Available at www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

Pattern Recognition Letters 24 (2003) 1555–1562

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Choosing k for two-class nearest neighbour classifiers with unbalanced classes

David J. Hand *, Veronica Vinciotti

Department of Mathematics, Imperial College, 180 Queen's Gate, London SW7 2BZ, UK

Received 11 July 2002; received in revised form 20 November 2002

Abstract

Supervised classification problems in which the class sizes are very different are common. In such cases, nearest neighbour classifiers exhibit a non-monotonic relationship between the number of nearest neighbours and misclassification rate of each of the two classes separately.

Keywords: Nearest neighbour methods; Classification; Choice of k; Unbalanced classes

1. Introduction

We assume we have available a design (or training) set consisting of measurement vectors and known class membership labels for a set of n cases, and the aim is to use these data to construct a rule to assign new cases to their correct class when only their measurement vector is observed. Many tools are available for constructing such assignment rules (see, for example, Hand (1997), McLachlan (1992), Ripley (1996), Webb (1999)). In this article we are concerned with the k-nearest neighbour method applied in two-class problems where one of the classes is much larger than the other. We begin by outlining some background on unbalanced class problems and on nearest neighbour methods.

In many practical problems, the class sizes are very different. Denoting the probability that a randomly chosen case will belong to class i by π_i, $i = 0, 1$, we will assume (w.l.g.) that $\pi_1 \gg \pi_0$. Examples of such situations arise in medical screening, where generally the disease class is much smaller than the non-disease class, in credit scoring, where the bad risk customers often form less than 10% of the applicant base, and in fraud detection and money laundering, again where the 'bad' class is much smaller than the 'good' class. For example, in a study of credit card transactions, Brause et al. (1999) say that '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%,' and Hassibi (2000) remarks 'Out of some 12 billion transactions made

* Corresponding author. Tel.: +207-594-8521; fax: +207-594-8561.

E-mail addresses: d.j.hand@ic.ac.uk (D.J. Hand), v.vinciotti@ic.ac.uk (V. Vinciotti).

0167-8655/03/$ - see front matter
doi:10.1016/S0167-8655(02)00394-X

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.'

We can distinguish two situations: those in which the classes are perfectly separable, and those in which they are not. 'Perfectly separable' here means that the space $\boldsymbol{X}$ in which the measurement vectors lie can be partitioned into two (each not necessarily connected) regions, such that $p(\boldsymbol{x}|0)$, the probability density function or mass function for class 0, is zero in one region and $p(\boldsymbol{x}|1)$ is zero in the other. This means that points at any given $\boldsymbol{x}$ can come from only one possible class, and all we need to do is to characterise the regions (or the decision surface separating them). If the classes are perfectly separable, then the disproportionate class sizes causes no problem in principle. Of course, this does not mean that finding the decision surface in the space of measurement vectors, $\boldsymbol{X}$, may not be difficult in practice. The machine learning community has often worked from the basis that the classes are perfectly separable (even going so far as to describe situations where the $p(\boldsymbol{x}|i)$ distributions overlap as 'noisy'). In contrast, the statistical community has usually seen the problem as one in which the probability distributions $p(\boldsymbol{x}|i)$ overlap. Hand (1997, Section 11.7) speculates that this may be because the computer science and machine learning community has often worked on artificial or man-made problems (e.g., character recognition), in contrast to the statistical community. In practice, real problems where the classes are perfectly separable are rare.

A simplistic classification rule will assign a new point to class 1 if

$$p(1|\boldsymbol{x}) > 1/2 \tag{1}$$

and to class 0 otherwise. Here $p(i|\boldsymbol{x})$ is the probability that a point with measurement vector $\boldsymbol{x}$ will belong to class i, and in practice the $p(i|\boldsymbol{x})$ will be estimated from the design set. However, such a rule makes the implicit assumption that the two different kinds of misclassification (classifying a class 0 case to class 1 and vice versa) are equally serious. Hand (1997) has argued that this is very rarely an appropriate assumption. The inappropriateness of such an assumption is easily illustrated in the extreme situation in which $p(1,\boldsymbol{x}) > p(0,\boldsymbol{x})$ for all $\boldsymbol{x}$, where $p(i,\boldsymbol{x})$ is the joint probability of i and $\boldsymbol{x}$. This means that it is more likely that an observation with measurement vector $\boldsymbol{x}$ will have come from class 1 *for all* $\boldsymbol{x}$. In such a situation, rule (1) would assign all points to class 1. This would not be appropriate in medical screening or identifying fraudulent credit card transactions, since such a rule would simply assign everyone to the disease-free or non-fraudulent classes respectively, on the grounds that this minimises the overall number of misclassifications.

Instead, it is necessary to weight the two types of misclassification according to their relative severities, or costs. Let the cost of misclassifying a class i case be c_i. Now, if points at $\boldsymbol{x}$ are assigned to class 1, the loss at $\boldsymbol{x}$ is $c_0 p(0|\boldsymbol{x})$. Similarly, if points at $\boldsymbol{x}$ are assigned to class 0, the loss at $\boldsymbol{x}$ is $c_1 p(1|\boldsymbol{x})$. The minimum loss at $\boldsymbol{x}$ is thus achieved by assigning points at $\boldsymbol{x}$ to class 1 if $c_0 p(0|\boldsymbol{x}) < c_1 p(1|\boldsymbol{x})$ and to class 0 otherwise. This is equivalent to the condition

$$p(1|\boldsymbol{x}) > c_0/(c_0 + c_1). \tag{2}$$

It follows that the overall loss, $L = \pi_0 c_0 m_0 + \pi_1 c_1 m_1$, with m_i the probability that a class i case is misclassified, will be minimised by assigning cases to class 1 if $p(1|\boldsymbol{x}) > c_0/(c_0 + c_1)$ and to class 0 otherwise. We call the ratio $c_0/(c_0 + c_1)$, with which $p(1|\boldsymbol{x})$ is compared, the *classification threshold*. In unbalanced problems, of the kind with which we are concerned in this paper, misclassifications of the smaller class will be much more serious than that of the larger class, so that $c_0 \gg c_1$. In what follows, without loss of generality we will rescale the costs so that $(c_0 + c_1) = 1$, so that the classification rule becomes

$$\text{Assign points at } \boldsymbol{x} \text{ to class 1 when } p(1|\boldsymbol{x}) > c_0 \text{ and to class 0 otherwise.} \tag{3}$$

Rules of this form will be applicable whether or not the classes are perfectly separable.

In practice, the $p(i|\boldsymbol{x})$ must be estimated from the data. Some strategies adopt the 'sampling' paradigm, in which the class conditional distributions, $p(\boldsymbol{x}|i)$, are separately estimated from the appropriate design set points, and then combined

to yield an estimate of $p(i|\boldsymbol{x})$ using Bayes theorem. For example:

$$p(1|\boldsymbol{x}) = \pi_1 p(\boldsymbol{x}|1)/[\pi_1 p(\boldsymbol{x}|1) + \pi_0 p(\boldsymbol{x}|0)].$$

A popular example of such a method is linear discriminant analysis. This is based on first and second moments of the two class conditional distributions, and assumes that they have the same covariance matrix. This common covariance matrix is estimated from the weighted sample matrices in each of the classes. This means that the estimate of the matrix will be dominated by the sample matrix of the larger class. If the matrices in the two groups are not really the same, this can lead to substantial bias. Other strategies adopt the diagnostic paradigm, in which $p(i|\boldsymbol{x})$ is estimated directly—for example, logistic discriminant analysis. Ripley (1996) points out that methods in which estimated parameters are used in parametric forms for $p(i|\boldsymbol{x})$ are likely to be biased.

There are also practical considerations when one class is much larger than the other. If one class has a million design set points and the other a thousand, it seems computationally inefficient, and largely superfluous, to retain all of the first class.

Several strategies to ease these difficulties have been proposed. A common one is to subsample from the design points in the larger class, so that the two classes contribute similar numbers of points. Alternatively, each design set point from the larger class can be downweighted in the likelihood, so that it contributes less in the parameter estimation. In each case, appropriate counter-adjustments are made at the classification stage. Such strategies seem rather unsatisfying—it would be preferable not to discard information. A third alternative is to assign different weights to the different kinds of misclassification, as described above. This seems the strategy with the soundest theoretical base, and is the strategy we adopt below. Most of the work on this problem seems to be algorithmic rather than model centred, and much of it has appeared in the data mining literature. Examples of papers discussing the issue are Fawcett (1996), Cardie and Howe (1997), Kubat and Matwin (1997), Lee (2000) and Ling and Li (1998).

Nearest neighbour methods estimate the $p(i|\boldsymbol{x})$ by the proportion of class i points amongst the k nearest neighbours to the point $\boldsymbol{x}$ to be classified. This requires a choice of a distance metric and a choice of the parameter k. Euclidean distance is often used for the distance metric, but this is only appropriate if the variables are commensurate. If, as will typically be the case, they are measured in different units or on different scales, it is necessary to adopt an appropriate standardisation. The optimum standardisation, and hence optimum metric, is defined in terms of the unknown contours of $p(1|\boldsymbol{x})$, so an iterative procedure can be used in which a first approximation to $p(1|\boldsymbol{x})$ yields a first approximation to the best metric, which serves to give better estimates of $p(1|\boldsymbol{x})$, which in turn yields a better metric, and so on. Discussions of optimal choice of metric are given in (Fukunaga and Flick, 1984; Hastie and Tibshirani, 1994; Henley and Hand, 1996; Myles and Hand, 1990).

The choice of k in nearest neighbour methods determines the bias/variance trade-off in the estimator—larger values will generally have more bias but less variance. Moreover, k must also be much smaller than the smaller class—which can be relevant when one class is much smaller than the other. A study by Enas and Choi (1986) led to the suggestion that $k \approx n^{2/8}$ or $k \approx n^{3/8}$ was reasonable, and, more empirically, choice by cross-validation is often adopted. Holmes and Adams (2002) describe a Bayesian approach which integrates over the choice of k.

The computational load can often be substantial with nearest neighbour methods, since it is necessary to search the entire design set to identify the nearest neighbours to $\boldsymbol{x}$. For this reason, several authors have proposed strategies for discarding superfluous points—those points which can be discarded without adversely affecting nearest neighbour classification performance. Such strategies are especially appropriate in situations where one class is much larger than the other, and lead to a much better strategy for discarding points than the simple random subsampling mentioned above. Methods for this have been proposed by Hart (1968), Gates (1972), and Hand and Batchelor (1978).

The next section describes the impact of the choice of k on nearest neighbour rule classification performance when the classes are unbalanced.

Section 3 gives an illustration using some personal loan data. Section 4 presents some conclusions.

2. Effect of extreme cost ratios

From Section 1, the k-nearest neighbour classification rule assigns a point with measurement vector $\boldsymbol{x}$ to class 1 if $k_1/k > c_0$, and otherwise to class 0, where k_1 is the number of class 1 points amongst the k design set points closest to $\boldsymbol{x}$. If $p(1|\boldsymbol{x})$ is constant in the neighbourhood of $\boldsymbol{x}$, then the probability that a class 1 point at $\boldsymbol{x}$ is correctly classified by this rule is

$$M_{1|1}(\boldsymbol{x}) = \sum_{\Omega} \binom{k}{k_1} p(1|\boldsymbol{x})^{k_1} (1 - p(1|\boldsymbol{x}))^{k-k_1}, \qquad (4)$$

where $\Omega = \{k_1 : k_1 > kc_0\}$. Here $M_{i|j}(\boldsymbol{x})$ is to be read as the probability that a class j point with measurement vector $\boldsymbol{x}$ will be classified into class i.

More generally, when $p(1|\boldsymbol{x})$ is not constant in the neighbourhood of $\boldsymbol{x}$, $M_{1|1}(\boldsymbol{x})$ is the asymptotic probability that a class 1 point at $\boldsymbol{x}$ will be correctly classified, when k is fixed and the design set size increases. The overall asymptotic probability of correctly classifying class 1 points is thus

$$M_{1|1} = \int M_{1|1}(\boldsymbol{x}) p(\boldsymbol{x}|1)\,\mathrm{d}x \qquad (5)$$

and we can take this as an approximation to the proportion correctly classified in the large sample case. Similar expressions, $M_{0|0}(\boldsymbol{x})$ and $M_{0|0}$ apply for class 0 points.

We are concerned with the unbalanced class size case, so that $c_0 \gg c_1$, and explore what happens as k increases.

(i) Points at $\boldsymbol{x}$ are classified into class 1 if $k_1 > kc_0$. When k is such that $(k-1)/k \leqslant c_0$, $k_1 > kc_0$ if and only if $k_1 > k - 1$. That is, when k is such that $k \leqslant (1 - c_0)^{-1}$, points will only be classified into class 1 if $k_1 = k$—if *all* of the k-nearest neighbours belong to class 1. The probability of this is $p(1|\boldsymbol{x})^k$ (assuming $p(1|\boldsymbol{x})$ to be constant near $\boldsymbol{x}$). We see immediately from this that the probability of correctly classifying a class 1 point will decrease as k increases, provided $k \leqslant (1 - c_0)^{-1}$. When c_0 is large, as we would expect in the unbalanced case with class 1 the larger class, k will be less than $(1 - c_0)^{-1}$ for k which takes large values. For example, if misclassifying a class 0 point is regarded as 99 times as serious as misclassifying a class 1 point, then the probability of correctly classifying class 1 points decreases monotonically (from $p(1|\boldsymbol{x})$) as k increases from 1 all the way up to 100. Such a large value of k may be of the order of, or possibly even larger than, the number of smaller class design set points in unbalanced problems.

Of course, the probability of correctly classifying class 0 points is inversely related to the probability of misclassifying class 1 points. In this situation, class 0 points are correctly classified whenever $k_1 < k$, and this will happen with probability $[1 - p(1|\boldsymbol{x})^k]$, increasing monotonically with k for $k \leqslant 1/(1 - c_0)$.

(ii) Now consider what happens when k exceeds $(1 - c_0)^{-1}$. In particular, suppose that $k > (1 - c_0)^{-1}$, but $k \leqslant 2(1 - c_0)^{-1}$. It can be easily shown that now a class 1 point will be correctly classified, that is $k_1 > kc_0$, if $k_1 = k$ or $k_1 = k - 1$, but not $k_1 = k - 2$. Firstly, if $k_1 = k$ then $k_1 > kc_0$ immediately, since $c_0 < 1$. Secondly, from $k > (1 - c_0)^{-1}$ it follows that $(k-1) > kc_0$, so that if $k_1 = k - 1$, then, $k_1 > kc_0$. Thirdly, if $k \leqslant 2(1 - c_0)^{-1}$ it follows that $k - kc_0 \leqslant 2$, from which $(k-2) \leqslant kc_0$, so that if $k_1 = k - 2$ then $k_1 \leqslant kc_0$. It follows that the probability that a class 1 point will be correctly classified is $p(1|\boldsymbol{x})^k + kp(1|\boldsymbol{x})^{k-1} \times (1 - p(1|\boldsymbol{x}))$. Now

$$\begin{aligned} &p(1|\boldsymbol{x})^k + kp(1|\boldsymbol{x})^{k-1}(1 - p(1|\boldsymbol{x})) \\ &\quad = p(1|\boldsymbol{x})^{k-1}[p(1|\boldsymbol{x}) + k(1 - p(1|\boldsymbol{x}))] \end{aligned}$$

and this is greater than $p(1|\boldsymbol{x})^{k-1}$ when $k > 1$, so that the probability of correctly classifying a class 1 point *increases* as k increases through $(1 - c_0)^{-1}$, before it starts decreasing again.

Once again, the probability of correctly classifying class 0 points changes inversely to this, jumping down as k crosses the $(1 - c_0)^{-1}$ threshold.

Let us summarise the phenomena described in (i) and (ii). First, as k increases from 1 up to $(1 - c_0)^{-1}$, so the probability of correctly classifying class 1 points decreases monotonically. However, as k increases through $(1 - c_0)^{-1}$, so this probability jumps up.

Similar calculations to the above show that this pattern is then repeated. The probability of correctly classifying a class 1 point decreases monotonically with increasing k, until increasing k by 1 leads to an increase in the number of possible values of k_1 for which k_1 can be greater than kc_0. When this happens the probability of correctly classifying a class 1 point increases, before beginning another decline. The probability of correctly classifying class 0 points changes in the reverse direction. Plots of the probability of correct classification (or incorrect classification, of course) against k, for each of the two classes separately, will therefore show a sawtooth pattern.

All of the above is all very well, but it is based on the assumption that $p(1|\boldsymbol{x})$ is constant. In fact, of course, this probability varies with $\boldsymbol{x}$. We need to integrate the probability of correct classification over the $\boldsymbol{X}$ space, as shown in Eq. (5). Now, although the values of the decreases and increases in the probability of correctly classifying class 1 points will depend on the local probability $p(1|\boldsymbol{x})$, the sign of the change (i.e., an increase or a decrease in the probability) is the same for all such probabilities. Furthermore, since the patterns will be based on the same k the changes will occur in phase. This means that, when they are aggregated as in (5), the overall sawtooth pattern will still be evident.

How should we expect changing c_0 to influence this effect? Again, for simplicity, for the moment assume that $p(1|\boldsymbol{x})$ is constant. Then the effect will be most pronounced when c_0 is near the mode of the binomial distribution $B(k, p(1|\boldsymbol{x}))$. For problems with unbalanced classes, we would generally expect small classes to be associated with large c_0, as discussed in Section 1. That is, problems with unbalanced classes are likely to lead to this phenomenon.

We have remarked that the sawtooth pattern will have the opposite shape for the two classes—deterioration in proportion correctly classified for one class is associated with improvement in proportion correctly classified for the other class. This means that when one puts the two classes together the two effects will tend to cancel out. Perhaps this is why the effect appears not to have been reported before: by far the most popular measure of predictive accuracy in theoretical studies of classification rule performance is misclassification rate, which sets $c_0 = c_1$ (even though this is seldom appropriate—see Hand (1997)) and aggregates the numbers misclassified in the two classes.

All of the above discussion has been in terms of the proportions of the two classes which are misclassified by the rule. However, sometimes it is more useful to work in terms of the proportion of those predicted to lie in a class which actually do lie in that class. For example, in consumer credit scoring (Hand, 2001; Hand and Henley, 1997; Thomas, 2000) a common measure is the 'bad rate amongst accepts'. Applicants for loans are classified as good or bad risks, with the former being offered loans, and the proportion of these which turn out to be bad risks is the key measure. The same sort of argument as that followed above also applies in this case. Moreover, since one generally is not concerned with the proportion of goods rejected (this is unmeasurable), there is no opportunity for cancellation. This means that the phenomenon can be especially important with such measures.

3. An example

We illustrate the ideas using a data set of unsecured personal loans, which follow-up has shown to have 11% 'bads' (according to a particular definition of default) and 89% 'goods'. We take the 'bads' as being the smaller class, class 0. We split the data into a design set of 3088 and a test set of 18,530 customers.

Fig. 1 shows the proportion of class 1 and class 0 points correctly classified, using $c_0 = 0.9$, for k from 1 to 99 in steps of 2. The sawtooth pattern is very striking. It is also noticeable that the oscillations remain large, even for k as large as 99 and beyond—that is, about a third of the size of the smaller class in the design set.

Figs. 2 and 3 show, respectively, similar plots using thresholds $c_0 = 0.8$ and $c_0 = 0.7$. Again the sawtooth pattern is evident, though not as pronounced as with more extreme c_0, as expected. The differences in absolute values of the traces for the smaller class (the solid line) between Figs. 1–3 is

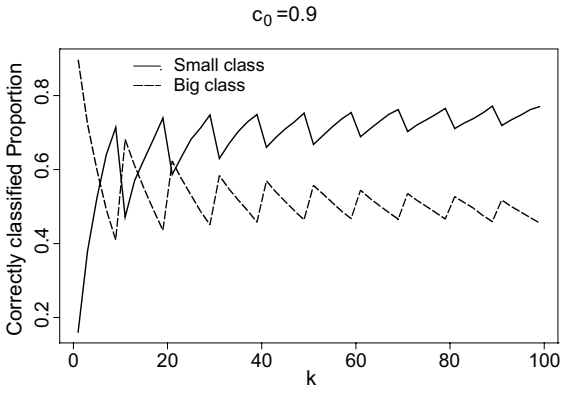

Fig. 1. The proportion of class 1 (broken line) and class 0 (solid line) points correctly classified, using $c_0 = 0.9$ for k from 1 to 100.

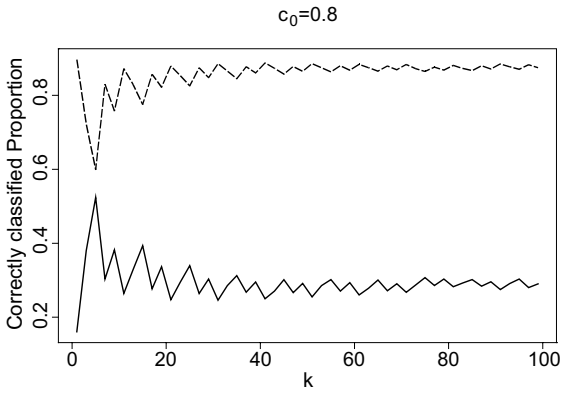

Fig. 2. The proportion of class 1 (broken line) and class 0 (solid line) points correctly classified, using $c_0 = 0.8$ for k from 1 to 100.

striking. This is simply a consequence of the different values of c_0. In Fig. 1, for example, misclassifying elements of the smaller class carries a heavy penalty, so the threshold is arranged so that few are misclassified. In Figs. 2 and 3, however, the penalty is not so great, so it is acceptable to have a lower proportion of the smaller class correctly classified.

Fig. 4 shows the total loss $L = \pi_0 c_0 m_0 + \pi_1 c_1 m_1$ (see Section 1) for different values of k, with $c_0 = 0.9$. As predicted, this does not show a sawtooth pattern—the losses arising from the misclassifications of elements of the two classes tend

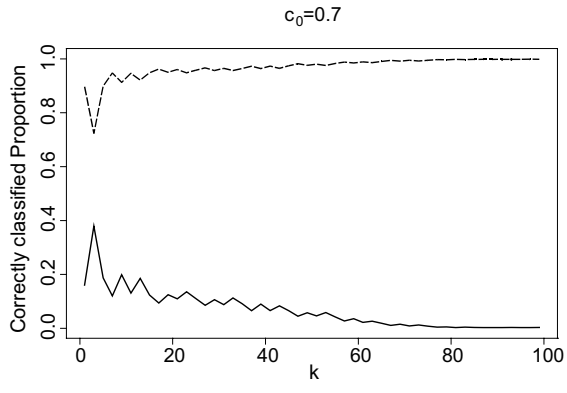

Fig. 3. The proportion of class 1 (broken line) and class 0 (solid line) points correctly classified, using $c_0 = 0.7$ for k from 1 to 100.

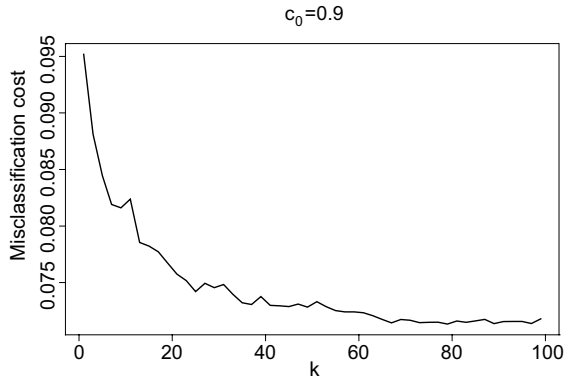

Fig. 4. Total loss, $L = \pi_0 c_0 m_0 + \pi_1 c_1 m_1$, for different values of k.

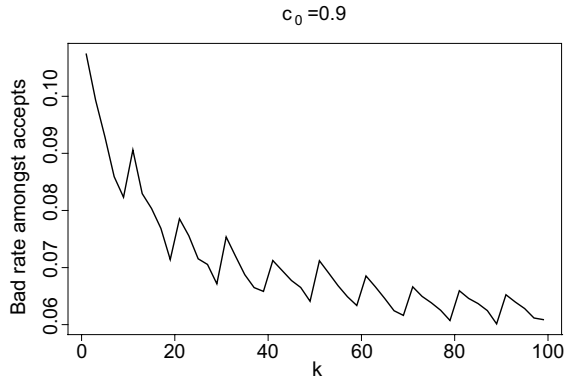

Fig. 5. Bad rate amongst accepts when $c_0 = 0.9$, for k from 1 to 100.

to cancel out, an improvement for one class being matched by a deterioration for the other.

In Section 2 we remarked that a similar oscillating path should be expected using measures based on the proportion of those predicted to be in each class which really were in that class. Fig. 5 shows the bad rate amongst accepts for the unsecured personal loan data using $c_0 = 0.9$. The sawtooth pattern is clear—all the way up to and beyond $k = 100$.

4. Conclusion

Nearest neighbour supervised classification methods have many attractive properties. However, many of these are based on asymptotic arguments. In problems where the classification threshold lies near 0 or 1, as generally is the case in situations where one class is much larger than the other, the fact that k is finite results in a non-monotonic relationship between the proportion of each class correctly classified as k varies. This means that, in general, larger k may not yield better performance than smaller k. In the data set used in Section 3, for example, Fig. 5 shows that a k value of 49 leads to a smaller bad rate amongst accepts than a k value of 91.

Taking this to an extreme, with unbalanced class problems there is the real possibility that the probability of correctly classifying a class 1 point will decrease as k increases right up to and beyond the number of points from the smaller class in the design set. This happens if the number of points from the smaller class in the design set is less than $(1 - c_0)^{-1}$. In such case, the best classification rule for predicting class 1 membership will be based on $k = 1$.

The phenomenon is less important when the misclassifications from the two classes are aggregated, since then the effects on the two classes cancel out, and a performance curve (of overall loss) which is relatively smooth results, showing performance improving with increasing k. This may be one of the reasons why the phenomenon appears not to have been reported before: loss, and in particular misclassification rate, is a very popular measure of the performance of classification rules (even though misclassification rate is based on the assumption of equal costs, which is seldom appropriate).

Acknowledgements

The work of Veronica Vinciotti on this project was partially supported by grants from the Institute of Actuaries and from Fair, Isaac. We would like to express our thanks to Gordon Blunt for drawing our attention to the phenomenon investigated in this paper.

References

Brause, R., Langsdorf, T., Hepp, M., 1999. Neural data mining for credit card fraud detection. In: Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence. IEEE Press, Chicago, pp. 103–106.

Cardie, C., Howe, N., 1997. Improving minority class prediction using case-specific feature weights. In: Proceedings of the 14th International Conference on Machine Learning. Morgan Kaufmann, San Mateo, pp. 57–65.

Enas, G.G., Choi, S.C., 1986. Choice of the smoothing parameter and efficiency of k-nearest neighbour classification. Comput. Math. Appl. 12A, 235–244.

Fawcett, T., 1996. Learning with skewed class distributions: summary of responses. Machine Learning List 8 (20).

Fukunaga, K., Flick, T.E., 1984. An optimal global nearest neighbour metric. IEEE Trans. Pattern Recognition Machine Intell. 6, 314–318.

Gates, G.W., 1972. The reduced nearest neighbour rule. IEEE Trans. Inf. Theory 18, 431.

Hand, D.J., 1997. Construction and Assessment of Classification Rules. Wiley, Chichester.

Hand, D.J., 2001. Modelling consumer credit risk. IMA J. Manage. Math. 12, 139–155.

Hand, D.J., Batchelor, B.G., 1978. An edited condensed nearest neighbour rule. Inf. Sci. 14, 171–180.

Hand, D.J., Henley, W.E., 1997. Statistical classification methods in consumer credit scoring: a review. J. Roy. Statist. Soc. Ser. A 160, 523–541.

Hart, P.E., 1968. The condensed nearest neighbour rule. IEEE Trans. Inf. Theory 14, 515–516.

Hassibi, K., 2000. Detecting payment card fraud with neural networks. In: Lisboa, P.J.G., Vellido, A., Edisbury, B. (Eds.), Business Applications of Neural Networks. World Scientific, Singapore.

Hastie, T.J., Tibshirani, R.J., 1994. Discriminant adaptive nearest neighbour classification. IEEE Trans. Pattern Anal. Machine Intell. 18, 607–616.

Henley, W.E., Hand, D.J., 1996. A *k*-nearest neighbour classifier for assessing consumer credit risk. The Statistician 44, 77–95.

Holmes, C., Adams, N.M., 2002. A probabilistic nearest neighbour method for statistical pattern recognition. J. Roy. Statist. Soc. Ser. B 64, 1–12.

Kubat, M., Matwin, S., 1997. Addressing the curse of imbalanced training sets: one-sided selection. In: Proceedings of the 14th International Conference on Machine Learning, pp. 217–225.

Lee, S.S., 2000. Noisy replication in skewed binary classification. Comput. Statist. Data Anal. 34, 165–191.

Ling, C.X., Li, C., 1998. Data mining for direct marketing: problems and solutions. In: Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining. AAAI Press, New York, pp. 73–79.

McLachlan, G.J., 1992. Discriminant Analysis and Statistical Pattern Recognition. John Wiley and Sons, New York.

Myles, J.P., Hand, D.J., 1990. The multi-class metric problem in nearest neighbour discrimination rules. Pattern Recognition 23, 1291–1297.

Ripley, B.D., 1996. Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge.

Thomas, L.C., 2000. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. Internat. J. Forecasting 16, 149–172.

Webb, A.R., 1999. Statistical Pattern Recognition. Arnold, London.