

Draft - comments welcomed

Scorecard construction with unbalanced class sizes

David J. Hand and Veronica Vinciotti
Department of Mathematics
Imperial College
London

{d.j.hand, v.vinciotti@ic.ac.uk}

Abstract:

A long-running issue in scorecard construction is how to handle dramatically unbalanced class sizes. This is important because, in many applications, the class sizes are very different. For example, it is common to find that 'bad' customers constitute less than 10% of the customer base and even more extreme situations often arise: Brause *et al* (1999) remark 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%,' while Hassibi (2000) comments that 'out of some 12 billion transactions made annually, approximately 10 million – or one out of every 1200 transactions – turn out to be fraudulent. Also, 0.04% (4 out of every 10,000) of all monthly active accounts are fraudulent.'

In coping with unbalanced classes, there are two issues to be considered. Firstly, what performance criterion is appropriate? And, secondly, how should the scorecard be constructed, and any parameters estimated, from such unbalanced data?

We look at each of these problems.

For the first problem, we illustrate the effect that marked lack of balance has on performance criteria, demonstrating how easy it is to be misled. The lack of balance means that simple error counts are inappropriate as performance criteria. Rather, misclassifications of customers from the smaller class must be regarded as more serious than the converse: different costs must be adopted for the two different kinds of misclassification. We examine some of the implications of this.

For the second, we examine both classical linear scorecards and more powerful k-nearest-neighbour nonparametric methods, such as are used in fraud detection. In the case of linear scorecards (and, more generally, for any simple parametric form) improved classification accuracy is achieved by focusing classification performance in particular parts of the data space, with the relevant parts being implied by the relative misclassification costs. We describe a new tool for constructing scorecards which takes this fact into account. In the case of k-nearest-neighbour methods, we draw attention to a phenomenon we believe has not previously been reported, and which has an important effect on choice of k . We illustrate both methods using a large data set of unsecured personal loan data.

Keywords: Credit scoring, unbalanced classes, classification performance

May 2002

1. Introduction

A long-running issue in scorecard construction is the issue of how to handle dramatically unbalanced class sizes. This is important because, in many applications, the class sizes are very different. For example, it is common to find that 'bad' customers constitute less than 10% of the customer base, and in mass promotion campaigns a response rate of 1% or less is common. Situations which are even more extreme arise in fraud detection (Bolton and Hand, 2002). Brause *et al* (1999) remark 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%,' while Hassibi (2000) comments that 'out of some 12 billion transactions made annually, approximately 10 million – or one out of every 1200 transactions – turn out to be fraudulent. Also, 0.04% (4 out of every 10,000) of all monthly active accounts are fraudulent.' A common problem with such situations is that, as we explain below, often the minimum number of misclassifications is achieved simply by assigning everyone to the larger class. Thus, for example, if 0.1% of a set of transactions are fraudulent, then if all transactions are taken as legitimate then only 0.1% of the transactions will be misclassified. Such a course of action is seldom acceptable.

For simplicity, in this paper we will restrict ourselves to the situation in which the aim is to predict which of just two classes the customer lies (or will lie) in. We assume that we have available a retrospective data set, consisting of information about a set of customers (e.g., application form data, data on past behaviour with a financial product, or bureau data) and that for these customers we know which of the two

classes they eventually fell into. Using this information, we aim to construct a scorecard which will allow us to predict the class of a new customer using their descriptive information. The way we will treat a new customer (the action we will take with them) will depend upon which class we predict them to lie in. Examples of different action pairs are (grant loan, do not grant loan), (treat as normal, investigate for possible fraud), (treat as normal, send warning letter), and so on. At this point it is useful to distinguish between (i) those action pairs for which one of the actions means that the final class of a customer is never known and (ii) those action pairs for which the classes of all customers eventually become known. An important example of the former is when the class pair is (default on loan, do not default) and the action pair is (grant loan, do not grant loan). In this case one never discovers whether those not granted loans will default or not. An example of the latter would be when the class pair relates to credit card delinquency, with the action pair being (send reminder letter, do not send letter). In this paper we will concentrate on the second of these two situations, assuming that the true classes of all individuals in our retrospective data set are known.

Let x represent the information describing a customer, and $s = s(x)$ the customer's score on some scorecard. This score represents a position on a continuum, for which, without loss of generality, we will take high scores as generally being associated with class 1 and low scores as generally being associated with class 0. The score, s , is thus (monotonically increasingly) related to $\hat{p}(1 | x)$, an estimate of the probability $p(1 | x)$ that someone with characteristic vector x will belong to class 1. Also, at this stage, we should note that we are assuming that the data are drawn (either entire populations of customers, or subsamples from these populations) in such a way that the proportion

of customers which belong to each class are unbiased estimates of the probability of belonging to that class. That is, we assume that the class proportions in the available data are not distorted away from the true priors. Later we will consider sampling from the available data in a way which will distort these proportions.

Given the estimate $\hat{p}(1|x)$, to assign customers to classes one could assign them to the class to which they are estimated as being most likely to belong. That is, one could adopt:

Rule 1: Assign to class 1 if $\hat{p}(1|x) > 0.5$ and to class 0 otherwise.

Now suppose that class 0 is much larger than class 1 - that the classes are *unbalanced*.

Then it can easily happen that $p(1|x) \ll 0.5$ and $\hat{p}(1|x) \ll 0.5$ for all vectors x .

Using Rule 1 would mean that no customers will be assigned to class 1. The 'classification threshold' 0.5 will minimise the overall number of customers misclassified (the number from class 0 assigned to class 1 plus the reverse) but may do this by the simple expedient of assigning all of the smaller class (class 1) to the larger class. If the smaller class represents potential frauds or potential bad risk customers, this is not at all what we want. The straightforward overall number of customers misclassified is thus an inappropriate measure of performance, so that Rule 1, which minimises this measure, is an inappropriate classification rule. As it happens, many other performance measures are used in retail banking, but most of them are also inappropriate. We describe such measures in Section 2.

The problem with Rule 1 arises from the failure to recognise that different types of misclassification carry different penalties. For example, assigning a fraudulent customer to the non-fraud class is more serious than the reverse: we wish to avoid this if at all possible, even if it means that we might misclassify some non-fraudulent customers as potentially fraudulent (and take some action, such as phoning them to see if their credit card has been stolen). If we suppose that misallocating a class 1 customer to class 0 is r times as serious as the reverse, and weight such misclassifications r times as heavily as the class-0-to-class-1 misclassifications (and also, for simplicity, assume that correct classifications incur no cost), then it is easy to show that comparing $\hat{p}(1 | x)$ with classification threshold $(1 + r)^{-1}$ minimises the overall weighted number of customers misclassified. That is, the overall weighted misclassification rate is minimised by

Rule 2: Assign to class 1 if $\hat{p}(1 | x) > (1 + r)^{-1}$ and to class 0 otherwise.

The term $(1 + r)^{-1}$ will be very different from the 0.5 used above if r is large.

Four broad strategies have been developed for implementing Rule 2.

- (1) Introduce misclassification costs at the scorecard classification stage, so that misclassifications of the smaller class are explicitly regarded as more serious than the reverse. This simply adopts rules of the form of Rule 2 directly, choosing the threshold $(1 + r)^{-1}$ appropriately.

- (2) Ignore the lack of balance, and use performance assessment measures which focus on the separability between the distributions of the estimates $\hat{p}(1 | x)$ for customers from the two classes. While this can be used for choosing which scorecards are likely to be effective, one still needs to choose a threshold in order to make actual classifications.
- (3) Preprocess the data to adjust the class sizes, either by subsampling from the larger class or by oversampling from the smaller class. Thus, for example, by applying Rule 1 to a data set in which the larger class has been reduced in size one can achieve the same results as applying Rule 2 to the unmodified data.
- (4) Introduce misclassification costs at the scorecard construction stage, so that again misclassifications of the smaller class are explicitly regarded as more serious than the reverse. In contrast to this, the traditional statistical approach has been to separate the model building and decision making phases (strategy 1 above). We demonstrate that such separation is not always a wise strategy, and that improved performance can be attained by taking the misclassification severities into account at the model building stage.

In Section 3, we review each of these in turn, and draw some conclusions about their relative merits. Before that, however, in Section 2, we examine scorecard performance measures, indicating how lack of balance impinges on them. Section 4 describes an illustrates a method for implementing strategy 4 which has the practical merits of the popular logistic regression approach, but which leads to improved scorecard performance.

2. Assessing scorecard performance

Logically, the choice of performance criterion precedes the selection, construction, and estimation of a scorecard. One measures the quality of a scorecard by seeing how effective it is at doing what one wants it to do. Only once one has determined how one will measure and compare performance can one rationally choose between alternatives or decide that a certain parameter setting is a good one. In view of this, one might expect it to be unusual to use one criterion for constructing the scorecard, and then another for assessing performance since it is possible that the two criteria might take their optima with very different scorecards. Benton (2002) illustrates the large differences which can arise between simple linear scorecards when different criteria are used to choose them. It is therefore surprising that, in almost all practical implementations of scorecards in the retail credit industry, the criteria used for constructing scorecards are different from those used for assessing performance. For example, likelihood is a common estimation criterion used when constructing scorecards, but likelihood is of no interest as a performance criterion. Indeed, likelihood is an overall measure of how well a model fits a data set and we show, in Sections 3.4 and 4, that this is not really relevant in many credit scoring contexts.

Broadly speaking, we can distinguish between two types of performance measure:

Type 1: Those which compare the score distribution of customers in class 0 with the score distribution of customers in class 1.

Type 2: Those which recognise that the aim is to carry out an action, with different actions being taken for those predicted to be in class 0 and those predicted to be in class 1.

Type 2 criteria are the more powerful of the two types, in the sense that they more properly reflect the use of the scorecard and its classification rule. However, they require not only the scorecard to have been constructed, but also a classification threshold to have been chosen. Now, clearly, to use the scorecard to make decisions (and hence to carry out actions) a classification threshold must have been chosen. However, it may not be clear what threshold to choose at the time the scorecard is being constructed. Indeed, the classification threshold may vary over the course of time, as economic conditions change. This means that circumstances do arise when one would like to be able to evaluate how effective is a scorecard without having an explicit classification threshold. Type 1 criteria are appropriate for such circumstances. Familiar examples of type 1 rules are the Gini coefficient, the Kolmogorov-Smirnov statistic, the information value, and the mean difference.

The *Gini coefficient*. This is usually defined as twice the area between a Receiver Operating Characteristic (ROC) curve and the diagonal of an ROC square (Hanley and McNeil, 1982). It is equivalent to the area under the ROC curve, the *AUC*. These measures can be given a natural interpretation, since they are also equivalent to the two sample *Mann-Whitney-Wilcoxon test statistic*, which estimates the probability that a randomly drawn member from class 0 will have a lower score than a randomly drawn member of class 1. All of these measures are equivalent in the sense that there are direct mathematical functions relating them: given any one, it is straightforward to

calculate the others. We see that their definitions make no reference to any classification threshold. In fact, it can be shown that these measures are equivalent to integrating the misclassification rate (see below) over the entire range of possible thresholds (Hand and Till, 2001; Hand, 2002), so that these measures do not, in a very real sense, require one to choose a classification threshold.

The *Kolmogorov-Smirnov test statistic*. This is an estimate of the maximum difference between the cumulative distribution functions of the scores of class 0 customers and class 1 customers. Hand (2002) shows that this measure is equivalent to choosing a classification threshold which minimises a particular weighted combination of the proportion of class 0 misclassified and the proportion of class 1 misclassified, using a classification threshold which is a function of the data. This is potentially very misleading, since the classification threshold should be chosen on the basis of the relative misclassification costs, as outlined above.

The *information value* is a symmetric Kullback-Leibler distance, and as such it integrates over all possible values of the score. This means that it ignores the classification threshold and using irrelevant information about the absolute value of the scores.

The *mean difference* score is the test statistic used in Student's *t*-test: the standardised difference between the means of the class 0 and class 1 score distributions. This ignores the classification threshold and the class priors, and also uses information about the absolute values of the scores. This information is typically irrelevant since the same action will be taken for all those customers scoring above the threshold,

regardless of their actual score. The same applies to those scoring below the threshold. The Gini coefficient and Kolmogorov-Smirnov statistic are invariant to nonlinear monotonic transformations of the score scale. The mean difference, however, is not. This means that before this measure can sensibly be used it is necessary to transform it to a standard form. A common way of doing this is to transform this scale so that the $\log(\text{odds})$ of being in class 1 (say) is a linear function of the transformed score. In fact, if this is done on with score scales which are comparable, scorecards can be conveniently compared by using the slope of the $\log(\text{odds})$ line.

Note that the Gini coefficient, the KS statistic, and the information value do not take account of the class priors. They simply compare the distributions for the class 0 and class 1 points, regardless of how many there are of each type. The same is true of the mean difference, except that the two classes may be weighted differentially when calculating the common standard deviation by which to standardise the mean difference. In particular, this will mean that the estimate of common standard deviation will be similar to the standard deviation of the larger class, and influenced relatively little by the smaller class. This raises a more general point, described in Section 3.4: scorecards of certain types will be biased by failing to accord proper weight to the smaller class during construction. Since type 1 measures ignore the class sizes, they will not be influenced by lack of balance. That is, they can be calculated and used as estimation and selection criteria with no adjustment for the fact that one class is much smaller than the other (with the possible exception of the mean difference measure).

Type 2 measures use only the true classes of the customers and whether their score is above or below the classification threshold t . Hand (1997, Chapter 8) discusses such measures in detail. A common simplification is to suppose that correct classifications incur no cost (because the appropriate action is taken), but that incorrect classifications of class 0 (to class 1) incur a cost c_0 (due to taking the action corresponding to class 1) while incorrect classifications of class 1 incur a cost c_1 .

Then the appropriate performance measure to use is the overall misclassification cost, defined as $n_0c_0 + n_1c_1$, where n_k is the number of class k points which are misclassified. Various other aspects of the ‘true-by-predicted’ table of class cross-classifications are sometimes used, including sensitivity and specificity (proportions of the true class correctly classified), and positive and negative predicted value (proportions of those classified which turn out to be correct). Definitions are given in Hand, 1997). These last four terms are also used in epidemiological contexts, but elsewhere other terms are used for the same concept (e.g. precision and recall in data mining). However, one looks at the ‘true-by-predicted’ cross-classification table, it involves three degrees of freedom, which need to be reduced to one to yield an overall measure which can be used to compare scorecards. Overall cost and misclassification rate (which assumes the misclassification cost to be equal) are two ways of doing this, but other, more exotic but less theoretically justified methods have also been developed (e.g. the geometric mean of precision and recall, or the geometric mean of sensitivity and specificity). Given that the aim of the classification is to choose which of two possible actions to take, and that different costs are incurred if inappropriate actions are taken, then the overall cost seems the most appropriate measure. This is developed in detail in Hand (2002).

Overall misclassification cost can be reinterpreted as a combination of the specificity and sensitivity, weighted appropriately by class sizes and the relative misclassification costs for members of the two classes. As such, this measure takes proper account of balance or lack of balance. Other ways of summarising the true-by-predicted table may not do this. That is, they may handle lack of balance inappropriately.

Type 1 measures are equivalent to aggregating values of particular type 2 measures over all possible choices of the classification threshold (see, for example, Hand, 2002). But this is clearly unrealistic. Different choices of the classification threshold correspond to different values for the relative costs of misclassification, and aggregating over all values is equivalent to an assertion that one has no idea at all what might be an appropriate measure. This is seldom the case. Typically one knows that certain values are possible and others not. Adams and Hand (1999) explore this, and define a measure which permits one to use information about likely values for the relative sizes of misclassification costs.

3. Strategies for unbalanced classes

3.1 Using costs at the classification stage

In the introductory section, we described the most obvious approach to handling unbalanced classes. This is a two stage approach. Stage 1 involves estimating $p(1|x)$ using all of the available data. For example, linear and logistic regression and tree classifiers are particularly common, but other methods include neural networks, tree classifiers, and so on. Linear and logistic regression methods have the property

that the resulting scores are simply weighted sums of the raw customer characteristics (perhaps partitioned, or combined in some way - see Hand and Adams (2000), for an example), which is often desirable in consumer credit applications.

Stage 2 involves comparing the estimate $\hat{p}(1 | x)$ with a threshold. Ideally, the threshold will reflect some measure of the relative severity of the two kinds of misclassification.

Many authors have followed this two stage procedure. It is perhaps the most natural, from a traditional perspective, which often regards such problems as two stage processes: estimate the distributions involved (model building) and then make the decision (by comparing the model with a threshold). However, despite the simplicity and popularity of this approach, it is often not ideal. In particular, if the estimate $\hat{p}(1 | x)$ is misspecified in some way (for example, it is based on a parametric form which does not properly reflect the true distributions), then it may lead to a decision surface which is a poor approximation to that for the desired threshold. We discuss the implications of this, and how to avoid it, in Sections 3.4 and 4. For now, however, a simple example will illustrate.

In standard linear discriminant analysis the assumed common covariance is estimated as a weighted average of estimates of the two within-class covariance matrices. The weights are normally taken to be the observed class sizes in the data. If one class is much larger than the other, then the estimate will be biased towards that class. In linear discriminant analysis, the decision surface is taken to be linear, and such surfaces will have the same orientation for all classification thresholds. This

orientation is a function of the vector difference between the sample centroids of the two classes and the assumed common covariance matrix. If the estimate of the latter is determined essentially by the larger class, then it may lead to a suboptimal decision surface for unbalanced cases.

3.2 Separability criteria based on within class score distributions

As described in Section 2, many measures in common use adopt this approach. Unfortunately, since the entire purpose of the classification is to take some action, such measures are not ideal. They aggregate in some way over all possible threshold (Hand, 2002). The practical consequence is that such measures may lead to the choice of a rule which is globally optimal in some average sense, but in fact performs poorly for the actual situation facing one. Adams and Hand (1999) describe strategies for taking into account information on likely threshold values, even if one cannot assert these values with precision.

In fact, the use of such global separability measures is widespread. Every retail bank and credit rating agency uses such measures, despite their disadvantages. An example for unbalanced data is given in Ling and Li (1998), who study a marketing application in which as little as 1% of the population responds to a promotion. If those likely to respond can be identified *a priori*, then a more targeted and hence cost-effective promotions strategy can be adopted. If the scores, s , are categorised into groups, $s_i, i = 1, \dots, g$, then Ling and Li (1998) use a measure of separability equivalent to a weighted sum of the estimated probability of being in class 1 in each of the groups:

$$\sum_i w_i p(1 | s_i).$$

3.3 Preprocessing the data

Many studies adopt the strategy of preprocessing the data to (roughly) equalise the numbers of elements in the two classes - to achieve better balance. Kubat and Matwin (1997), for example, preprocessed the data by removing unnecessary instances from the majority class. Isolated points from the majority class in regions dense with points from the other class, and examples which are redundant in the sense that their removal does not affect the decision surface, or those that are close to the decision boundary can all be considered as candidates for removal. The ideas parallel those developed some two decades previously, in attempts to speed up the processing time of nearest neighbour classification rules (see, for example, Hart, 1968; Gates, 1972; Hand and Batchelor, 1978).

An *et al* (2001) adopted the opposite approach. Rather than subsampling the larger class, they experimented with duplicating the elements of the smaller class (so that this class comprised from 4% to 50% of the training data). Lee (1999, 2000), also duplicated elements of the smaller class, but added small random perturbations to the replicated points.

Chan and Stolfo (1998) studied a credit card fraud data set, which had about 20% in the smaller (fraudulent) class. This is exceptionally large for the proportion of fraudulent customers in a data set, and arises because a pre-processing stage has been applied which eliminated many of those thought very unlikely to be fraudulent. Chan and Stolfo tackled the lack of balance by randomly partitioning the larger set into four

non-overlapping samples, and combining each of these with the smaller set, to yield four smaller data sets with equal numbers from each class. The four resulting classification rules were merged to yield a meta-classifier. This might not be a very effective strategy when the imbalance is marked, or if very few points are available from the smaller class.

Many of the subsampling or oversampling procedures are rather ad hoc. Elkan (2001) describes what sampling fractions are appropriate for given cost ratios.

Rule 2 can be alternatively written as: assign a customer with characteristic vector x to class 1 if

$$\hat{p}(x|1)\hat{p}(1)/\hat{p}(x|0)\hat{p}(0) > 1/r \quad (1)$$

and to class 0 otherwise, where $\hat{p}(x|k)$ is the estimated probability that a customer from class k will have characteristic vector x , and $\hat{p}(k)$ is the estimated overall probability of belonging to class k .

(1) is equivalent to

$$\hat{p}(x|1)/\hat{p}(x|0) > \hat{p}(0)/r\hat{p}(1) \quad (2)$$

Ideally, subsampling would reduce the class 0 prior by a factor of $1/r$, so that the new class priors become $\pi(k)$, $k = 0,1$, estimated by $\hat{\pi}(k)$, the proportion of the (sampled) data set which belong to class k . Expression (2) then becomes

$$\hat{p}(x|1)/\hat{p}(x|0) > \hat{\pi}(0)/\hat{\pi}(1) \quad (3)$$

so that the classification rule is simply: assign a customer with characteristic vector x to class 1 if

$$\frac{\hat{p}(x|1)\hat{\pi}(1)}{\hat{p}(x|0)\hat{\pi}(0)} = \frac{\hat{p}(1|x)}{\hat{p}(0|x)} > 1 \quad (4)$$

and to class 0 otherwise, where the $\hat{p}(k|x)$ are based on the sampled data. This is, of course, equivalent to Rule 1, but using the $\hat{p}(k|x)$ in place of the raw (unsampled) data, so that an appropriate threshold is used.

A similar derivation applies if class 1 is oversampled, rather than class 0 subsampled.

The derivation of (4) assumed that the sampling fraction was $1/r$, this being the fraction which is appropriate to balance the relative severities of the two kinds of misclassification. If a different sampling fraction is used then a poor rule could result. For example, many authors simply try roughly to equalise the sizes of the two classes. This confounds differences between class sizes with the relative severities of the two kinds of misclassification. The two need have no relationship at all.

The sampling approach (assuming the correct sampling fraction is used) has the merit that it focuses attention on the correct decision surface. That is, it is equivalent to using the optimal threshold $(1+r)^{-1}$ of Rule 2, rather than the inappropriate threshold 0.5 of Rule 1. However, one might have doubts about the subsampling strategy, on the grounds that it sacrifices information. Likewise, the oversampling strategy either fails to model the variation of the smaller class properly (if the data from this class are simply replicated) or attempts to model this variation, but in a way which is not proven to be correct (by perturbing the replicates). Section 3.4 describes an alternative approach which sidesteps these problems.

3.4 Using costs when building the scorecard

The strategy described in Section 3.1 is based on the assumption that the relative misclassification severities, equivalently the particular threshold to use in the classification rule, should not affect the estimate $\hat{p}(1|x)$. This is a reasonable assumption if one believes that the model form underlying the estimate $\hat{p}(1|x)$ is sufficiently flexible to include the true distributions. For example, if one believes that the contours of the function $p(1|x)$ really are linear in the raw characteristics, then linear and logistic regression models are appropriate to consider. Of course, the fact is that one will seldom have such confidence, although one might believe that a given parametric model form provides a reasonable approximation. In contrast, nonparametric approaches, such as kernel and nearest neighbour methods, do not restrict the model (indeed, subject to certain regularity conditions, they can be shown to be able to model any distribution, at least asymptotically). However, in the credit scoring context, there is a premium on simplicity. Often it is necessary to explain the reasoning behind the rules to fairly non-numerate people, and often there are legal requirements that one must be able to indicate on what grounds a decision has been made. Such considerations lead to an emphasis on simple parametric models (see Adams and Hand, 1999).

In these circumstances, model fitting procedures typically aggregate the quality of fit over the entire data space. For example, least squares regression is based on a criterion which combines the sum of squared residuals from all data points. Likewise, maximum likelihood methods combine the contribution to the likelihood from all the observed data points. Such aggregation will yield a model which is the best overall

model, where the meaning of ‘best’ depends on the particular criterion chosen - sum of squares, likelihood, etc. However, since they do aggregate over all data points - over the entire data space - they combine the accuracy of the model in the particular regions of interest (those given by the threshold) with all other regions (those far from the threshold). In particular, it is entirely possible that the fit in the region of interest is not very good, even though the overall average fit is the best that can be achieved. It means that a better local model, in the region which matters, might be possible. Put another way, it means that the relative severities of the two kinds of misclassification should be taken into account when the model is constructed. These ideas are described in more detail in Hand and Vinciotti (2002) and are illustrated in Section 4 below.

More generally than the particular model we have developed, several authors have explored the use of relative misclassification penalties when constructing classifiers, including Pazzani *et al* (1994), Turney (1995), Cardie and Howe (1997), Bradford *et al* (1998), Fan *et al* (1999), Domingos (1999), Verlopoulos *et al* (1999), Wan *et al* (1999), and Ting (2000).

4. Local scorecard models

Hand and Vinciotti (2002) give an example in which the contours of $p(1|x)$ are linear, but not parallel (the support of $p(x)$ is specified as zero in regions where different contours may cross, so that there are no conflicts). Logistic regression assumes parallel linear contours. In effect, such a model ‘averages’ the non-parallel contours of the example over the entire data space. If, by accident, the particular

contour corresponding to the threshold $(1+r)^{-1}$ is parallel to this ‘average’ contour, then the model will yield good predictions. On the other hand, if the contour of interest is not parallel to this ‘average’ contour then the predictions could be poor. In the case when one of the classes is very small, the data points from this class may lie in a relatively small region of the data space. If this happens, the aggregation process when a global parametric model is fitted will yield a model which has greatest accuracy in the vicinity of the data points from the smaller class. This will generally not correspond to threshold values which weight the relative misclassifications appropriately, so that the effect will be more marked in unbalanced situations.

If the problem with the standard models is that it aggregates over all data points, ‘averaging’ over all the different contours, then one can ease the problem by focusing attention around the contour which matters. Data far from this contour are at best irrelevant, and at worst misleading, leading to a poor estimate. Of course, one cannot take this principle too far. Thus one might weight the data so that points close to the relevant contour are weighted more heavily. The problem is, of course, that one cannot identify the relevant data because one does not know the position of the contour. This suggests an iterative, or at least several-stage process, in which one uses an estimate of the relevant contour to provide information on the weights, which in turn leads to an improved estimate, and so on. Hand and Vinciotti (2002) describe such an approach using a modified likelihood function, and we use method in the examples below, using real credit data sets.

The appropriate performance measure to use here is the overall misclassification cost, $n_0c_0 + n_1c_1$, defined above, where we have taken $c_0 + c_1 = 1$. For a given cost pair,

this measure was calculated for both the standard (global) logistic regression model and the local logistic model described above. As the cost varies (as one chooses different costs) so, of course, different contours become the most important contour. Thus, for both global and local models, a threshold will be chosen to match the costs: this threshold will be $(1+r)^{-1} = (1+c_1/c_0)^{-1} = c_0$. For global logistic models the same probability estimate $\hat{p}(1|x)$ will be used for all costs. In contrast, for the local model different estimates will be used - estimates which are tuned to the cost. To compare the two models, we used the difference between the global and local costs. A positive value of the (global-local) difference means that the global model has greater cost - that the local model yields superior cost weighted classifications.

Figures 1 to 3 show the values of global-local cost for three examples. As can be seen, in all cases, over the entire range of relative misclassification costs, the local method usually yields a smaller overall cost: the local method is usually, though not always, superior.

Example 1: These data were supplied by a major UK bank. They consist of 21618 unsecured personal loans with a 24-month term, collected over the two year period January 1995 to December 1996. An account is defined as bad if it is at least three months in arrears. With this definition, 11% of customers turn out to be bad. 16 variables describe the application for the loan.

Example 2: These data were supplied by a major UK credit card company. The aim of the analysis was to predict the future behaviour of a customer based on their

previous behaviour. There were 772 observations on 8 variables, with 9% of the data in class 1 and 91% in class 0.

Example 3: These data were supplied by a major UK bank. They describe customers who have defaulted on a loan in some well-defined sense and from whom the bank is trying to recover the loan. A bad account is defined as one that has spent more than a month in this “collections” state. The data consists of 6811 observations on 11 variables. 9% of the data are in the smaller class.

In fact, further data were also available for examples 2 and 3 above, so we also explored the effectiveness of the local model in more balanced situations.

Example 4: These data arise from the same situation as *Example 2*, but consist of 1490 observations on 8 predictor variables, with 47% of the customers in class 1.

Example 5: These data arise from the same situation as *Example 3*, but consist of 10102 observations on 11 variables, with 39% of the customers being in class 1.

In both of these (more balanced) cases, the local model is never worse than the global model.

Figure 1: Global-local costs for Example 1.

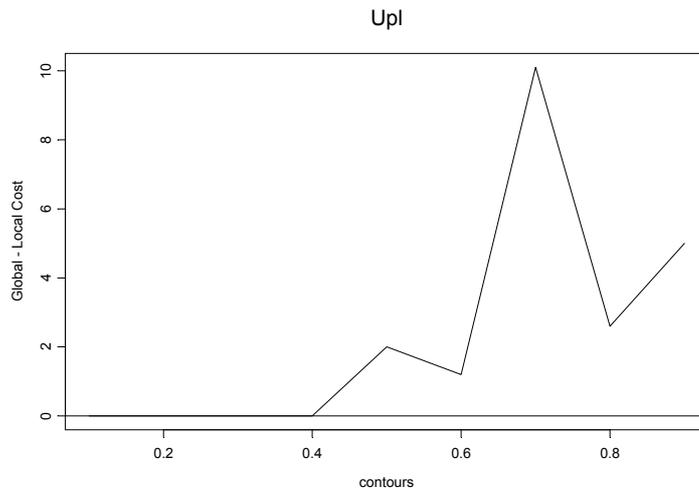


Figure 2: Global-local costs for Example 2.

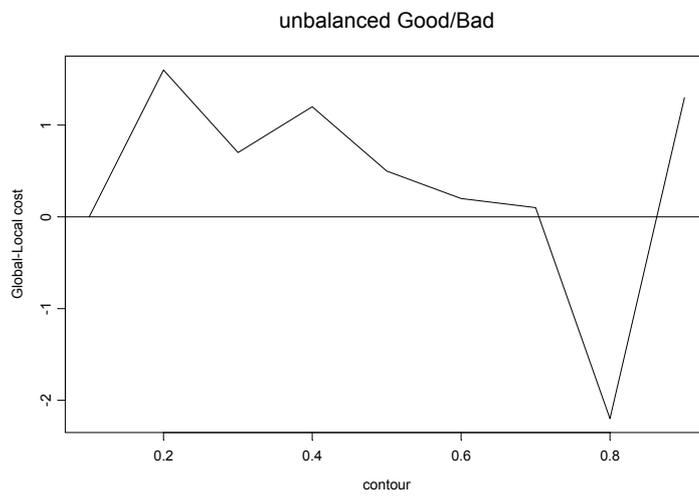


Figure 3: Global-local costs for Example 3.

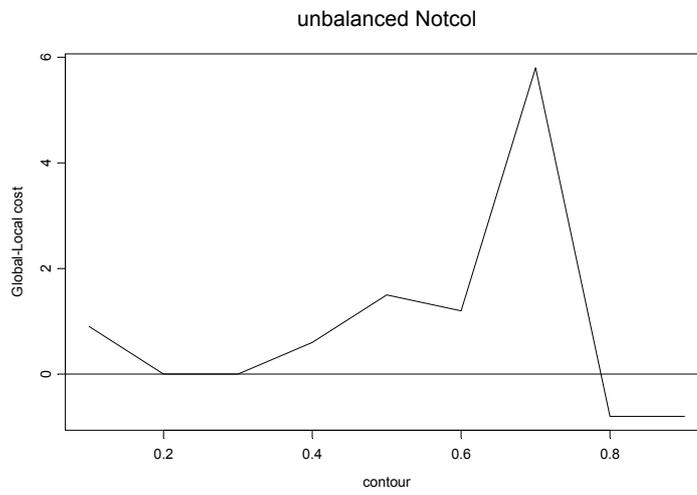


Figure 4: Global-local costs for Example 4.

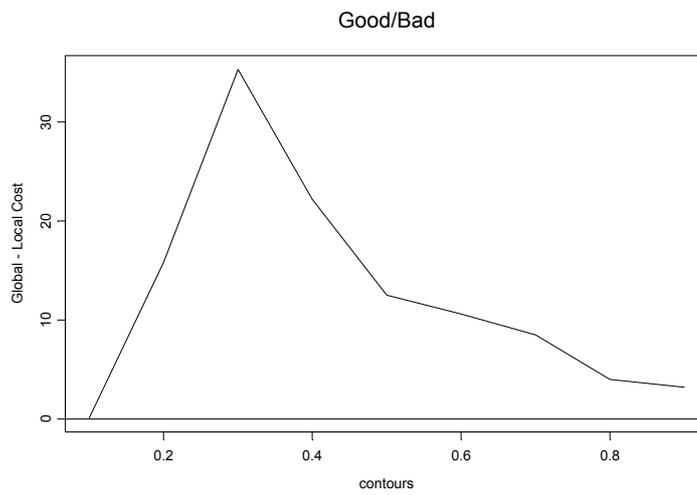
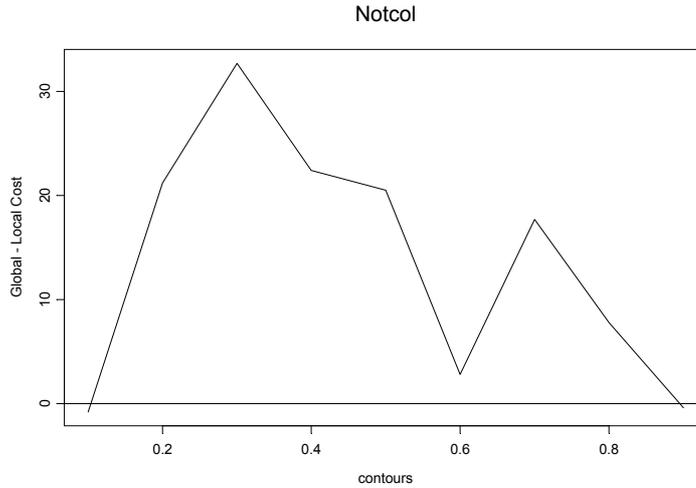


Figure 5: Global-local costs for Example 5.



5. Conclusion

Unbalanced data sets, those in which one of the (two) classes is much larger than the other, are common in retail banking applications, and several strategies have been proposed for building scorecards using such data sets. The most straightforward approach is simply to adjust the threshold with which $\hat{p}(1|x)$ is compared. If $\hat{p}(1|x)$ is based on a model which is thought to properly reflect the true probability structure $p(1|x)$, then this is fine. However, often, especially in retail banking applications, a simple form is adopted for the estimate, and it is difficult to argue that this is likely to properly reflect the truth. In this case, the probability estimate $\hat{p}(1|x)$ is obtained as an aggregate value over the entire data set, so that it may not yield a very good estimate for any particular value. In particular, it may not yield an accurate estimate of the contour of $p(1|x) = t$ which is to be used for classification.

Another class of methods is based on selectively sampling from the two classes, either to reduce the size of the larger class or to increase the size of the smaller. Often the

sampling fraction (which is larger than unity in the second case) is taken to be such as to yield approximately equal class sizes. This, however, is unlikely to be the optimal sampling fraction. If this method is adopted, then sampling should be such as to yield a class size ratio determined by the relative costs of misclassification for customers from the two classes.

The sampling approach is equivalent to adjusting the relative misclassification costs of customers from the two classes. This is easily seen from (2), which may be rewritten as

$$\hat{p}(x | 1)/\hat{p}(x | 0) > c_0\hat{p}(0)/c_1\hat{p}(1)$$

From this, we see that artificially distorting the $\hat{p}(k)$ yields an effect equivalent to adjusting the c_k .

Even if an optimal sampling fraction is chosen, the sampling methods leave one with the suspicion that something better could be done. After all, subsampling appears to discard information, while oversampling either ignores natural variability or artificially introduces it. In any case, just as with the simple method based on adjusting the threshold, sampling methods are global. They aggregate information from the entire data set and do not concentrate attention where it matters.

The final strategy is to take account of the misclassification costs - of which contour matters - when the probability estimates $\hat{p}(1 | x)$ are made. In particular, we describe such an approach which is based on logistic regression, and so preserves the simple linear form of such models. This method then concentrates estimation power in the region of this contour, so that irrelevant contours, which may merit a model with

completely different parameters, albeit of the same form, do not influence the estimate. This strategy is appropriate whether or not the classes are unbalanced, though it may be particularly pertinent in the unbalanced case. Our empirical investigations show that this method generally improves on straightforward logistic regression.

References

Adams N.M. and Hand D.J. (1999) Comparing classifiers when the misallocation costs are uncertain. *Pattern Recognition*, **32**, 1139-1147.

An A., Cercone N., and Huang X. (2001) A case study for learning from imbalanced data sets. *Advances in Artificial Intelligence Proceedings of the 14th Conference of the Canadian Society for Computational Studies of Intelligence*, 1-15.

Benton T. (2002) *Theoretical and empirical models*. Unpublished PhD thesis, Department of Mathematics, Imperial College, London.

Bolton R.J. and Hand D.J. (2002) Statistical fraud detection: a review. To appear in *Statistical Science*.

Bradford J., Kunz C., Kohavi R., Brunk C., and Brodley C.E. (1998) Pruning decision trees with misclassification costs. *Proceedings of the Tenth European Conference on Machine Learning*, 131-136.

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*.

Cardie C. and Howe N. (1997) Improving minority class prediction using case-specific feature weights. *Proceedings of the Fourteenth International Conference on Machine Learning*, 57-65.

Chan P.K. and Stolfo S.J. (1998) Towards scalable learning with non-uniform class and cost distributions: a case study in credit card fraud detection. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, 164-168.

Domingos P. (1999) Metacost: a general method for making classifiers cost sensitive. *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining*, 155-164.

Fan W., Stolfo S.J., Zhnag J., and Chan P.K. (1999) AdaCost: misclassification cost-sensitive boosting. *Proceedings of the Sixteenth International Conference on Machine Learning*, 99-105.

Gates G.W. (1972) The reduced nearest neighbour rule. *IEEE Transactions on Information Theory*, **18**, 431.

- Hand D.J. (2002) Measuring scorecard performance in retail credit applications. Technical Report, Department of Mathematics, Imperial College, London.
- Hand D.J. and Adams N.M. (2000) Defining attributes for scorecard construction. *Journal of Applied Statistics*, **27**, 527-540.
- Hand D.J. and Batchelor B.G. (1978) An edited condensed nearest neighbour rule. *Information Sciences*, **14**, 171-180.
- Hand D.J. and Till R.J. (2001) A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, **45**, 171-186.
- Hand D.J. and Vinciotti V. (2002) Local versus Global Models for Classification Problems. Fitting Models where it Matters. Technical Report, Department of Mathematics, Imperial College.
- Hanley J.A. and McNeil B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29-36.
- Hart P.E. (1968) The condensed nearest neighbour rule. *IEEE Transactions on Information Theory*, **14**, 515-516.
- 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.

Kononenko I. and Bratko I. (1991) Information-based evaluation criterion for classifier's performance. *Machine Learning*, **6**, 67-80.

Kubat M. and Matwin S. (1997) Addressing the curse of imbalanced data sets: one-sided sampling. *Proceedings of the Fourteenth International Conference on Machine Learning*, 179-186.

Kubat M., Holte R., and Matwin S. (1997) Learning when negative examples abound. *Proceedings of the 9th European Conference on Machine Learning, ECML'97, Prague*.

Lee S.S. (1999) Regularization in skewed binary classification. *Computational Statistics*, **14**, 277-292.

Lee S.S. (2000) Noisy replication in skewed binary classification. *Computational Statistics and Data Analysis*, **34**, 165-191.

Lewis D. and Gale W. (1994) Training text classifiers by uncertainty sampling. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Pazzani M., Merz C., Murphy P., Ali K., Hume T., and Brunk C. (1994) Reducing misclassification costs. *Proceedings of the Eleventh International Conference on Machine Learning*, 217-225.

Ting K.M. (2000) Cost-sensitive classification using decision trees, boosting, and MetaCost. *Workshop on Cost-Sensitive Learning at the Seventeenth International Conference on Machine Learning*, Stanford, USA.

Turney P.D. (1995) Cost-sensitive classification: empirical evaluation of a hybrid genetic decision tree induction algorithm. *Journal of Artificial Intelligence Research*, **2**, 369-409.

Veropoulos K., Campbell C., and Cristianini N. (1999) Controlling the sensitivity of support vector machines. *Proceedings of the Sixteenth International Conference on Artificial Intelligence*.

Weiss G.M. and Provost F. (2001) The effect of class distribution on classifier learning. *Technical Report ML-TR-43, Department of Computer Science, Rutgers University*.