The Statistician (1996)
45, *No.* 1, *pp.* 77–95

A *k*-nearest-neighbour classifier for assessing consumer credit risk

By W. E. HENLEY

Abbey National PLC, Milton Keynes, UK

and D. J. HAND†

The Open University, Milton Keynes, UK

[Received June 1994. Revised September 1995]

SUMMARY

The last 30 years have seen the development of credit scoring techniques for assessing the creditworthiness of consumer loan applicants. Traditional credit scoring methodology has involved the use of techniques such as discriminant analysis, linear or logistic regression, linear programming and decision trees. In this paper we look at the application of the *k*-nearest-neighbour (*k*-NN) method, a standard technique in pattern recognition and nonparametric statistics, to the credit scoring problem. We propose an adjusted version of the Euclidean distance metric which attempts to incorporate knowledge of class separation contained in the data. Our *k*-NN methodology is applied to a real data set and we discuss the selection of optimal values of the parameters *k* and *D* included in the method. To assess the potential of the method we make comparisons with linear and logistic regression and decision trees and graphs. We end by discussing a practical implementation of the proposed *k*-NN classifier.

Keywords: Classification rule; Credit risk; Nearest neighbour methods

1. Introduction

Consumer credit is granted by banks, building societies, retailers, mail order companies and various other lending institutions and is a sector of the economy that has seen rapid growth over the last 30 years. Traditional methods of credit risk assessment involved the use of human judgment, based on experience of previous decisions, to determine whether to grant credit to a particular individual. The economic pressures resulting from the increased demand for credit and the emergence of new computer technology have led to the development of sophisticated statistical models to aid the credit granting decision. Credit scoring is the name used to describe the process of determining how likely an applicant is to default with repayments. In this paper we look at the application of the *k*-nearest-neighbour (*k*-NN) method, a standard technique in pattern recognition and nonparametric statistics, to the credit scoring problem.

Traditional credit scoring methodology has focused on using techniques such as discriminant analysis and linear regression to discriminate between applicants who are assumed to belong to one of two classes, namely good and bad credit risks. Previous papers which include the application of these methods to credit scoring include Durand (1941), Myers and Forgy (1963), Apilado *et al.* (1974), Eisenbeis (1978), Grablowsky and Talley (1981), Reichert *et al.* (1983) and Srinivasan and Kim (1987). Despite the fact that both these techniques can provide good discrimination between the good and bad classes, they are subject to the conceptual problems outlined by Eisenbeis (1978) and Reichert *et al.* (1983). Logistic regression is a more appropriate technique for credit scoring and has been applied with success by Wiginton (1980), Gilbert *et al.* (1990) and Leonard (1993).

†*Address for correspondence*: Department of Statistics, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK.
E-mail: d.j.hand@open.ac.uk

0039–0526/96/45077

However, as we shall see later in this paper, it has given similar predictive accuracy to linear regression in comparisons and, like that method, suffers from the disadvantages of assuming a parametric model form. This provides our motivation for looking for alternative classification techniques for credit scoring. Other techniques which are commonly used in practice are linear programming and decision trees. (Both are discussed in Srinivasan and Kim (1987) and Henley (1995).)

The k-NN method is a standard nonparametric technique used for probability density function estimation and classification and was originally proposed by Fix and Hodges (1952) and Cover and Hart (1967). It was chosen as a suitable method for applying to consumer credit data for several reasons:

(a) the nonparametric nature of the method enables modelling of irregularities in the risk function over the feature space;

(b) the k-NN method has been found to perform better than other nonparametric techniques such as kernel methods when the data are multidimensional (see Terrell and Scott (1992));

(c) the k-NN method is a fairly intuitive procedure and as such could be easily explained to business managers who would need to approve its implementation and it can be used dynamically.

A simple NN classifier was applied to credit scoring data in Fogarty and Ireson (1993) in a general comparison of techniques. The results were promising despite the simplicity of the NN classifier used.

The k-NN method involves estimating the good or bad risk probabilities for an applicant to be classified by the proportions good and bad among the k 'most similar' points in a training sample. The similarity of points is assessed by using a suitable distance metric. An important part of our analysis will be the selection of suitable distance metrics for the k-NN method. An adjusted version of the Euclidean metric which incorporates knowledge of underlying equiprobability contours for class membership is proposed in Section 2. These contours are estimated by using regression weights, calculated from the data, with the intention of incorporating into the metric knowledge of class separation in the sample. A general transformation of the data which has the same property is defined in Henley (1995). This allows us to transform the data before using any standard metric, thus providing a general framework for selecting a data-dependent metric.

The aim of this research is to provide a practical classification model that can improve on traditional credit scoring techniques. In Section 3 we apply our k-NN methodology to a data set supplied by a large mail order company. Attention is given to the nature of the data and the criterion for assessing performance. We include a discussion of several interesting features of the k-NN results and consider the selection of optimal values of the parameters included in the method. The performance of the k-NN classification rule is compared with a range of other discrimination techniques including linear regression, logistic regression and decision trees. The sensitivity of the k-NN method to changes in the population priors is considered in Section 3.5. Issues relating to the practicality of implementation are discussed in Section 3.6.

2. Selection of metrics

The selection of an appropriate distance measure is an important part of the k-NN method but one that has not received proportional attention in the literature. The aim in choosing a metric is to improve classification performance according to some specified criterion. The work that has been published on metric selection has concentrated on the NN rule and has used minimization of the difference between the finite sample misclassification rate and the asymptotic misclassification rate as the performance criterion.

A standard distance measure is the Euclidean metric given by

$$d_1(\mathbf{x}, \mathbf{y}) = \{(\mathbf{x}-\mathbf{y})^{\mathrm{T}}(\mathbf{x}-\mathbf{y})\}^{1/2}$$

where $\mathbf{x}$ and $\mathbf{y}$ are points in the feature space. Myles (1991) gave two simple examples to show why d_1 may not always be the most sensible distance measure to use. Fukanaga and Flick (1984) and Short and Fukanaga (1982) have considered the problem of selecting data-dependent versions of the Euclidean metric in the NN case. Fukanaga and Flick introduced a general approach to incorporating knowledge from the data through the metric

$$d_2(\mathbf{x},\ \mathbf{y}) = \{(\mathbf{x}-\mathbf{y})^{\mathrm{T}}\mathbf{A}(\mathbf{x}-\mathbf{y})\}^{1/2}$$

where $\mathbf{A}$ can be any $p \times p$ symmetric positive definite matrix and p is the dimension of the feature space. There are two possible approaches to selecting $\mathbf{A}$: local metrics are defined to be those for which $\mathbf{A}$ can vary with $\mathbf{x}$ and global metrics are defined to be those for which $\mathbf{A}$ is independent of $\mathbf{x}$. In the global case the distance between two points depends only on their relative position in Euclidean space. Fukanaga and Flick (1984) proposed to use a global metric that is optimized with respect to the mean-squared error between the asymptotic and finite sample NN risk. They presented a simulation experiment which showed that a global metric can be a practical way of improving over the standard Euclidean metric.

The danger of using a local metric approach is that the metric will incorporate local features of the training set that are not representative of the population from which it is drawn. It is also difficult to determine the metric accurately since for each $\mathbf{x}$ the metric must be determined from a small region around $\mathbf{x}$. We choose to adopt the global approach to metric selection based on the Euclidean metric described above. Our approach differs in two ways from that adopted by Fukanaga and Flick (1984): first, we deal with the general case of k NNs and, secondly, we use a different assessment criterion (the minimization of the bad risk rate among accepted applicants: see Section 3.2 for more details).

We define the distance between two points as being the separation between them in the direction orthogonal to equiprobability contours for $P(g\,|\,\mathbf{x})$, the probability of belonging to class g (e.g. the class of good credit applicants in our problem) given $\mathbf{x}$, with allowance for random variation. If we did know the equation of the true equiprobability contours then the best metric to use would be the distance in the orthogonal direction. The squared distance between two points $\mathbf{x}$ and $\mathbf{y}$ in this direction (denoted by $\mathbf{w}$) is given by

$$\{\mathbf{w}^{\mathrm{T}}(\mathbf{x}-\mathbf{y})\}^2 = (\mathbf{x}-\mathbf{y})^{\mathrm{T}}\mathbf{w}\mathbf{w}^{\mathrm{T}}(\mathbf{x}-\mathbf{y}). \tag{2.1}$$

In practice we estimate the equiprobability contours from the data via a regression model and so our metric needs to take into account random variation in the estimates. Moreover the true contours are unlikely to be exactly linear in practice. For these reasons our proposed metric includes a contribution from the squared Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$ given by

$$(\mathbf{x}-\mathbf{y})^{\mathrm{T}}(\mathbf{x}-\mathbf{y}) = (\mathbf{x}-\mathbf{y})^{\mathrm{T}}\mathbf{I}(\mathbf{x}-\mathbf{y}). \tag{2.2}$$

Combining terms (2.1) and (2.2) gives the metric

$$d_3(\mathbf{x},\ \mathbf{y}) = \{(\mathbf{x}-\mathbf{y})^{\mathrm{T}}(\mathbf{I}+D\mathbf{w}\mathbf{w}^{\mathrm{T}})(\mathbf{x}-\mathbf{y})\}^{1/2}.$$

Our metric is in the form of the generalization of the Euclidean metric d_2 proposed by Fukanaga and Flick (1984), with the matrix $\mathbf{A}$ given by

$$A_{D,\mathbf{w}} = (\mathbf{I}+D\mathbf{w}\mathbf{w}^{\mathrm{T}}).$$

The metric d_3 includes the distance parameter D, and the selection of a suitable value for this parameter will be an important part of the implementation of the methodology.

In this paper we concentrate on the use of linear regression weights to give the direction orthogonal to assumed equiprobability contours. Logistic regression weights may be a more natural choice, because of their direct relationship with the predicted probabilities of class membership, but in our experiments they produced results that were barely distinguishable from those produced by the linear weights. This is probably because the probabilities of being good lie within the range 0.2–0.8 for most parts of the measurement space. Detailed comparisons are considered by Henley (1995).

3. Empirical study of the *k*-nearest-neighbour method with data-dependent metrics

3.1. *Consumer credit data*

The data used in our analysis consist of a sample from the full population of applicants for credit from a large mail order company. All the applicants in the sample, including applicants who would normally have been rejected, were given credit and observed over a period of a year. A standard procedure was used to define creditworthiness. Applicants who defaulted for three consecutive months were classified as bad and the remaining applicants in our data set were classified as good. This allows us to treat creditworthiness as a two-state problem and to use techniques for dealing with binary response data. In particular we chose not to consider creditworthiness as a continuous property. Alternative definitions of credit default are considered by Crook *et al.* (1992).

The predictor variables used for constructing credit scoring models were extracted from the application form and other sources, such as a historical database. These variables, which we call *characteristics*, are usually nominal or ordinal in nature. The values that they take are referred to as *attributes*. An example of a suitable characteristic 'time on electoral roll' is described in Table 1.

Because of the categorical nature of the data and the fact that the characteristics are often measured on different scales, the data were transformed before carrying out the analysis. The selected transformation is widely used in the credit industry and is described in Crook *et al.* (1992). The values of the predictor variables are replaced by *weights of evidence* where the weight of evidence of the jth attribute of the ith characteristic is given by

$$w_{ij} = \ln(p_{ij}/q_{ij})$$

where p_{ij} is the proportion of those classified good in attribute j of characteristic i and q_{ij} is the proportion of those classified bad in attribute j of characteristic i.

By using the above transformation we put the data into a ratio form allowing us to exploit techniques such as logistic regression. This is also beneficial to the k-NN method because its performance deteriorates rapidly as the number of dimensions increases and the alternative of using indicator variables would lead to a very high dimensionality.

Variables were selected for inclusion in the analysis by considering the information value

$$\text{IV} = \sum_{i,j} (p_{ij} - q_{ij}) w_{ij}$$

where p_{ij}, q_{ij} and w_{ij} are defined above. This is a common measure of the discriminatory power of a characteristic and is used by many score-card developers in the credit industry. 16 characteristics were chosen on the basis of high information values and expert knowledge of the data. The suitability of this set of characteristics was checked by building linear regression models with extra characteristics added. A factor in the selection of characteristics was the need to justify the chosen subset on business grounds.

TABLE 1
Description of the characteristic 'time on electoral roll'

Attribute	*Description*
1	Not on electoral role
2	1 year or less
3	2 years
4	3 years
5	4 years
6	5+ years

TABLE 2
Description of the data set used for assessing the performance of the *k*-NN classifier

	No. of variables	*No. of classes*	*No. of cases*	*% bad in full sample*
Design set	16	2	15054	54.49
Test set	16	2	4132	54.70

A decision tree combining several of the other available characteristics was included in the analysis as a characteristic. The aim of doing this was to help the linear and logistic regression methods to take account of interactions between variables. This approach is similar to the hybrid classifier of discriminant analysis and decision trees proposed by Boyle *et al.* (1992). We would expect that by adopting this approach we would remove one of the advantages of the *k*-NN method and thus set a more difficult base-line for it to beat.

A summary of the data set is given in Table 2.

To obtain a more robust estimate of the performance of the *k*-NN method our analysis involves taking five random design–test sample splits from the combined design and test samples mentioned above (using the same ratio between sample sizes). The method is then applied to each of the design sets in turn and the results are assessed by using the corresponding test set. If the results are then averaged our assessment procedure becomes similar to that adopted by Leonard (1993).

3.2. *Criterion for assessing performance*

The criterion used in this study is rather unusual and deserves some discussion. Whereas most applications of assignment procedures work with error rates (perhaps appropriately weighted by a loss function), for commercial reasons in this application the proportion to be accepted is prespecified and the aim is to minimize the number of bad risk applicants accepted. This criterion is thus involved with only half of the misclassification space. This has some interesting consequences for the assignment rule.

One immediate consequence is that it imposes bounds on the best and worst performances that can be achieved. Suppose that a proportion a of applicants must be accepted and that the population contains a proportion p of good risk applicants. The *best* bound for the bad risk rate among those accepted is given by

$$\text{best BR} = \begin{cases} 1 - p/a & a > p, \\ 0 & a \leqslant p. \end{cases}$$

It is obtained by accepting all good applicants and making up the rest of the $100p\%$ acceptances from the bad risk applicants (note that it does not depend on the number of applicants in the sample). The *worst* bound for the bad risk rate among the acceptances is given by

$$\text{worst BR} = \begin{cases} (1 - p)/a & a > 1 - p, \\ 1 & a \leqslant 1 - p. \end{cases}$$

It comes from accepting all the bad applicants and making up the rest of the acceptances from good applicants.

Corresponding bounds on the best and worst good risk rates for the rejects and error rates can also be found:

$$\text{best GR} = \begin{cases} 0 & a > p, \\ (p - a)/(1 - a) & a \leqslant p; \end{cases}$$

$$\text{worst GR} = \begin{cases} 1 & a > 1-p, \\ p/(1-a) & a \leqslant 1-p; \end{cases}$$

$$\text{best ER} = \begin{cases} a-p & a > p, \\ p-a & a \leqslant p; \end{cases}$$

$$\text{worst ER} = \begin{cases} 2-(a+p) & a > 1-p, \\ a+p & a \leqslant 1-p. \end{cases}$$

Our particular problem has $p = 45.3\%$ and $a = 0.70$. Using the above expressions yields the following bounds on the bad risk rate: best, 35.3%, worst, 78.1%. The expected bad risk rate of a rule which classifies points randomly is 54.7%. These results allow us to put the performance of the classification methods that we consider into perspective. For example, whereas in general a bad risk rate of 36% might represent a disappointingly poor performance, in this context it is only just above the theoretical best.

3.3. *Estimation of k and D*

The *k*-NN method with metric d_3 was implemented by using a C routine (see Henley (1995)). To apply the method to future samples it is necessary to choose values for the two parameters *k* and *D*. We propose a procedure for doing this which uses the design sample and assess performance by using an independent test sample from the same population.

In principle the estimation of suitable *k* and *D* is straightforward, requiring a simple bivariate optimization (e.g. by using a steepest descent algorithm) of the criterion. In what follows we present a more detailed discussion of the properties of this criterion for various choices of *k* and *D* so that better understanding is gained.

Figs 1–3 show plots of bad risk rate against *k* for three of the five randomly selected design–test

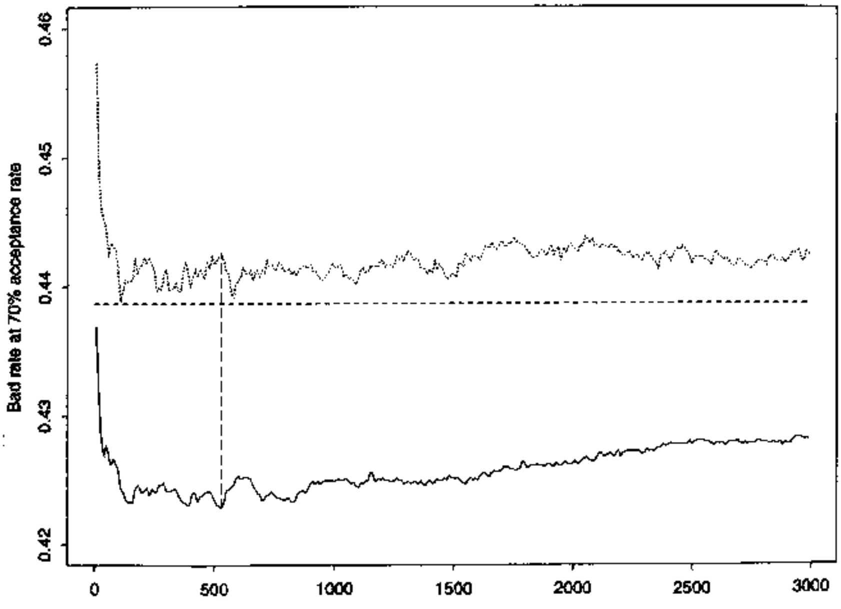

Fig. 1. *k*-NN method using data set 1 and $D = 0.00$: ———, design set; · · · · · ·, test set; - - - - -, linear regression

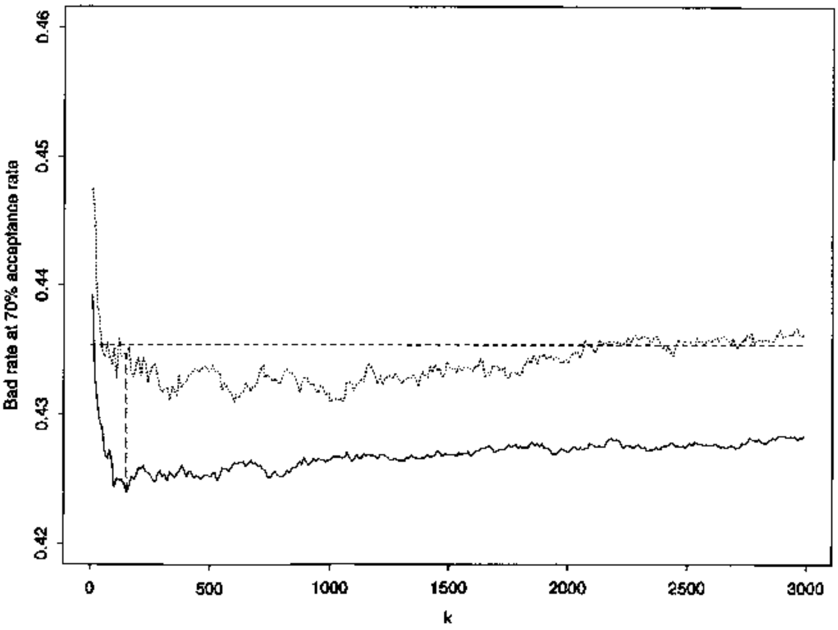

Fig. 2. *k*-NN method using data set 3 and $D = 1.40$: ——, design set; ·····, test set; - - - - -, linear regression

sets using various choices of *D*. The figures were chosen to illustrate the subtle variations in the bad risk rate curves across various samples and values of *D*. For each curve the bad risk rate among the accepted applicants is shown using both the design and the test samples. The design sample curve comes from classifying each design set point using its *k* nearest design set points (excluding itself). The design set curve could be used to select the appropriate value of *k* to consider for the corresponding test set. The resulting *k* are shown by the vertical lines joining the lowest points on the design set curves with the appropriate test curves (we consider alternative ways of selecting *k* after a discussion of properties of these curves). The horizontal broken lines represent the performance of a classifier using linear regression.

Fig. 4 shows the averaged bad risk rate curve for the five samples, using, in each case, the optimal value of *D* as estimated from the design set.

3.3.1. *Properties of bad risk rate curves.* Several observations may be made about Figs 1–4.

(a) There is a difference in level between the bad risk rate curves for the design and test samples in Figs 1–4. This is explained by the difference in bad risk rates among the full design and test samples in these cases. For example, sample 1 had a bad risk rate of 54.24% in the design sample and 55.59% in the test sample. We chose not to fix the proportion of good applicants in each randomly selected test sample because we were interested in predicting the performance of future applicants and the proportion of bad applicants in such a sample will vary.

(b) The curves appear very jagged. In fact, the bad risk rate criterion was evaluated for $k = 10, 20, 30, \ldots$ up to the indicated maximum values. If it had been evaluated for every value of *k* the curves would have an even greater fractal nature. Marked jaggedness clearly has implications for choosing a value of *k*, since it implies that slight differences will produce large consequences. However, an examination of the vertical axis shows that even the largest jumps in the body of the curve are produced by no more than about 10 data

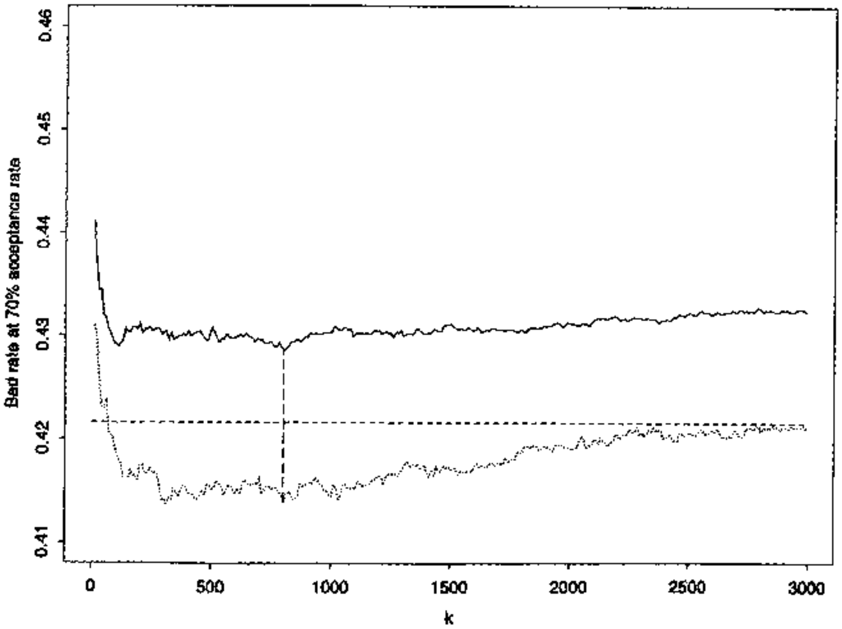

Fig. 3. *k*-NN method using data set 4 and $D = 1.80$: ———, design set; ······, test set; - - - - -, linear regression

points, i.e., although the curve appears irregular, the consequences are fairly small. Despite this, we propose a smoothed version of the *k*-NN estimator as outlined later.

(c) The early parts of the curves are more jagged—even less regular—than the later parts. These curves are the end product of two averaging processes: first, the averaging inherent in using the *k* NNs to estimate probabilities and, second, the averaging over test set points to produce an overall estimate of performance at each particular value of *k*. For the test set this second part is always an average of 4132 estimated probabilities but the first part involves averaging over samples of size *k*. Since early parts of the curve are based on smaller *k*-values the associated probability estimates will be expected to have large variances (the variance of the estimated probabilities from the *k*-NN rule is given by $\text{var}\{\hat{P}(i|\mathbf{x})\} = pq/k$ where i is the class, p is the true probability of belonging to class i and $q = 1 - p$). There may also be a dependence effect since, the larger the value of *k*, the more design set points each classification is likely to have in common. (With *k* equal to the entire design set all classifications would be based on the same data set and so the curve would rise to the full sample bad risk rate.)

(d) We found that for samples 2–5 there were values of D and ranges of k for which the test sample curves drop below the line representing the performance of a classifier using linear regression. This indicates that our method has the potential to lead to improved discrimination. Furthermore, as we shall consider later, the adjusted Euclidean metric appears to give improved performance over the standard Euclidean metric for a range of D-values.

(e) The *k*-NN curves remain below the linear regression line for a considerable range of values of *k* and, in general, the range of 'best' values of *k* is large—the curves have broad flat valleys. The breadth of these valleys came as a surprise to us. We might reasonably expect that as *k* is increased the rising bias of the probability estimates will overcome any improvement in the variance of the estimates, resulting in a consequent loss of classification accuracy.

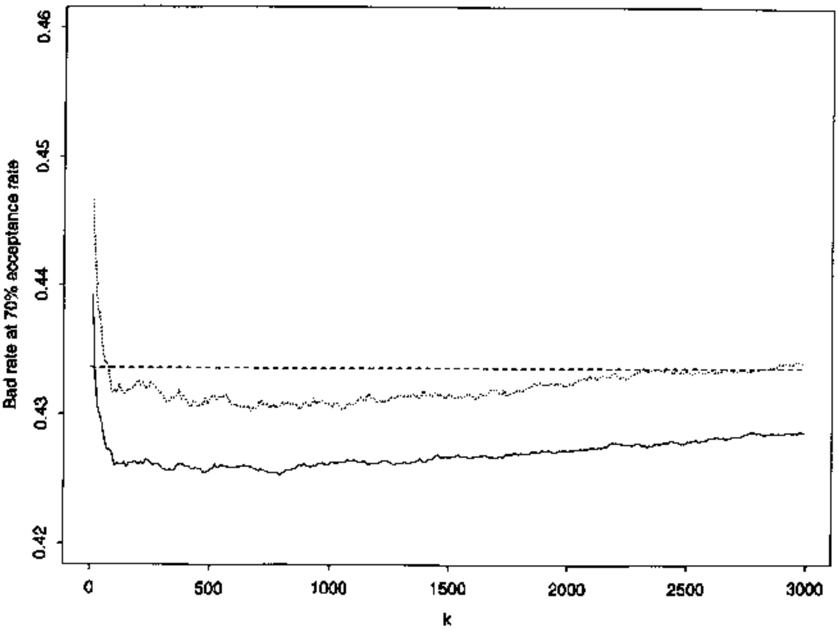

Fig. 4. *k*-NN method averaged over samples: ———, design set; · · · · · ·, test set; - - - - -, linear regression

To explore this we considered plots of $\hat{P}(g|\mathbf{x}, k_1)$ against $\hat{P}(g|\mathbf{x}, k_2)$ for the points in the original test set, using different values of k_1 and k_2. Fig. 5 shows such a plot when $k_1 = 100$ and $k_2 = 1000$. The increased variation in the ordering of points in the middle of the range is due to the higher variance of the probability estimates in this region.

Fig. 5 shows that, although there is variation in the ordering of the test set points for the *k*-NN method with the two *k*-values, most points are placed in the same class by the two classifiers. This means that the points accepted by using both *k*-values swamp the points only accepted by using one.

Table 3 shows the numbers of good and bad applicants accepted by the *k*-NN classifier using either one or both of $k_1 = 100$ and $k_2 = 1000$. The proportions of bad applicants among those accepted in the first and second rows are very similar, indicating the similar creditworthiness of the swap sets. This, combined with the 'swamping' effect of the points accepted under both classifiers helps to explain the 'flat valley' phenomenon.

(f) Fig. 6 shows what happens when a very large value (100) is taken for *D*. Such a large *D* means that the distance is effectively being measured orthogonally to the equiprobability contours of $P(g|\mathbf{x})$, i.e. we would expect the results to be almost identical with those obtained by imposing a threshold on the linear regression. Fig. 6 shows that the prediction is correct. Another feature of this curve is the large fluctuations in the bad risk rate for high *k*. From (c) we know that the bad risk rate curve must reach the sample bad risk rate as *k* tends to the number of points in the design set. Owing to the criterion in this problem the convergence does not begin until very high values of *k* are reached. See Henley (1995) for a further discussion of this issue.

To complement the above plots of *k*, Fig. 7 shows a curve of bad risk rate against *D* for the original test sample. This curve shows a global minimum, as we hoped, although the differences in bad risk rate are quite small. We conclude that the bad risk rate is fairly insensitive to the choice of *D*.

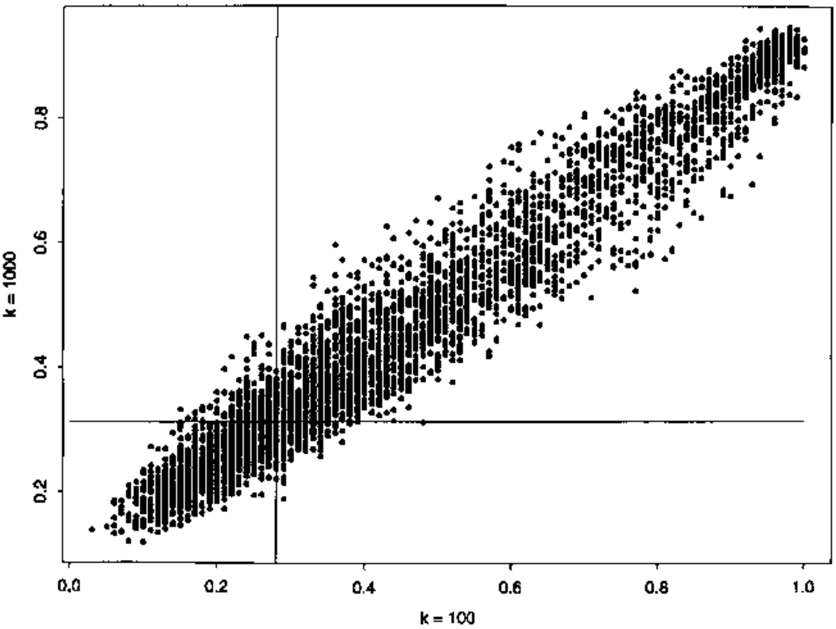

Fig. 5. $P(g|\mathbf{x})$ plotted for two values of k

To bring the discussion of the raw data results together, Fig. 8 shows the bad risk rate values for the range of k and D-values together (using the original test sample). This highlights the insensitivity of the minimum to the choice of both k and D.

3.3.2. *Smoothing of bad risk rate curves.* As has already been suggested, we could simply use the values of k and D which give the lowest bad risk rate for the design set. (We obviously cannot pick k and D from the test set because we are assuming that it is a future sample for which the true classes are not available at the time of classification.) However, this approach is risky because of the fluctuations in the bad risk rate curves for small k described in (b) in Section 3.3.1. The result that we have observed for the design set may be due to random variation and not attributable to a feature of the underlying population. To try to obtain a more robust choice of k we propose to smooth the bad risk rate curves for the design set. To find the estimated bad risk rate for a particular value of k we average the raw bad risk rates for a range of values of k around the value in question. To

TABLE 3
Number of good and bad applicants accepted by using the k-NN method with $k = 100$ and/or $k = 1000$

	Good	*Bad*
Accepted ($k = 100$ only)	49	123
Accepted ($k = 1000$ only)	53	119
Accepted ($k = 100$ and $k = 1000$)	1594	1127

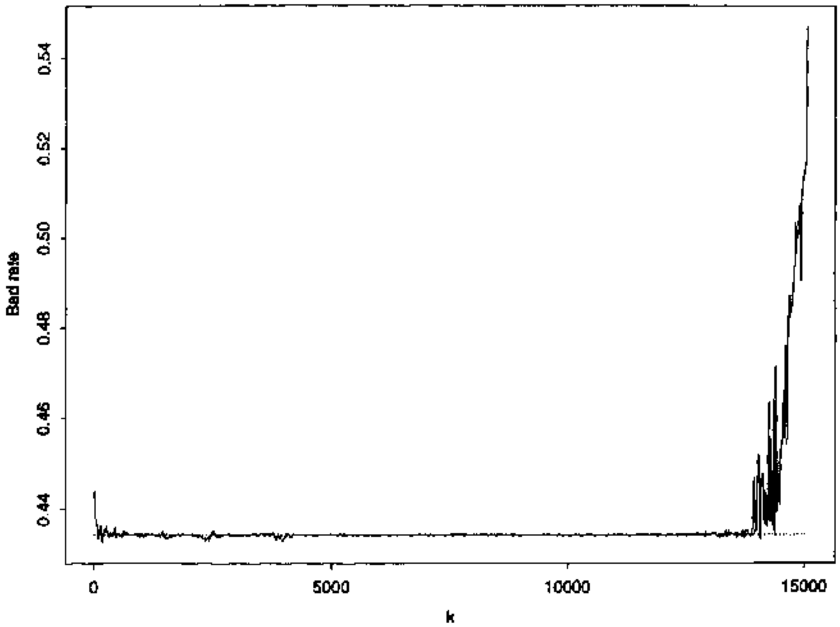

Fig. 6. k-NN method with $D = 100$: ———, k-NN; ·······, linear regression

formalize the procedure we choose a smoothing parameter h and estimate the bad risk rate for $k = k_0$ by

$$S(k_0) = \begin{cases} \dfrac{1}{2h+1} \sum_{i=k_0-h}^{k_0+h} R(i) & \text{for } h < k_0 < 3000 - h, \\ R(k_0) & \text{otherwise,} \end{cases}$$

where $R(i)$ is the raw bad risk rate for $k = i$.

Smoothing was not performed for very low and high values of k because we are only interested in finding estimates of the optimum values, which can be seen from Figs 1–4 to occur for $100 < k < 2000$. We choose to use this smoothing function because of its intuitive appeal and simplicity.

This approach to selecting k and the choice of a suitable smoothing parameter h is analogous to kernel density estimation (see for example Hand (1982)). In choosing k we wish to balance the conflicting aims of ironing out anomalies in the design set and preserving the structure of the data. For example, Fig. 1 shows a sharp minimum in the design and test sets for $k = 100$ when $D = 0$. Because this k-value is close to values giving much higher bad risk rates, the minimum is smoothed out by the above function.

Fig. 9 shows an example of a smoothed bad risk rate curve for the design set from sample 5. The raw curves for the design and test set have been added. The vertical lines represent the selected values of k using the smoothed and raw design curves.

By the same reasoning as above we could choose to use an average of the predictions from a range of values of k to predict the true class for the test sets. This is equivalent to considering smoothed versions of the test set curves. More complex weighted averages could be adopted. (This would be like working out a kernel density estimate based on the value of the smoothing parameter selected from the design set.)

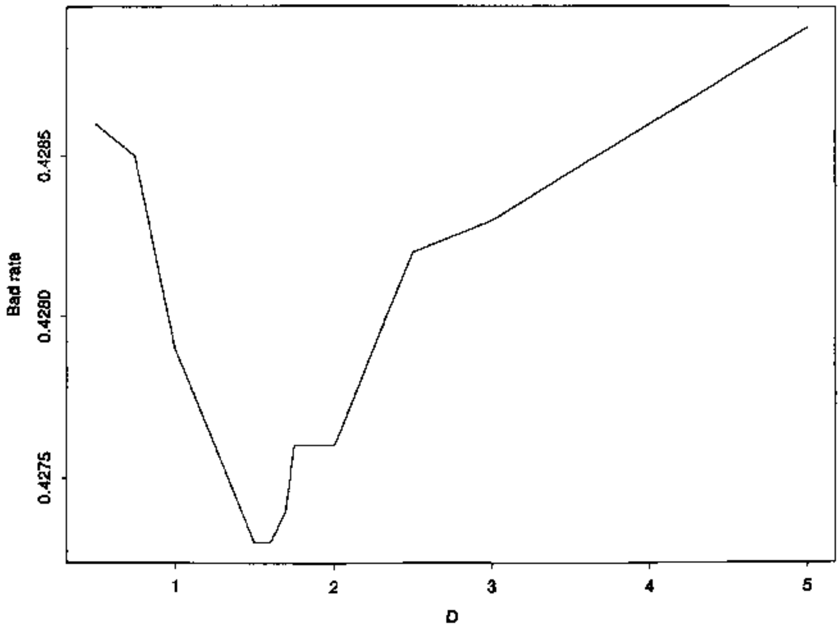

Fig. 7. Lowest bad risk rate against *D* for the original test set

3.4. *k-nearest-neighbour results*

We now bring together the results of applying the various methods of selecting *k* and *D* discussed in the previous section. For each sample we begin by selecting the value of *D* which gives the lowest overall bad risk rate in the design set (we choose to use smoothed results). Having focused on a particular value of *D*, we select a value (or range of values) of *k* from the design set (with or without smoothing) and use this value to classify the test set. The results for the test set could then be smoothed as explained in the previous section.

Table 4 shows the bad risk rates at a 70% acceptance rate for each sample for the four combinations of smoothing or non-smoothing of the design and test set curves. The selected values of *k* are shown and it can be seen that the choice of *k* is highly sensitive to the sample and whether smoothing is employed. However, since the bad risk rate curves are fairly flat over a wide range of *k*- and *D*-values, a large change in *k* does not give rise to much change in performance.

TABLE 4
Bad risk rates and selected values of *k* using smoothing or no smoothing of the design and test set curves for each sample

Sample	*Rates for the following combinations and values of k:*						*D*
	Design (smoothing)			*Design (no smoothing)*			
	Test (smoothing)	*Test (no smoothing)*	*k*	*Test (smoothing)*	*Test (no smoothing)*	*k*	
1	44.02	43.96	1260	44.01	44.01	1160	1.6
2	43.66	43.63	690	43.90	43.90	340	1.8
3	43.38	43.45	200	43.42	43.46	150	1.4
4	41.55	41.55	750	41.85	41.85	100	1.4
5	42.86	42.84	820	42.85	42.83	780	1.4

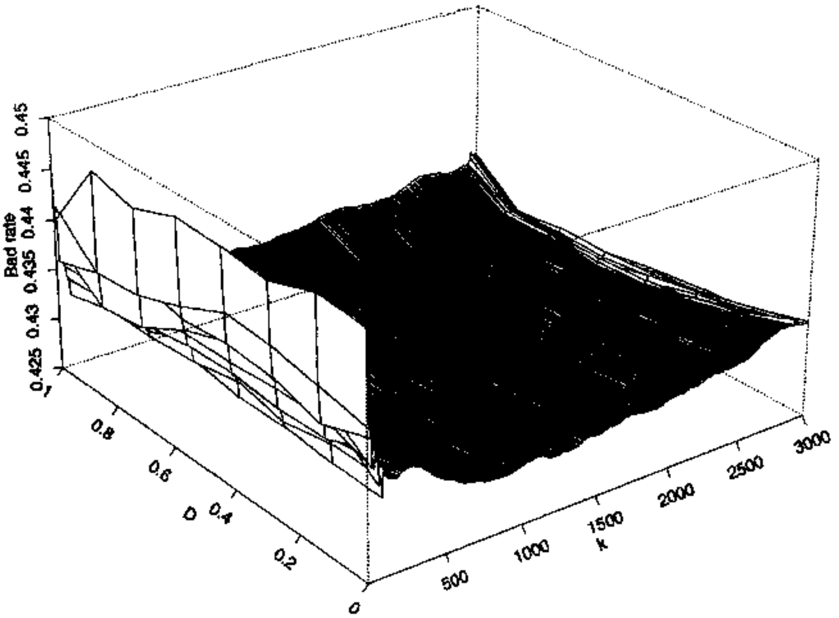

Fig. 8. Bad risk rate against k and D for the original test set

Although the method of smoothing employed is fairly crude it does lead to more consistent results than when no smoothing is used. It is clear from the results that smoothing of the design set curves is more effective than smoothing of the test set curves. In particular, for samples 2 and 4 the results are between 0.24% and 0.3% better when the value of k is chosen from the smoothed design set results. For the other samples there is little difference in the results whether smoothing is employed or not. This indicates that if the raw curve minimum represents a real feature of the population then smoothing the bad risk rates will not lead to a significant change in performance; however, smoothing will help to reduce the chance of selecting a k-value which is only a minimum in one sample (owing to random variation).

We could consider more complex smoothing functions and this might lead to improvements in the overall performance of the method. However, we chose not to do this because the shallow nature of the bad risk rate curves means that the performance of the method is not particularly sensitive to the choice of k. In the rest of this paper we focus on using a smoothed version of the design set curves to select k, but we do not perform smoothing on the test set curves when assessing performance.

To provide a base-line against which to compare our k-NN predictions Table 5 shows the bad risk rates for each sample from linear regression models (and from the k-NN method with k selected from the smoothed design set curve). The k-NN classifier with adjusted Euclidean metrics performs better for samples 2–5 and linear regression performs slightly better for sample 1. The result for sample 1 gives grounds for being cautious in our interpretation of the results (although the difference in this case is unlikely to be significant).

The averaged bad risk rates for linear regression, the k-NN method with the standard Euclidean metric (i.e. $D = 0$) and the adjusted Euclidean metric are 43.36, 43.25 and 43.09 respectively. These results indicate that the use of the k-NN method with adjusted Euclidean metrics can lead to improvements over the use of the standard Euclidean metric and linear regression. Although the difference in bad risk rates is quite small, if it represents a real difference then it could result in large savings for the credit grantor.

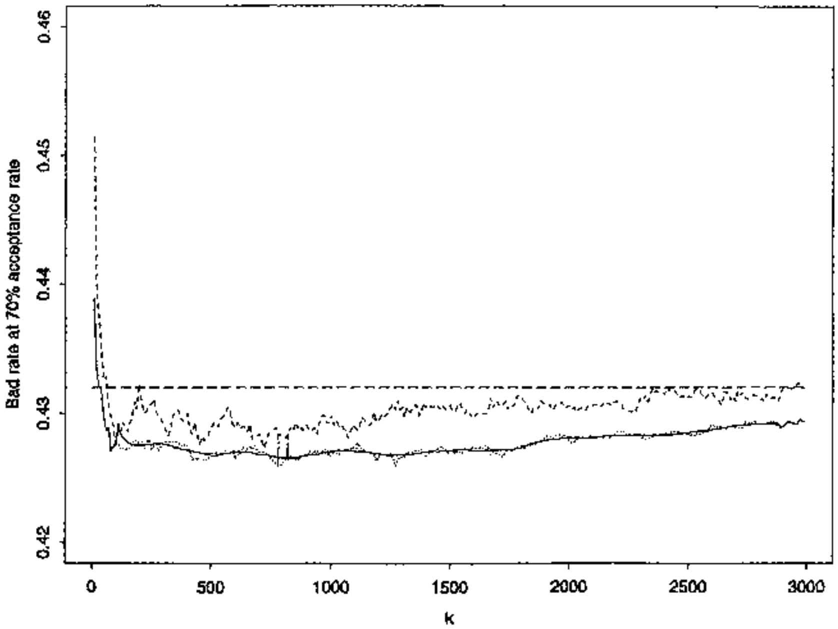

Fig. 9. Smoothed design set k-NN method using data set 5 and $D = 1.40$: ———, smoothed design; ·······, raw design; - - - - - -, test set; — — —, linear regression

We note that removing the decision tree characteristic does not lead to a significant change in relative performance of the k-NN and linear regression methods. For example, if we repeat the analysis on sample 4 without the decision tree characteristic, the linear regression gives a bad risk rate of 42.64 (among the 70% accepted) and the k-NN method gives a bad risk rate of 41.94. The difference of 0.70 is only slightly higher than the difference of 0.61 in Table 5. This result was found to be consistent for other samples. We conclude that including the decision tree characteristic in the analysis gives similar extra benefit to the two classification methods. In the rest of this paper it will be included.

To assess further the potential of the k-NN method with adjusted Euclidean metrics for credit scoring, we extended the comparisons to include logistic regression, decision trees (e.g. Breiman *et al.* (1984)) and decision graphs. Decision graphs are an extension of decision trees which permit paths to merge and are described by Oliver (1992).

Table 6 shows the bad risk rates averaged over samples for the range of classification techniques considered. The results show that the k-NN method with adjusted Euclidean metrics gives the lowest bad risk rate. For a more detailed discussion of the comparisons see Henley (1995).

3.5. *Sensitivity of k-nearest-neighbour method to changes in population priors*

In the comparisons described in this paper the sample proportion of bad accounts is approximately 50%. This is typical of the design sample bad risk rates used in many previous studies, such as Myers and Forgy (1963), Orgler (1970), Apilado *et al.* (1974) and Chandler and Coffman (1983). We can justify this approach, independently of the true population bad risk rate, because it enables the most accurate estimation of the ratio of class conditional probabilities for a fixed total sample size. In practice the proportion of bad risks in the full population will vary according to the credit product and will commonly be less than 20%. To try to maximize the range of validity of our results we undertook some further analysis to investigate the relative performance of the k-NN method and linear regression for various population priors.

TABLE 5
Bad risk rates for linear regression and *k*-NN models for samples 1–5

Sample	*Rates for the following methods:*	
	Linear regression	*k-NN*
1	43.87	43.96
2	44.03	43.63
3	43.54	43.45
4	42.16	41.55
5	43.20	42.84

In this paper we restrict attention to the results obtained when the full population bad risk rate is assumed to be 20%. To simulate this scenario, we applied weights of 4 and 1 respectively to the good and bad applicants in the original test sample and classified the reweighted test sample by using the original design sample. This approach was adopted for both the *k*-NN method and linear regression. (The unweighted design sample was used to maximize the number of applicants from each class available for estimating $P(g \mid \mathbf{x})$, as argued above. Furthermore, the *k*-NN parameters were estimated from the unweighted design sample. In fact, using a weighted design sample gave inferior performances for both methods.)

Table 7 shows the bad risk rates at a 70% acceptance rate for each of the five reweighted test samples by using linear regression and the *k*-NN method. The bad risk rate curves for the design sets (but not the test sets) were smoothed as before. The results shown in Table 7 are consistent with the conclusions of Section 3.4. In particular, we note than the *k*-NN method gives lower bad risk rates than does linear regression for four of the five samples. As we found previously the differences are fairly small in each case but may be sufficiently large to have commercial implications. The averaged bad risk rates for the *k*-NN method and linear regression are 12.57 and 12.74 respectively.

It was also found that the curves of bad risk rate against *k* and *D* exhibited similar properties to the curves shown in Figs 1–4. For example, Fig. 10 shows design and test curves of the bad risk rate against *k* for sample 4 by using the weighted test sample. (The values of the design curve have been translated to fit on the graph.) In particular we note that the test set curve has a broad flat valley, indicating that this property does not result from a high population bad risk rate.

To conclude, our sensitivity analysis has indicated that the *k*-NN method can give a consistently good performance for a wide range of values of the population bad risk rate. Furthermore, we assert that, regardless of the population priors, the *k*-NN design set should consist of equal proportions of good and bad applicants. The *k*-NN parameters *k* and *D* should also be estimated from this data set.

TABLE 6
Summary of the averaged bad risk rate results for various classification techniques

Method	*Bad risk rate*
k-NN (any *D*)	43.09
k-NN ($D = 0$)	43.25
Logistic regression	43.30
Linear regression	43.36
Decision graphs or trees	43.77

TABLE 7
Bad risk rates for linear regression and k-NN models when the population bad risk rate is 20%

Sample	*Rates for the following methods:*	
	Linear regression	*k-NN*
1	13.18	12.95
2	13.13	12.84
3	12.47	12.44
4	12.27	11.78
5	12.65	12.82

3.6. *Practicality of implementation*

In this paper we have shown that the k-NN method with adjusted Euclidean metrics can give slightly improved prediction of consumer credit risk than traditional techniques such as linear and logistic regression. However, for this result to be useful to credit grantors, it needs to be established that the methodology can be implemented in practice. We consider this issue below with reference to four specific points.

3.6.1. *Time taken to score or classify new applicant for credit.* Although assessing an applicant by using the k-NN method requires considerably more computations than linear or logistic regression (because it requires the calculation of the distance to all points in the design set), given the power of a modern 486 personal computer there is little practical difference. If the design set is

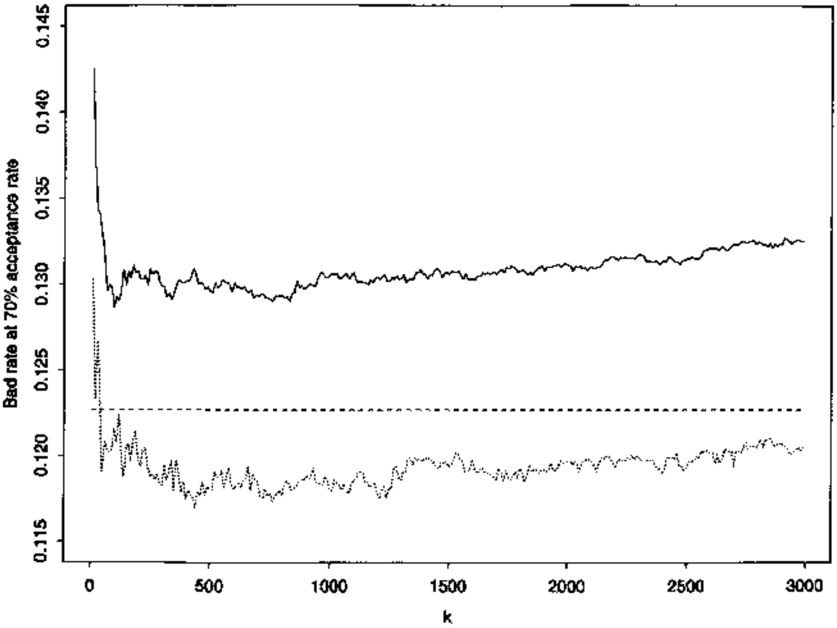

Fig. 10. k-NN method using weighed data set 4 and $D = 1.40$

available on line then the *k*-NN method (with or without design set smoothing) can classify an applicant in less than 1.5 s for any value of *k* up to 3000. Furthermore, we have not optimized the time efficiency of our algorithm, so it is potentially possible to reduce this value further.

If an applicant's predicted class membership probabilities are averaged for a range of *k* values (equivalent to smoothing the test set) then this makes a negligible difference to the time to classify an applicant. This is because, having worked out the $k+h$ NNs, the $k+h-1 \ldots k-h$ NNs are automatically available in our algorithm.

There are variants of the standard *k*-NN methodology which enable processing time to be cut still further by reducing the number of distances which need to be recalculated for each new applicant. For example, consider the simple case of one NN. By storing all the pairwise differences between points in the design sample we can significantly reduce the number of points for which distances need to be recalculated. This is because, given a distance *d* between the new applicant and a point in the design sample, any point more than distance 2*d* from this design set point cannot be the NN. Fukanaga and Narenda (1975) presented a method of searching for the *k*-NNs by using the branch-and-bound algorithm (see Hand (1981)). Further reductions in the number of computations required can be achieved by using an edited or condensed *k*-NN approach (see Gates (1972) and Hand and Batchelor (1978)).

3.6.2. *Automatic updating of classification rule.* One potential attraction of the *k*-NN method is that dynamic updating of the design set is easy. As (recently) accepted customers reach the point where their performance is assessed as good or bad, they can be added to the design set to replace the oldest surviving observation.

3.6.3. *Reasons for refusal of credit.* The *k*-NN method with adjusted Euclidean metrics could provide a reason for refusal of credit by exploiting the information about class separation in the data provided by the regression weights. To do this we could calculate the (standardized) distance between a new applicant and the mean for the design set bad applicants for each characteristic in the model. This gives an approximation to the contribution of each characteristic to the decision. Characteristics with low values of this distance are the characteristics which identify the applicant with previous bad applicants and, thus, ensure that the applicant is rejected. Of course, with any multivariate classification technique (including linear or logistic regression) it is not possible to reduce a decision explicitly to the value of a single variable.

3.6.4. *Red lining of applicants.* A charge sometimes levelled at credit grantors is that their methods of screening applicants red line people (refuse credit on the basis of one characteristic regardless of all other attributes). We believe that the *k*-NN method is less susceptible to this criticism than linear or logistic regression because our proposed distance metric takes into account random variation as well as the distance along the equiprobability contours (estimated from the regression). However, as the value of *D* increases, the *k*-NN decision becomes increasingly dependent on characteristics with large weights w_i from the regression. One way to reduce this problem would be to put a limit on the ratio of the regression weights w_i. This constraint would be likely to reduce the classification accuracy of the regression score-card by more than that of the *k*-NN classifier.

4. Conclusions

In this paper we have considered the problem of choosing an appropriate technique for discriminating between a population of good and bad credit risks. The criterion used for assessing performance, the minimization of the bad risk rate among those accepted, is rather unusual for a classification problem and so was discussed in detail. In particular, it places bounds on the level of success that can be achieved.

We proposed on application of the k-NN method using an original distance measure, an adjusted version of the standard Euclidean metric, and set out a strategy for selecting values of the parameters k and D. We have seen that, in fact, our k-NN classification rule is fairly insensitive to the choice of these parameters and, in particular, the curves of bad risk rate against k have surprisingly flat valleys. Other interesting features of the results have been discussed.

A comparison was made between the performance of the k-NN method and a range of other classification techniques. Linear and logistic regression and decision trees were selected to represent the accepted credit scoring techniques and decision graphs were included to represent a recent development in the classification literature. It was found that the k-NN method performed well, achieving the lowest expected bad risk rate. It was also found that the adjusted Euclidean metric led to an improvement over the standard Euclidean metric. In Section 3.5 we confirmed the validity of our results for populations with lower proportions of bad risks in the full population. It was argued that a design set with equal proportions of good and bad risks should be used to classify future applicants regardless of the population bad risk rate.

Finally, in Section 3.6 we asserted that it is practical to implement a k-NN classification rule for credit scoring new applicants in realtime. It was found that, given the power of modern computers, it is feasible to classify an applicant within seconds and that the k-NN method can provide reasons for refusing credit, thus satisfying future legal requirements. Furthermore, the k-NN method has the advantage over traditional score-card approaches that the classification rule can be dynamically updated to reflect changes in the nature of the population.

Acknowledgement

The first author was supported in this research by a grant from Littlewoods Plc.

References

Apilado, V. P., Warner, D. C. and Dauten, J. J. (1974) Evaluative techniques in consumer finance—experimental results and policy implications. *J. Finan. Quant. Anal.*, Mar., 275–283.

Boyle, M., Crook, J. N., Hamilton, R. and Thomas, L. C. (1992) Methods for credit scoring applied to slow payers. In *Proc. Conf. Credit Scoring and Credit Control* (eds L. C. Thomas, J. N. Crook and D. B. Edelman), pp. 75–90. Oxford: Clarendon.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees.* Belmont: Wadsworth.

Chandler, G. C. and Coffman, J. Y. (1983) Applications of performance scoring of accounts receivable management in consumer credit. *J. Ret. Bank.*, **5**, no. 4, 1–10.

Cover, T. M. and Hart, P. E. (1967) Nearest neighbour pattern classification. *IEEE Trans. Inform. Theory*, **13**, 21-27.

Crook, J. N., Hamilton, R. and Thomas, L. C. (1992) A comparison of discriminators under alternative definitions of credit default. In *Proc. Conf. Credit Scoring and Credit Control* (eds L. C. Thomas, J. N. Crook and D. B. Edelman), pp. 217–245. Oxford: Clarendon.

Durand, D. (1941) Risk elements in consumer instalment financing. In *Financial Research Program*, *Study 8*. New York: National Bureau of Economic Research.

Eisenbeis, R. A. (1978) Problems in applying discriminant analysis in credit scoring models. *J. Bank. Finan.*, **2**, 205–219.

Fix, E. and Hodges, J. (1952) Discriminatory analysis, nonparametric discrimination: consistency properties. *Report 4, Project 21-49-004.* US Air Force School of Aviation Medicine, Randolph Field.

Fogarty, T. C. and Ireson, N. S. (1993) Evolving Bayesian classifiers for credit control—a comparison with other machine learning methods. *IMA J. Math. Appl. Bus. Ind.*, **5**, no. 1, 63–75.

Fukanaga, K. and Flick, T. E. (1984) An optimal global nearest neighbour metric. *IEEE Trans. Pattn Anal. Mach. Intell.*, **1**, 25–37.

Fukanaga, K. and Narendra, P. M. (1975) A branch and bound algorithm for computing k-nearest neighbours. *IEEE Trans. Comput.*, **24**, 750–753.

Gates, G. W. (1972) The reduced nearest neighbour rule. *IEEE Trans. Inform. Theory*, **18**, 431.

Gilbert, L. R., Menon, K. and Schwartz, K. B. (1990) Predicting bankruptcy for firms in financial distress. *J. Bus. Finan. Account.*, **17**, 161–171.

Grablowsky, B. J. and Talley, W. K. (1981) Probit and discriminant functions for classifying credit applicants: a comparison. *J. Econ. Bus.*, **33**, 254–261.

Hand, D. J. (1981) Branch and bound in statistical data analysis. *Statistician*, **30**, 1–13.

——— (1982) *Kernel Discriminant Analysis*. Letchworth: Research Studies Press.

Hand, D. J. and Batchelor, B. G. (1978) An edited nearest neighbour rule. *Inform. Sci.*, **14**, 171–180.

Henley, W. E. (1995) Statistical aspects of credit scoring. *PhD Dissertation*. Department of Statistics, The Open University, Milton Keynes.

Leonard, K. J. (1993) Empirical Bayes analysis of the commercial loan evaluation process. *Statist. Probab. Lett.*, **18**, 289–296.

Myers, J. H. and Forgy, E. Q. (1963) The development of numerical credit evaluation systems. *J. Am. Statist. Ass.*, **58**, 799–806.

Myles, J. (1991) The use of *k*-nearest neighour methods in statistical pattern recognition. *PhD Dissertation*. Department of Statistics, The Open University, Milton Keynes.

Oliver, J. J. (1992) Decision graphs—an extension of decision trees. *Technical Report 92/173*. Computer Science Department, Monash University, Clayton.

Orgler, Y. E. (1970) A credit scoring model for commercial loans. *J. Money Credit Bank.*, Nov., 435–445.

Reichert, A. K., Cho, C. C. and Wagner, G. M. (1983) An examination of the conceptual issues involved in developing credit scoring models. *J. Bus. Econ. Statist.*, **1**, 101–114.

Short, R. D. and Fukunaga, K. (1982) The optimal distance measure for nearest neighbour classification. *IEEE Trans. Inform. Theory*, **27**, 622–627.

Srinivasan, V. and Kim, Y. H. (1987) Credit granting: a comparative analysis of classification procedures. *J. Finan.*, **92**, 665–681.

Terrell, G. R. and Scott, D. W. (1992) Variable kernel density estimation. *Ann. Statist.*, **20**, 1236–1265.

Wiginton, J. C. (1980) A note on the comparison of logit and discriminant models of consumer credit behaviour. *J. Finan. Quant. Anal.*, **15**, 757–770.