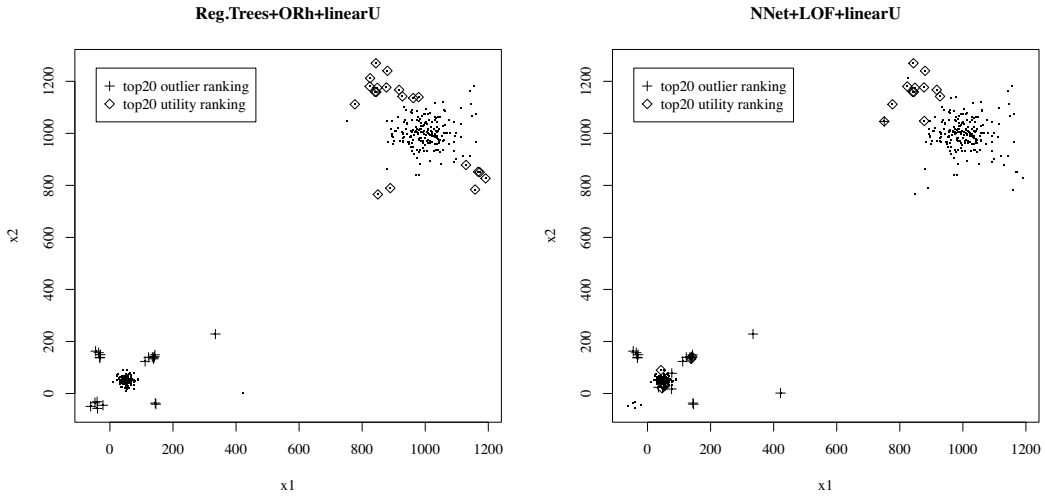

Figure 2: Results with the Artificial Data Set.

a wide range of rather diverse products. There are varying but limited human resources allocated for this task. In this context, it is of high importance to allocate these resources to the "right" reports. Given the diverse range of products the impact of error/frauds on official statistics is not uniform due to the amounts involved. This impact is a function of both the unit price of the product in question but also of the reported quantity involved in the transaction.

Among the information in the transaction reports, domain experts recommend the use of the cost/weight ratio as the key variable for finding suspicious transactions. According to these experts transactions of the same product should have similar cost/weight ratio. The available data set contains around 410 000 transaction reports concerning 8 months and covering a wide range of products. The information of past inspection activities on these transactions is rather limited. In effect, the sole information we have is a label on transactions that were found to be errors. We do not have any information on transactions that were inspected but found correct neither on the corrected values for transactions labelled as errors. This limited information creates serious difficulties for applying our methodology because we assume a relatively rich set of information on prior inspection activities.

According to domain experts the inspection cost of each report is roughly constant and was set to 150 Euros (approximately 2 hours of work). This means that we do not need a prediction model to estimate this cost. Regarding the benefits we do not have any information on past corrections so we can not build a training set to obtain a model for predicting benefits. However, we can try to estimate the benefit of inspecting a certain transaction report by confronting the values in the form with the expected values for transactions of the same product. The difference between the two can provide us with clues on the eventual corrections that are necessary to the report and thus on the benefit of the correction in monetary terms. For instance, if we have a transaction of 1000 Kg of product A with a reported cost/weight ratio of 10, and we observe that the typical cost/weight of this product is 4 then if we decide to inspect this transaction and confirm it as an error the benefit can be approximated by $(10-4)\times 1000$. In summary, in the application of our method to this problem we have used a constant value of $\hat{C}_i = 150$, and have estimated the benefit of a detection, $\hat{B}_i$, as $|c/w_i - \widetilde{c/w}| \times w_i$, where $\widetilde{c/w}$ is the median cost/weight of the past transactions of the same product, and w_i, c/w_i are the weight and cost/weight reported in transaction i, respectively.

Our experimental analysis follows the procedure used in the institute of analysing each month data independently. We also analyse each product transactions independently, given the rather diverse prices involved. Given the transactions of a product in a month we calculate the expected utility of inspecting each of the transactions using our method, according to the procedure explained above concerning the estimates of costs and benefits. Using the estimated utilities for all transactions of all products in the current month we obtain an utility-based ranking of that month transactions.

This ranking of the transactions of a month obtained using our methodology was compared to the ranking of the same transactions obtained using solely the probabilities of being outlier, i.e. not taking into account neither benefits nor costs. The comparison was carried out for different inspection effort levels. Namely, for inspecting the top $10\%, 15\%, \cdots, 30\%$ transactions of the month, according to each ranking method. So for inspection effort $x\%$ we compare the top $x\%$ cases according to the two rankings. This is done by calculating the net utility in the respective $x\%$ set of cases for each of the rankings. The net utility is given by the sum of the benefits of the cases in the set that have the error label set (i.e. the ones experts told us that are errors), subtracted by the total inspection cost (i.e. $150 \times \|set\|$). Comparing the net utility values of each of the two rankings we obtain % gain in utility of our proposal over the simple probability of outlier rankings. This comparison was carried out twice for each of the two meth-

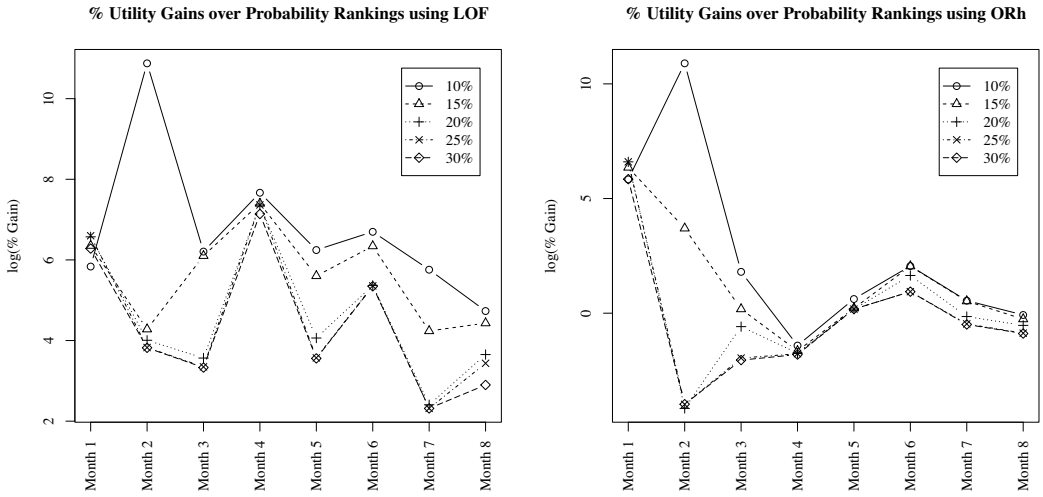

Figure 3: Results with the Foreign Trade Transactions Data Set.

ods that we considered to estimate these probabilities: LOF and OR_h. Figure 3 presents the results of these experiments. The lines on each graph represent the % utility gain of our method over the respective outlier probability ranking (results are shown on log scale for better visualisation), for different inspection effort levels. Once again we have observed a clear superiority of our rankings in terms of the net results of the inspection activities. In effect, in all setups the % utility gain is positive. We have also observed that our advantage tends to decrease for larger inspection efforts, which is expectable. Moreover, the advantage also holds across the two alternative outlier probability estimators.

5 Conclusions

We have presented a new methodology for supporting fraud detection activities. The main distinguishing feature of this method is its focus on trying to optimise the available inspection resources. This is achieved by producing an inspection ranking that is ordered by decreasing expected utility of the posterior inspection. We use the theoretical framework provided by Utility Theory to integrate several learning steps that obtain probabilities of being outlier and also forecast the expected inspection costs and resulting payoffs.

We have presented a series of experimental tests of our proposal with both artificially created data and a real world fraud detection application. These experiments provide clear evidence of the advantages of our proposal in terms of the optimisation of the available inspection resources.

Further work should focus on investigating the contributions of the different steps of our methodology as well as the exploration of different variants in its components.

Acknowledgments

This work is financially supported by FEDER funds through COMPETE program and by FCT national funds in the context of the project PTDC/EIA/68322/2006.

References

[Bolton and Hand, 2001] R.J. Bolton and D. J. Hand. Unsupervised profiling methods for fraud detection. In *Credit Scoring and Credit Control VII*, 2001.

[Breunig *et al.*, 2000] M. Breunig, H. Kriegel, R. Ng, and J. Sander. Lof: identifying density-based local outliers. In *ACM Int. Conf. on Management of Data*, pages 93–104, 2000.

[Fawcett and Provost, 1997] Tom Fawcett and Foster Provost. Adaptive fraud detection. *Data Mining and Knowledge Discovery*, 1(3):291–316, 1997.

[Friedman and Sandow, 2011] C. Friedman and S. Sandow. *Utility-based Learning from Data*. CRC Press, 2011.

[Ghosh and Reilly, 1994] S. Ghosh and D. Reilly. Credit card fraud detection with a neural-network. In *Proceedings of the 27th Annual Hawaii International Conference on System Science*, volume 3, Los Alamitos, CA, 1994.

[Hand, 2002] Richard J. Bolton ; David J. Hand. Statistical fraud detection: A review. *Statistical Science*, 17(3):235–255, 2002.

[Nigam *et al.*, 2000] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classication from labeled and unlabeled documents using em. *Machine Learning*, 39:103–134, 2000.

[Phua *et al.*, 2005] C. Phua, V. Lee, K. Smith, and R. Gayler. A comprehensive survey of data mining-based fraud detection research. *Artificial Intelligence Review*, (submitted), 2005.

[Torgo, 2010] L. Torgo. *Data Mining with R, learning with case studies*. CRC Press, 2010.

[von Neumann and Morgenstern, 1944] J. von Neumann and O. Morgenstern. *Theory of games and Economic Behavior*. Priceton University Press, 1944.