Biometrika (2003), **90,** 1, *pp.* 99–112
© 2003 Biometrika Trust
Printed in Great Britain

Likelihood inference in nearest-neighbour classification models

By CHRISTOPHER C. HOLMES and NIALL M. ADAMS
Department of Mathematics, Imperial College, London, SW7 2BZ, U.K.
c.holmes@ic.ac.uk n.adams@ic.ac.uk

Summary

Traditionally the neighbourhood size k in the k-nearest-neighbour algorithm is either fixed at the first nearest neighbour or is selected on the basis of a crossvalidation study. In this paper we present an alternative approach that develops the k-nearest-neighbour algorithm using likelihood-based inference. Our method takes the form of a generalised linear regression on a set of k-nearest-neighbour autocovariates. By defining the k-nearest-neighbour algorithm in this way we are able to extend the method to accommodate the original predictor variables as possible linear effects as well as allowing for the inclusion of multiple nearest-neighbour terms. The choice of the final model proceeds via a stepwise regression procedure. It is shown that our method incorporates a conventional generalised linear model and a conventional k-nearest-neighbour algorithm as special cases. Empirical results suggest that the method out-performs the standard k-nearest-neighbour method in terms of misclassification rate on a wide variety of datasets.

Some key words: Maximum pseudolikelihood; k nearest neighbour; Nonparametric classification; Probabilistic nearest neighbour.

1. Introduction

Within the field of statistical pattern recognition and machine learning the k-nearest-neighbour algorithm, also known as instance-based learning or lazy learning (Aha, 1997), is one of the most commonly used tools for prediction. The algorithm is remarkably simple and has remained largely unchanged since its introduction in an unpublished USAF School of Aviation Medicine report by E. Fix and J. L. Hodges. The k-nearest-neighbour procedure uses a training dataset $\{y_i, x_i\}_{i=1}^n$ to make predictions on new unlabelled data, where $y_i \in \{C_1, \ldots, C_Q\}$ denotes the class label of the ith point and x_i denotes a vector of p predictor variables. The prediction for a new point, $y_{n+1} | x_{n+1}$ is reported as the most common class found amongst the k nearest neighbours of x_{n+1} in the set $\{x_i\}_{i=1}^n$. The neighbours of a point are defined via a distance metric $\rho(x_{n+1}, x_i)$ which is commonly taken to be the Euclidean norm. A degree of confidence in the prediction can be provided by the relative counts of each category within the k neighbours. Dasarathy (1991) provides an overview of methods and a comprehensive collection of around 140 key papers.

The k-nearest-neighbour algorithm is a nonparametric procedure in that it makes no assumption about the distribution of the underlying class conditional density $\rho(x | y)$, and it can be shown for a suitable choice of k that the error rate converges with n to the Bayes risk; see Ripley (1996, Ch. 6) for details. The 'vanilla' algorithm with Euclidean norm has just one parameter k and this method remains the most popular, partly because of its simplicity but also because of empirical evidence that shows the approach to be effective at prediction on a wide variety of datasets (Michie et al., 1994, Ch. 9).

Traditionally, the value of k is chosen by crossvalidation on the misclassification rate (Mitchell, 1997, Ch. 4; Ripley, 1996). In this paper we propose an alternative approach by defining a conditional probability model for the data $(y_1, \ldots, y_n)$ that is specified by a function of the k nearest neighbours and a single interaction parameter. Optimisation of k can then proceed by a method of maximum pseudolikelihood. We illustrate how the crossvalidation approach and our method both seek to maximise an additive cost function of the data. The form of the cost function highlights the difference between the approaches.

The probability model we suggest has the form of a generalised linear model (McCullagh & Nelder, 1989) on a set of nonlinear autocovariates defined by the class labels of the nearest neighbours to each point. The method is shown to possess all of the properties of the standard k-nearest-neighbour algorithm together with a number of key advantages: the introduction of a probability model for $(y_1, \ldots, y_n)$ facilitates the use of the model within a formal decision process where predictions lead to actions with associated costs on outcomes; we can consider the choice of k as a generalised linear model variable-selection problem for which a large body of theory exists; we can consider the inclusion of the original predictors as linear terms within the model; and we can consider multiple values for k which can be used to capture nonlinear effects that operate at different scales within the data.

Nearest-neighbour models implicitly assume that the expectation $E(y|x)$ is adequately approximated by a local mean fit to the data, using the nearest k neighbours. The extension to include linear terms assumes that $E(y|x)$ is smoothly changing around possible global linear effects. A further potential advantage of our approach is that we can make use of generalised linear model diagnostic tools, such as those discussed in Hosmer & Lemeshow (2000, Ch. 5), to check the validity of our modelling assumptions; see § 3.

The problem of which linear, or which k-nearest-neighbour terms to include is handled using a stepwise model-building procedure. In this manner, our method incorporates as special cases a conventional generalised linear model, that uses just x, and the conventional k-nearest-neighbour algorithm that uses just nearest neighbours. We demonstrate that the model-fitting procedure for the two-class classification problem can be carried out using any standard statistical software package that supports stepwise variable-selection for the generalised linear model.

Our model is closely related to the autologistic model discussed in Besag (1974, 1986) and the coloured lattice model presented in Strauss (1977); see also Geman & Geman (1984). In fact our model can be viewed as an extension of these methods to include multiple neighbourhood terms and to deal with nonspatial data in general high-dimensional classification problems. Besag (1974) proposes the method of maximum pseudolikelihood for estimation in these models and we use this idea here. Maximum pseudolikelihood estimators have similar large-sample properties to their maximum likelihood counterparts including asymptotic consistency and asymptotic normality but not asymptotic efficiency (Mase, 1995; Jensen & Kunsch, 1994).

This paper is a development of the approach of Holmes & Adams (2002), who consider two-class classification models with a single k-neighbourhood, using a Bayesian approach with a prior distribution on k. Predictions are made via numerical integration. Here we consider multi-class problems and, more importantly, multiple predictors including a mixture of multiple nearest-neighbour terms and global linear effects. Our estimation follows pseudolikelihood maximisation, which is computationally much more efficient than the

Bayesian analysis, especially in the context of selecting multiple predictors. Some of the benchmark datasets analysed in § 3 were previously analysed in Holmes & Adams (2002). The results here suggest that improvements in predictive accuracy can be obtained with the multiple-predictor method.

In § 2 we describe the model and discuss implementation issues. We illustrate the method on a number of benchmark datasets in § 3, and in § 4 we provide a brief discussion. A full Matlab version of our method is available on request from C. C. Holmes.

2. Nearest-neighbour classification and pseudolikelihood inference

2·1. *The model*

To begin, we consider the two-class classification problem, $y_i \in \{C_0, C_1\}$, using a single value for the neighbourhood size k. We propose a sampling distribution for y_i which for the two-class problem is taken to be the Bernoulli distribution, $y_i \sim \text{Ber}(\eta_i)$, with a mean parameter η_i that is determined by a function of the class labels of the k nearest neighbours of x_i. In particular, we rewrite the k-nearest-neighbour procedure as a generalised linear regression on a nonlinear nearest-neighbour autocovariate using the conditional distribution

$$\text{pr}(y_i = C_1 \mid y_{-i}, x_i) = \eta_i = \frac{\exp\{\beta z_i^{(k)}(C_1, x_i)\}}{1 + \exp\{\beta z_i^{(k)}(C_1, x_i)\}}, \tag{1}$$

where y_{-i} denotes the data with the ith observation removed, β is an interaction regression parameter and $z_i^{(k)}(C_1, x_i)$ is a k-nearest-neighbour autocovariate, defined by

$$z_i^{(k)}(C_1, x_i) = \frac{1}{k} \sum_{j \overset{k}{\sim} i} \{I(y_j = C_1) - I(y_j = C_0)\}. \tag{2}$$

In (2), $I(a) = 1$ if condition a is true, $I(a) = 0$ otherwise and $\sum_{j \overset{k}{\sim} i}$ denotes that the summation is over the k nearest neighbours of x_i in the set $\{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$, given the distance metric $\rho(.)$. The autocovariate, which from now on we simply write as $z_i^{(k)}$, records the proportion of class C_1's to class C_0's within the k nearest neighbours of x_i. In particular, $z_i^{(k)} = 1$ if all of the k nearest neighbours of x_i are of class C_1, $z_i^{(k)} = -1$ if all the neighbours are of class C_0, and $z_i^{(k)} = 0$ if there is a tie.

Restricting β to $\Re^+$ in (1), we find that the decision boundary, defined by $\text{pr}(y_i = C_1) = 0{\cdot}5$, is identical to that of the conventional k-nearest-neighbour algorithm for given k; that is, in (1) the most probable class for a new point y_{n+1} is given by the most common class among its k nearest neighbours and the level of confidence in the prediction is monotonic in the number of counts for that class.

Expression of the model in terms of the conditional predictive densities (1) is equivalent to the autologistic method for smoothing binary spatial fields (Besag, 1972, 1974). Besag (1974) discusses the advantages of defining the model via the conditional distributions rather than the full joint probability distribution.

In order to estimate the regression parameter β, conditional on k, we mimic the pseudolikelihood approach of Besag (1974) and set β to maximise

$$\prod_{i=1}^{n} \text{pr}(Y_i = y_i \mid y_{-i}, x_i).$$

Computational issues in fitting such models are well developed. For instance, the model

(1) has the form of a generalised linear model in Z and hence the calculation of β can be performed via iteratively reweighted least squares (McCullagh & Nelder, 1989, p. 114). We let $\hat{\beta}_k$ denote the maximum pseudolikelihood estimate of β conditional on k. It can be seen that $2\hat{\beta}_k$ is interpretable as the change in the log-odds of class C_1 relative to class C_0 when observing a point of class C_1 in the k nearest neighbourhood of x_i.

Having written the k-nearest-neighbour algorithm in generalised linear model form we then propose our best estimate for k as the value that maximises the profile pseudolikelihood:

$$\hat{k} = \arg\max_k \operatorname{lik}(z^{(k)}, \hat{\beta}_k),$$

where $z^{(k)} = (z_1^{(k)}, \ldots, z_n^{(k)})'$ is the set of k-nearest-neighbour autocovariate values at the n data points and k is defined over a suitable range, $k \in \{1, 2, \ldots, k_{\text{max}}\}$. From our assumption of Bernoulli conditional likelihood we obtain

$$\operatorname{lik}(z^{(k)}, \hat{\beta}_k) = \prod_{i=1}^{n} \eta_i^{y_i}(1 - \eta_i)^{(1-y_i)},$$

with

$$\eta_i = \frac{\exp(\hat{\beta}_k z_i^{(k)})}{1 + \exp(\hat{\beta}_k z_i^{(k)})},$$

where we take $y_i \in \{0, 1\}$ to be the class label.

Maximum pseudolikelihood selects a $\hat{k}$ and $\hat{\beta}_k$ that maximises the joint conditional density of the class indicators given the data. This is in contrast to the conventional approach using crossvalidation on misclassification rate which concentrates solely on trying to place points on the correct side of the decision boundary. The difference between the two approaches is best illustrated by examining the form of the additive cost function that both procedures are attempting to maximise: we have

$$\hat{k} = \arg\max_k \sum_{i=1}^{n} l\{f(x_i \mid k) \mid y_i\},$$

where $l(.)$ is the cost function and $f(x_i \mid k)$ is the prediction for y_i given x_i and k. For the maximum pseudolikelihood method the cost function $l(. \mid y_i)$ is the loglikelihood and $f(x_i \mid k)$ is the probability forecast $\operatorname{pr}(y_i = C_1)$. For the standard crossvalidation method the predictions are just the normalised counts of the number of C_1's among the k nearest neighbours, $f(x_i \mid k) = k^{-1} \sum_{j \overset{k}{\sim} i} I(y_j = C_1)$, and the cost function is

$$l\{f(.) \mid y_i\} = cI\{|f(.) - y_i| > 0{\cdot}5\}$$

for some constant $c < 0$. The cost of a prediction versus the actual prediction for a point belonging to class C_0, $y_i = 0$, is illustrated in Fig. 1 for both methods. Note that the change in cost associated with a change in prediction is a smooth function for the likelihood method, whereas for the crossvalidation procedure we have a discontinuous step function at $f(x_i \mid k) = 0{\cdot}5$. This has the potential to make the results of crossvalidation sensitive to points that lie near the decision boundary. Note that the scaling of the cost function for the crossvalidation method is arbitrary. We believe that the smoothness of the cost function for the maximum pseudolikelihood method is in part responsible for the improved empirical performance highlighted in § 3.

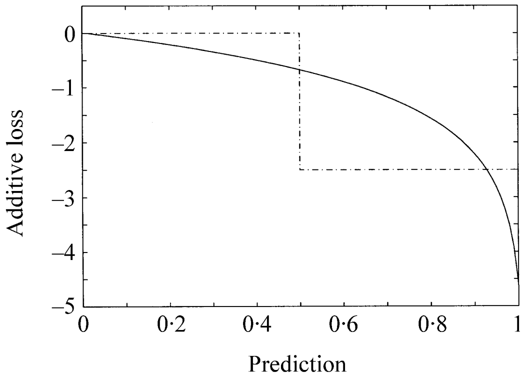

Fig. 1. Additive loss $l\{f(x_i\,|\,k)\,|\,y_i = C_0\}$ versus prediction $f(x_i\,|\,k)$ for the maximum pseudolikelihood method (solid) and crossvalidation method (dot-dashed) for a data point for which $y_i = C_0$.

2·2. *Extension to multiple predictors including linear covariates*

Having written the k-nearest-neighbour algorithm as a generalised linear model we find that our procedure of selecting $\hat{k}$ is equivalent to a generalised linear model variable-selection problem where the $n \times k_{\text{max}}$ design matrix is of the form

$$Z = (z^{(1)}\ \ z^{(2)}\ \ \ldots\ \ z^{(k_{\text{max}})}),$$

and the task previously was to select a single 'best' column from Z using profile pseudolikelihood. The selection of a single column of Z appears an artificial restriction when working with generic classification tasks within a generalised linear model paradigm, and hence we might like to extend the procedure to consider all of the $2^{k_{\text{max}}}$ subsets of Z as potential models for the data. This equates to allowing multiple values of k to be included within the model. This can be extremely useful when nonlinear effects are operating at different scales within the data. For instance, small values of k capture very local variations in the underlying class probability surface while larger values of k record smooth underlying trends. Both of these features may be present in a dataset and in this case the inclusion of both terms will improve the model. An example of this is provided in § 3·3.

Furthermore, the autocovariates in Z model nonlinear effects in the data. However, it may be that some of the original variables have a global linear relationship with the response. This leads us to consider the original predictor variables as possible predictors within the model. The augmented design matrix is then,

$$S = (1\ \ X\ \ Z),$$

where S is an $n \times p^*$ matrix with $p^* = p + k_{\text{max}} + 1$, starting with a column vector of ones, and X is the original matrix of p predictor variables. By entertaining all possible subsets of columns of S we include as special cases in the model space the conventional generalised linear model, using the first $p + 1$ columns of S, as well as a conventional k-nearest-neighbour method, using a single column of Z.

The extension to include multiple columns of S leads to the following model for the two-class problem:

$$\text{pr}(y_i = C_1) = \frac{\exp(s_i^{(\gamma)\text{T}}\beta)}{1 + \exp(s_i^{(\gamma)\text{T}}\beta)},$$

where γ is a $1 \times p^*$ indicator vector, $\gamma = \{\gamma_1, \ldots, \gamma_{p^*}\}$ with $\gamma_i \in \{0, 1\}$ such that $\gamma_i = 1$ if the ith column of S is to be included in the model and $y_i = 0$ if the ith column is excluded. We use $m = \sum_{i=1}^{p^*} \gamma_i$ to record the number of columns included, so that β is an $m \times 1$ vector of regression parameters and $s_i^{(\gamma)}$ is an $m \times 1$ vector of values extracted from the ith row of S using the columns indicated by γ.

The problem of which variables to include in a generalised linear model is usually solved by ranking models using a model choice criterion which takes into account the complexity of the model as well as the goodness of fit. The 'best' model under the criterion is then adopted. We use the Bayesian information criterion, BIC (Akaike, 1977, 1978; Schwartz, 1978), which is closely related to the principle of minimum description length (Hansen & Yu, 2001). The BIC is defined as

$$\text{BIC} = -2 \text{ maximised } \log(\text{pseudolikelihood}) + \log(n) \times \text{number of parameters}.$$

The user is free of course to adopt other criteria such as crossvalidation if they so wish. One advantage of using BIC is that it is a standard option in many statistical software packages for generalised linear model building. The significance of this last point is discussed below.

The variable selection procedure for our model is complicated because there are models to be tested. We adopt a stepwise selection approach (Venables & Ripley, 1999, p. 186; Draper & Smith, 1981, p. 337). We begin by ranking by BIC score all models formed by adding a single variable to the null model. If the addition of the 'best' single variable results in a lower BIC score then that variable is retained. The process is then repeated, seeking to add another variable, until no additional variable can be found to lower the BIC. We then attempt backward deletion by testing all of the variables in the model for deletion and the submodels are ranked by BIC. If the deletion of the 'worst' variable results in an overall lower BIC score then that variable is removed and the procedure is repeated, until no variable can be removed without causing an increase in BIC score. The final model, $S_{\hat{\gamma}}$, indexed by $\hat{\gamma}$, is then reported as the 'best' subset of S. It is extremely simple to implement this procedure using standard packages such as S-Plus.

2·3. *Extenstion to multinomial likelihood*

In extending the above method for the two-class classification problem to the more general case of multinomial data $y_i \in \{C_0, \ldots, C_Q\}$, it is instructive to begin by considering the model with a single k-nearest-neighbour component for which we write the probability model as

$$\text{pr}(y_i = C_j \mid y_{-i}, x_i) = \frac{\exp(z_i^{(k,j)}\theta)}{\sum_{v=0}^{Q} \exp(z_i^{(k,v)}\theta)}, \tag{3}$$

where θ is a single regression parameter and the multiple class autocovariate $z_i^{(k,v)}$ is in relation to class C_0 by

$$z_i^{(k,v)} = \frac{1}{k} \sum_{j \overset{k}{\sim} i} \{I(y_j = C_v) - I(y_j = C_0)\},$$

so that the autocovariate $z_i^{(k,v)}$ records the proportion of class C_v's to class C_0's in the k nearest neighbours of x_i and hence $z_i^{(k,0)} = 0$ for all i, k. As for the two-class model in (1) the multinomial model (3) produces identical decision boundaries to the conventional k-nearest-neighbour algorithm for given k. The model in (3) is equivalent to the spatial system considered by Strauss (1977) on coloured lattices, also known as the Potts model.

The construction of the multinomial model (3) reveals a difference between itself and a conventional generalised linear model. In the latter the covariate values remain constant across classes while the regression parameters differ. In contrast, in (3) it is the autocovariate that changes across classes while the regression parameter θ remains constant. This ensures that the predictive class boundaries match those of the conventional k-nearest-neighbour model.

We can extend the spatial model considered by Strauss (1977) in (3) to the general classification situation by taking multiple covariates using multiple columns of the original predictors x and multiple values for the autocovariates Z

$$\mathrm{pr}(y_i = C_j) = \frac{\exp(x_i^{(\gamma^x)}\beta_j + z_i^{(\gamma^z, j)}\theta)}{\sum_{v=0}^{Q} \exp(x_i^{(\gamma^x)}\beta_v + z_i^{(\gamma^z, v)}\theta)},$$

where γ^x and γ^z indicate which linear and which autocovariate variables are included respectively, and hence define the dimensions of β and θ. Following standard generalised linear model practice we fix $\beta_0 = 0$.

The departure from the conventional generalised linear model framework in (4) means that standard statistical software procedures cannot be used to obtain maximum pseudolikelihood estimates for the model parameters $\{\beta_1, \ldots, \beta_Q, \theta\}$. Instead, in our implementation we obtain the maximum pseudolikelihood estimates using the Matlab optimisation function `fminunc.m`, which performs a Broyden–Fletcher–Goldfarb–Shanno quasi-Newton method (Press et al., 1990, Ch. 10) with a mixed quadratic and cubic line search. This finds the maximum pseudolikelihood estimates for the model parameters and from there the selection of which subsets of linear predictors and nonlinear autocovariates to include proceeds, as before, by stepwise selection based on the BIC score.

3. Examples

3·1. *Preamble*

In this section we analyse five publicly available benchmark datasets and a real world classification problem. For the benchmark tests in §§ 3·2 and 3·3 we are particularly interested in how our method compares with the conventional k-nearest-neighbour algorithm. We use the standard performance measure of average out-of-sample prediction cost,

$$T = \frac{1}{n^{(t)}} \sum_{i=1}^{n^{(t)}} L(y_i, q) \times I\left\{ y_i \neq \arg\max_q \mathrm{pr}(y_i = q \mid x_i, Y, x) \right\},$$

where the summation is over a test set of $n^{(t)}$ points, given a training set (Y, X), and $L(y_i, q)$ is the cost of misclassifying class y_i as class q.

In each example we use the stepwise procedure described in § 2 to select the model with the parameters estimated by maximum pseudolikelihood. We take $k_{\max} = \min(n, 200)$, where n is the number of training points. Higher values for k could be considered if it appeared that the data supported this. The two-class datasets used in §§ 3·2 and 3·3 have previously been studied in Holmes & Adams (2002).

3·2. *Synthetic two-class problem*

The task is a binary classification problem where each class is drawn from a mixture of two bivariate normal distributions; see Ripley (1994, 1996, Ch. 1). The data can be obtained from `www.stats.ox.ac.uk/pub/PRNN/`, where a designated training set of 250 points and an out-of-sample test set of 1000 points are available.

In Ripley (1994) 15 classification methods were tested on this dataset including k-nearest-neighbour, logistic regression and a number of neural network models. We ran our stepwise procedure which terminated having chosen a single autocovariate with $\hat{k} = 66$ and no linear component. This suggests that the problem is intrinsically nonlinear. The model gave an average test error rate of 8·2% which would rank above all the 15 methods that Ripley analysed. The training data plus predictive probability contours $\mathrm{pr}(y_i = C_1) = \{0{\cdot}1, 0{\cdot}3, 0{\cdot}5, 0{\cdot}7, 0{\cdot}9\}$ for our maximum pseudolikelihood method are shown in Fig. 2(a), where we see that the model is able to capture the nonlinear class boundary of this problem. For comparison, in Fig. 2(b) we show the classification boundary $\mathrm{pr}(y_i = C_1) = 0{\cdot}5$ of our model, dashed line, alongside the 5-nearest-neighbour boundary, solid, reported by Ripley. The 5-nearest-neighbour method had a test error rate of 13·0%, placing it 12th overall. Clearly the higher value of $\hat{k}$ chosen by maximum pseudolikelihood produces a smoother curve and this leads, in this example, to a much improved error rate. The best performance quoted by Ripley on this dataset is 8·3%, for an edited version of 5-nearest-neighbour method. For these data then, our approach is highly effective, and, as shown in Fig. 2, has the advantage of giving probabilistic outputs which are not obtained using standard k-nearest-neighbour methods. The selection of what appears to be a high value of $\hat{k}$ for our model is typical of the method as a whole; see § 4. The decision boundary shown in Fig. 2 is not as smooth as that reported in Holmes & Adams (2002). However, the test error rate is lower, suggesting a better fit to the data.

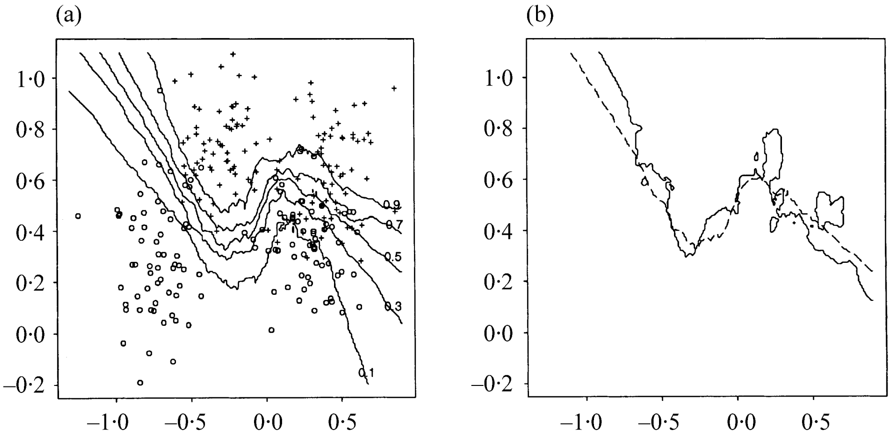

Fig. 2. Synthetic two-class problem (Ripley, 1994). (a) shows the training data, $y_i\{+, \mathrm{o}\}$, with predictive probability contours for the maximum pseudolikelihood method using a single autocovariate $\hat{k} = 66$. (b) shows the decision boundary $\mathrm{pr}(y_i = C_1) = 0{\cdot}5$ for the maximum pseudolikelihood method (dashed line) and 5-nearest-neighbour method (solid).

3·3. *Benchmark series*

We next consider a collection of five datasets available from the University of California, Irvine machine learning repository (Blake & Merz, 1998), four of which were used in the Statlog project (Michie et al., 1994). The datasets we used were Australian Credit, Diabetes, Heart, German Credit and the Vowel data. Characteristics of these datasets and the analysis that we undertook are given in Table 1. The first four datasets are two-class problems from Statlog.

To assess performance we replicated the evaluation procedure undertaken in Statlog

Table 1. *Benchmark dataset characteristics and algorithm performance. Test method refers to the manner in which performance is estimated, either N-fold crossvalidation or the size of an independent test set. The* MPL *column quotes estimated test performance for our maximum pseudolikelihood method, along with the rankings compared to the* 25 *models tested in Michie et al.* (1994), *except for the vowel data, where rankings come from Friedman* (1997). *The k*-NN *column has corresponding results for the k-nearest-neighbour method, where k is selected by crossvalidation. The p-value is provided from McNemar's test under the null hypothesis that the error rates of the two approaches are equal.*

Dataset	n	Test method	p	Q	MPL	k-NN	p value
Australian Credit	690	10-fold	14	2	0·133 (2)	0·149 (9)	0·0218
Diabetes	768	12-fold	8	2	0·239 (5)	0·267 (13)	0·1981
Heart*	270	9-fold	13	2	0·414 (5)	0·418 (5)	0·0053
German Credit*	1000	10-fold	24	2	0·548 (3)	0·591 (5)	0·0000
Vowel	528	462	10	11	0·493 (1)	0·658 (25)	0·0000

n, number of observations in each dataset; p, number of variables; Q, number of classes.
* Datasets have associated misclassification costs.

(Michie et al., 1994, Ch. 7). In particular, as highlighted in Table 1, we use N-fold crossvalidation (Mitchell, 1996, Ch. 5) to assess performance. Crossvalidation provides an unbiased assessment of prediction error and is widely used in nonlinear model assessment; see Hastie et al. (2001, Ch. 7) for a more detailed discussion.

Among the Statlog datasets, the Australian and German Credit datasets are both concerned with the problem of allocating new customers to a good or bad risk category according to application form and demographic data. The German dataset has misclassification costs with Type II errors, of classifying a bad debtor as good, being 5 times more costly than the reverse. The Australian dataset does not quote a misclassification cost, for reasons of commercial confidentiality.

The Heart dataset is concerned with predicting whether or not a patient suffers from heart disease based on a set of continuous variables, including blood testing and electrocardiogram derived measurements. The cost of misclassifying a sufferer of heart disease as healthy is five times as costly as the reverse misallocation. The Diabetes dataset is concerned with predicting whether or not patients are likely to test positive for diabetes according to a World Health Organization criterion, using eight variables, observed on a group of adult females of Pima Indian heritage.

The Vowel dataset is included as a multinomial example. The problem consists of classifying the 11 steady state vowel sounds in British English, using a collection of variables obtained in a speaker normalisation study; see a 1989 University of Cambridge Ph.D. thesis by D. H. Deterding. The training data refer to vowel sounds recorded by one group of speakers, while the test data refers to sounds recorded by a different group.

The performance of our method on these benchmark datasets is also given in Table 1, along with the performance of the conventional k-nearest-neighbour method with k chosen by a 10-fold crossvalidation study on each training sample. In all of the examples the maximum pseudolikelihood method with multiple predictors outperformed the conventional approach. Many of the models studied in Statlog are not able to incorporate costs into the predictions and hence the ranking of the model on the Heart and German Credit data may appear artificially high. However, real-world classification tasks often have unequal costs associated with false positive and false negative predictions and we see the ability to deal with this as an advantage of our approach.

The results in Table 1 hint at potential improvements in estimation by maximum pseudolikelihood over standard crossvalidation. To test this formally we applied McNemar's test (Ripley, 1996, Ch. 2). The p-values for the test are shown in Table 1; in all but one of the tests the difference in error rate is significant at the 2·5% level.

Insight into the data can be provided by examining the variables selected by the model. We re-ran the model using all of the data and in Table 2 we list the variables selected for each dataset alongside their corresponding values of $\hat{\beta}$ in brackets. In Table 2, $x^{(i)}$ indicates that the ith linear predictor was included and $z^{(j)}$ indicates that the j-nearest-neighbour autocovariate was included. For the synthetic dataset of § 3·2, the interaction parameter $\hat{\beta} = 5{\cdot}72$ indicates that there is a strong association between the classes of neighbouring points. The high values for $\hat{\beta}$ and $\hat{k}$ suggest that the classes are reasonably well separated in this example. In the Diabetes dataset, the second and sixth predictor variables were included as linear effects. These variables relate to blood glucose level and age respectively and it seems highly plausible that these could have a global positive linear effect on the probability of exhibiting diabetes. The single autocovariate $z^{(42)}$ in the Diabetes example has a fairly high value of k, which suggests that some smooth nonlinear effects remain after the linear terms are included. In the German Credit data we see that just two nonlinear terms are included relating to $k = 144$ and $k = 7$. This suggests that nonlinear effects are operating at different scales within the data; that is, the $k = 7$ term is capturing very local effects while the $k = 144$ is modelling smoother structure.

Table 2. *Variables used in Benchmark datasets. The term $x^{(i)}$ denotes that the ith predictor was included as a linear term and $z^{(j)}$ denotes that the j-nearest-neighbour autocovariate was included*

Dataset	Variables selected ($\hat{\beta}$)
Synthetic	$z^{(66)}$ (5·72)
Australian Credit	$x^{(9)}$ (−0·64), $x^{(14)}$ (1·95), $z^{(66)}$ (4·40)
Diabetes	$x^{(6)}$ (0·40), $x^{(2)}$ (0·48), $z^{(42)}$ (2·18)
Heart*	$z^{(12)}$ (3·09)
German Credit*	$z^{(144)}$ (5·14), $z^{(7)}$ (0·62)

* Datasets have associated misclassification costs.

3·4. *Osteoporosis example*

The final example in this section relates to part of the European Prospective Osteoporosis Study. The area of investigation is the automatic classification of osteoporosis sufferers using statistical pattern recognition models on X-ray data.

A preliminary detection method for osteoporosis usually involves an expert examining an X-ray for signs of vertebral fractures, often taken as evidence for osteoporosis. However, this has been criticised as being too subjective, and a number of methods of defining vertebral deformities based on measurements of the X-ray have been proposed (Minne et al., 1998; Melton et al., 1993; McCloskey et al., 1993). Such methods are widely used in clinical trials and epidemiological studies, although there is no agreement about which method performs best (Black et al., 1995).

Details of the procedure used to perform and measure the radiographs are given in Lunt et al. (1997). The raw X-ray image is digitised and the anterior, mid and posterior heights of 13 vertebrae are measured from the digitised image using a mouse-caliper system on a back-lit digitising board. Technically, the 13 vertebrae are the fourth to the twelfth thoracic vertebrae, T4 to T12, and the first four lumbar vertebrae, L1 to L4. The dataset consists of 667 records and the observed proportion of patients with vertebral deformity is 0·269.

We ran our model on the data and the variables selected, with the associated $\hat{\beta}$, were $x^{(6)}$ (0·41), $x^{(12)}$ (0·58), $x^{(18)}$ (0·40), $x^{(30)}$ (0·48), $z^{(43)}$ (11·07) and $z^{(10)}$ (3·59), in the notation of Table 2. The four linear terms selected relate solely to the posterior height recorded on the sequence T5, T7, T9 and L1. The fact that the estimated regression parameters associated with these terms are all positive suggests that the posterior height of the even vertebrates may be an important global feature in this classification task, though it should be noted that all variables enter into the autocovariates. The presence of the autocovariates suggests there is clearly some nonlinear structure remaining after the linear terms. The autocovariates selected record both small- and large-scale interaction between classes in feature space. These findings are the basis of on-going work.

We now turn to the question of model validity. As mentioned in § 1, one advantage of the generalised linear model framework is that it allows for standard diagnostic measures to be introduced to assess the validity of the modelling assumptions. One such statistic is the Pearson chi-squared residual

$$\chi^2 = \sum_{i=1}^{n} r_i^2,$$

where

$$r_i = \frac{y_i - \hat{y}_i}{\surd\{\hat{y}_i(1-\hat{y}_i)\}},$$

in which $\hat{y}_i$ is the model's prediction for y_i, is the Pearson residual for y_i. The contribution of the ith observation to the chi-squared residual can be measured by the decrease in the value of χ^2 brought about by the removal of the observation from the dataset, namely

$$\Delta\chi_i^2 = \frac{r_i^2}{1-h_{ii}},$$

where h_{ii} is the ith diagonal element of the 'hat' matrix

$$H = V^{\frac{1}{2}} S_{\hat{\gamma}} (S_{\hat{\gamma}}' V S_{\hat{\gamma}})^{-1} S_{\hat{\gamma}}' V^{\frac{1}{2}},$$

where V is a diagonal matrix with elements $v_{ii} = \hat{y}_i(1-\hat{y}_i)$ and, as before, $S_{\hat{\gamma}}$ denotes the selected predictors. The hat matrix is derived from the weighted least squares approximation to logistic regression; see Pregibon (1981) and Hosmer & Lemeshow (2000, Ch. 5).

Large values of $\Delta\chi_i^2$ highlight influential outliers that are poorly fitted by the model. In Fig. 3 we plot these effects against the prediction $\hat{y}_i$ for our model of the osteoporisis data; $y_i = 0$ denotes a fractured vertebra. Two highly influential outliers are observation number 239, which was predicted as a fracture when in fact it was a non-fracture, and observation 324, which was predicted as a non-fracture when it was a fracture. On further analysis we found that observation 239 had exceptionally low mid-height measures on T8 and T9,

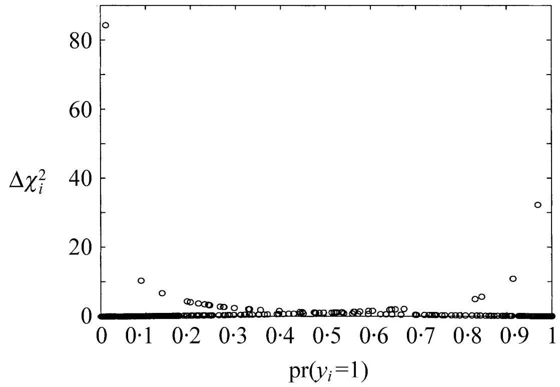

Fig. 3. Osteoporosis example. Plot of contribution to the Pearson chi-squared residual for each data point.

which is usually indicative of a fractured vertebra. These measurements on 239 were much lower than any other non-fracture sample. We have been unable to determine if this observation has been mislabelled by the radiologist. Observation 324 is an example of a 'compression fracture' where, for one or more vertebrae, the anterior, mid and posterior measurements are all low because of a partial collapse of the vertebrae. This makes 324 an outlier in our sample of fracture examples and its nearest neighbours turn out to be non-fractures. Though this class is rare, if necessary we should distinguish it as a separate category and form a multinomial model with labels non-fracture, compression fracture and non-crush fracture.

4. Discussion

Our analysis suggests there is often advantage, in the sense of predictive performance, in using more than one k term. This insight seems to have remained largely undiscovered in the k-nearest-neighbour literature. Furthermore, our method tends to favour higher values of k than are typically considered in conventional k-nearest-neighbour methods. The higher values of k will tend to produce smoother probability models for the data. One of the referees made the interesting observation that editing methods also tend to produce smoother boundaries. Editing methods reduce the number of observations used in constructing predictions in order to improve classification performance (Ripley, 1996, Ch. 6, p. 198). Many editing methods exist and which ones perform best in which situations is an on-going area of research. Dasarathy et al. (2000) provides a comparison of several editing procedures. We are currently investigating how to use maximum pseudolikelihood to infer the editing model.

While our examples suggest that our method can be more accurate than the standard k-nearest-neighbour method, this clearly will not be the case in all situations. In practical applications our recommendation would be to evaluate both models using some unbiased measure, such as 10-fold crossvalidation, and select the one that performs best. However, there are other issues surrounding the choice of procedure. The fitting of the multiple model by maximum pseudolikelihood is slower than the standard k-nearest-neighbour especially for the multinomial response data where a Broyden–Fletcher–Goldfarb–Shanno quasi-Newton optimiser is needed. In some applications the extra computation may prove prohibitive especially for on-line classification where new data are arriving over time. Of further note is that the k-nearest-neighbour algorithm is guaranteed to converge to the Bayes risk (Ripley, 1996, Ch. 6) whereas the maximum pseudolikelihood method's performance does not share this guarantee. This result may be worthy of consideration for very large datasets in low dimensions.

The osteoporosis example presented in § 3·4 suggests that the method has utility for elucidating structure in data, in addition to its utility for classification. Current work is concerned with investigation of weighted distance metrics, rather than the Euclidean norm reported here, in order to allow each predictor to have a different influence in the resulting autocovariates.

Acknowledgement

We are grateful to Mark Lunt for introducing us to the osteoporosis problem, and to anonymous referees for interesting and useful comments.

References

Aha, W. D. (1977). *Lazy Learning*. Dordrecht: Kluwer.

Akaike, H. (1977). On entropy minimization principle. In *Application of Statistics*, Ed. P. R. Krishnaiah, pp. 27–42. Amsterdam: North Holland.

Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Ann. Inst. Statist. Math.* **30A**, 9–14.

Besag, J. E. (1972). Nearest-neighbour systems and the auto-logistic model for binary data. *J. R. Statist. Soc.* B **34**, 75–83.

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems (with Discussion). *J. R. Statist. Soc.* B **36**, 192–236.

Besag, J. E. (1986). On the statistical analysis of dirty pictures (with Discussion). *J. R. Statist. Soc.* B **48**, 259–302.

Black, D. M., Palermo, L., Nevitt, M. C., Genant, H. K., Epstein, R., San Valentin, R. & Cummins, S. R. (1995). Comparison of methods for defining prevalent vertebral deformities: The study of osteoporotic fractures. *J. Bone Miner. Res.* **10**, 890–902.

Blake, C. L. & Merz, C. J. (1998). UCI repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.

Dasarathy, B. V. (1991). *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. Los Alamitos, CA: IEEE Computer Society Press.

Dasarathy, B. V., Sanchez, J. S. & Townsend, S. (2000). Nearest neighbour editing and condensing tools-synergy exploitation. *Pat. Anal. Applic.* 3, 19–30.

Draper, N. R. & Smith H. (1981). *Applied Regression Analysis*. New York: Wiley.

Friedman, J. H. (1997). On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining Know. Discov.* **1**, 55–77.

Geman, D. & Geman, S. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE. Trans. Pat. Anal. Mach. Intel.* **6**, 721–41.

Hansen, M. H. & Yu, B. (2001). Model selection and the principle of minimum description length. *J. Am. Statist. Assoc.* **96**, 746–74.

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning*. New York: Springer-Verlag.

Holmes, C. C. & Adams, N. M. (2002). A probabilistic nearest-neighbour method for statistical pattern recognition. *J. R. Statist. Soc.* B **64**, 295–306.

Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression*. New York: Wiley.

Jensen, J. L. & Kunsch, H. R. (1994). On asymptotic normality of pseudo likelihood estimates for pairwise interaction processes. *Ann. Appl. Prob.* **46**, 445–61.

Lunt, M., Felsenberg, D., Reeve, J., Benevolenskaya, L., Cannata, J., Dequeker, J., Dodenhof, C., Falch, J. A., Masaryk, P., Pols, H., Poor, G., Reid, D., Scheidt-Nave, C., Weber, K., Varlow, J., Kanis, J., O'Neill, T. & Silman, A. J. (1997). Bone density variation and its effects on risk of vertebral deformity in men and women studied in 13 European centres: the EVOS study. *J. Bone Miner. Res.* **12**, 1883–94.

Mase, S. (1995). Consistency of the maximum pseudo-likelihood estimator of the continuous state-space Gibbsian processes. *Ann. Appl. Prob.* **5**, 603–12.

McCloskey, E., Spector, T. D., Eyres, K. S., Fern, E. D., O'Rourke, N., Wasikaran, S. & Kanis, J. A. (1993). The assessment of vertebral deformity: a method for use in population studies and clinical trials. *Osteoporosis Int*. **3**, 138–47.

McCullagh, P. & Nelder, J. A. (1989). *Generalised Linear Models*, 2nd ed. London: Chapman and Hall.

Melton, L. J., Lane, A. W., Cooper, C., Eastell, R., O'Fallon, W. M. & Biggs, R. L. (1993). Prevalence and incidence of vertebral deformities. *Osteoporosis Int*. **3**, 113–9.

Michie, D., Spiegelhalter, D. J. & Taylor, C. C. (1994). *Machine Learning, Neural and Statistical Classification*. New York: Ellis Horwood.

Minnie, H. W., Leidig, G., Wüster, C., Siromachkostov, L., Baldauf, G., Bickel, R., Sauer, P., Lojen, M. & Ziegler, R. (1988). A newly developed spine deformity index (SDI) to quantitate vertebral crush fractures in patients with osteoporosis. *Bone Miner*. **3**, 335–49.

Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw Hill.

Pregibon, D. (1981). Logistic regression diagnostics. *Ann. Statist*. **9**, 705–24.

Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vettering, W. T. (1990). *Numerical Recipes in C*. Cambridge University Press.

Ripley, B. D. (1994). Neural networks and related methods for classification (with Discussion). *J. R. Statist. Soc*. B **56**, 409–56.

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

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist*. **6**, 461–4.

Strauss, D. J. (1977). Clustering on coloured lattices. *J. Appl. Prob*. **14**, 135–43.

Venables, W. N. & Ripley, B. D. (1999). *Modern Applied Statistics with S-PLUS*, **1**: *Data Analysis*. New York: Springer Verlag.

[*Received March* 2001. *Revised December* 2001]