Measuring Distance from a Training Data Set

Ilmari Juutilainen and Juha Röning

Computer Engineering Laboratory
PO BOX 4500
90014 University of Oulu, Finland
(e-mail: ilmari.juutilainen@ee.oulu.fi, juha.roning@ee.oulu.fi)

Abstract. In this paper, a new method is proposed for measuring the distance between a training data set and a single, new observation. The novel distance measure reflects the expected squared prediction error, when the prediction is based on the k nearest neighbours of the training data set. The simulation shows that the distance measure correlates well with the true expected squared prediction error in practice. The distance measure can be applied, for example, to assessing the uncertainty of prediction.

Keywords: Distance measure, Model uncertainty,
Distance weighted k-nearest-neighbour, Novelty detection.

1 Introduction

In some applications, such as in evaluation of the reliability of prediction at a query point, it is interesting to measure the information given by the training data set about a new observation via the current prediction model. In this work, we propose a novel measure for the distance between a single observation and a data set. The distance measure reflects the expected uncertainty of the new observation being predicted based on the data set. The distance measure is a linear function of the approximated expected squared prediction error, when the new observation is predicted with the distance weighted k-nearest-neighbour method.

There has been much discussion about measuring the distance between two observations. We refer to a review paper [Wettschereck *et al.*, 1997] that discusses the different methods. Often, Euclidean distance or Manhattan distance is used, and the problem lies in the weighting or scaling of the variables. The input variables that have a large effect on the response should have large weights in the distance measure. Global distance measures use constant weights, unlike local distance measures. Some distance measures take the correlations between the explanatory variables into account.

The measurement of the distance between a set of observations and a single observation has also been widely discussed. Different distance measures have been applied in clustering and in prototype methods. In these applications, the aim in defining the distance has been to assign the observation to the nearest cluster or prototype. Examples of the different methods include the average pairwise distance, the Mahalanobis distance and the Euclidian distance to the cluster centroid. We refer to [Kaufman and Rousseeuw, 1990] for these methods. However, these methods have been planned to measure the distance between a cluster and a single observation and not the distance between a data set and a single observation.

Novelty detection aims to find abnormal observations from a data set. Abnormal observations can indicate that the modelled system is in an abnormal state, which needs to be reported. In classification, detection of novel observations is needed to identify new classes and observations that cannot be classified reliably. Novelty detection can be used to differentiate novel information from existing information when only the novel information needs to be shown to the learners. For novelty detection methods, we refer to the review [Markou and Singh, 2003].

The usual approach in novelty detection is to measure somehow the similarity with the training data and to use some threshold to interpret the observations as novel. The most common method is to model the joint density function of input variables to judge the observations with low density as novel [Markou and Singh, 2003]. Our approach differs in that we do not construct any distribution model for the inputs. Our distance measure tries to measure the uncertainty about the expected response value at a new query point, which is quite a novel approach to the problem. The standard errors of predictions measure the uncertainty with variance, but we take also bias into account.

[Angiulli and Pittuzi, 2005] suggested a method for detecting outliers in a data set. They calculated the sum of the Euclidean distances to the k nearest neighbours for measuring the distance, which approach is quite similar to our proposal. [Mahamud and Hebert, 2003] discussed the optimal distance measures in k-nearest-neighbour prediction, and we constructed our distance measure using a similar optimality principle.

2 Distance between two single observations

Let $x_{(j)}$ refer to the jth explanatory variable and x_{ij} denote the ith observation of $x_{(j)}$, y_i denote the ith observation of the response and T denote the training data set consisting of N observations $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$. Let (x_0, y_0) be a new test data observation and $d_i = d(x_0, x_i)$ measure the distance between x_0 and $x_i \in T$. We assume that the response depends on the inputs via a regression function $f(\cdot)$, and that the additive error term has a constant variance

$$y_i = f(x_i) + \varepsilon_i,\ E(\varepsilon_i) = 0,\ \text{Var}\ (\varepsilon_i) = \sigma^2. \tag{1}$$

[Mahamud and Hebert, 2003] discussed the optimal distance measures in nearest-neighbour classification. The optimal distance measure in 1-nearest-neighbour prediction minimises the expected loss function $E_{y_0,x_0,T} L(y_0, y')$, where y' is the measured response at x', which is the nearest neighbour of x_0

using the distance measure d. The distance measure $d(x_0, x_i) = EL(y_0, y_i)$ is optimal, because the nearest neighbour $x' = \arg\min_{x_i} EL(y_0, y_i)$ minimises the expected loss $L(y_0, y')$ $\forall x_0$ $\forall T$ [Mahamud and Hebert, 2003]. The same reasoning holds for k-nearest-neighbour prediction. All order-preserving transformations of the expected loss function are optimal, because the nearest neighbours remain the same. We use the expected squared error loss $EL(\mu_0, y_i) = E(\mu_0 - y_i)^2 = E(y_0 - y_i)^2 - \sigma^2$ related to the true expectation $\mu_0 = E(y|x_0) = f(x_0)$ without losing optimality.

The optimal distance measure cannot be used directly because the conditional expectation of the response is not known, and the true expected loss cannot be solved. The optimal distance measure is not monotonic, which implies an interpretational disadvantage: The nearest neighbours may lie far away from the query point on the scale of explanatory variables. To eliminate this problems, we must be content with a coarse approximation of the expected loss: We use the sum of the expectations of squared differences in the true regression function, when one input variable at a time is set to the measured values x_0 and x_i, and other input variables are drawn randomly,

$$E(\mu_0 - y_i)^2 = \sigma^2 + [f(x_0) - f(x_i)]^2 \approx \sigma^2 + \sum_{j=1}^{p} E_x\big\{f[w_i^{(j)}(x)] - f[w_0^{(j)}(x)]\big\}^2. \tag{2}$$

In the formula, x is a randomly drawn input observation, $w_0^{(j)}(x)$ is otherwise identical with x but the jth element is altered $w_{0j}^{(j)} = x_{0j}$, and $w_i^{(j)}(x)$ is otherwise identical with x but the jth element $w_{ij}^{(j)} = x_{ij}$.

In the case of continuous input variables, we further approximate the squared differences in y with squared differences in the input variable values $E_x\big\{f[w_i^{(j)}(x)] - f[w_0^{(j)}(x)]\big\}^2 \approx \alpha_j(x_{0j} - x_{ij})^2$. [Mahamud and Hebert, 2003] proposed to estimate the α-coefficients by fitting a regression model to a data set of pairs of training data instances using the response $L(y_i, y_j)$. The advantage of their direct method is that the regression function need not be estimated. We propose a different method. Let our prediction model be $\widehat{y} = \widehat{f}(x) = \widehat{f}(x_{(1)}, x_{(2)}, \ldots, x_{(p)})$, and let $\widehat{\sigma}^2$ be the corresponding error variance estimate. Let now $x_c \in T$ denote a training data observation lying near x_0, and let $\widehat{f}'(x_c) = \Big(\frac{\partial \widehat{f}(x)}{\partial x_{(1)}}, \frac{\partial \widehat{f}(x)}{\partial x_{(2)}}, \ldots, \frac{\partial \widehat{f}(x)}{\partial x_{(p)}}\Big)_{(x=x_c)}$ denote the gradient of the fitted response surface at point x_c. Motivated by the first-order Taylor approximation around x_c, we suggest that $\alpha_1, \ldots, \alpha_p$ are defined as the average squared partial derivative over the training data set

$$\alpha_j = \frac{1}{N} \sum_{i=1}^{N} \Big(\frac{\partial \widehat{f}(x)}{\partial x_{(j)}}{}_{(x=x_i)}\Big)^2. \tag{3}$$

For large N, it is enough to calculate the average over a sample. The regression function can be fitted using any learning method, for example, neural networks or additive models. The partial derivatives of the fitted response surface can be approximated numerically with $\frac{\partial \widehat{f}(x)}{\partial x_{(j)}}\,_{(x=x_i)} = \frac{\widehat{f}(x_i) - \widehat{f}(x_i + o_j)}{|o_j|}$, where o_j is a vector of zeros elsewhere but a small constant at the jth element.

When $x_{(j)}$ is a categorical variable with class levels $\gamma_{j1}, \gamma_{j2}, \ldots, \gamma_{jq_j}$, we can estimate the expected squared difference $E_x\big\{f[w_i^{(j)}(x)] - f[w_0^{(j)}(x)]\big\}^2$ between each two class levels γ_{jl} and γ_{jm} using the fitted prediction model with $\frac{1}{|J|}\sum_{j \in J}\Big(\widehat{f}(x_i) - \widehat{f}(w_i^{(j)})\Big)^2$. The input vectors $w_i^{(j)}$ are otherwise identical to x_i but the jth element is altered: $w_{ij}^{(j)} = \gamma_{jm}$, if $x_{ij} = \gamma_{jl}$, and $w_{ij}^{(j)} = \gamma_{jl}$, if $x_{ij} = \gamma_{jm}$. The squared differences in the prediction are averaged over the index set $J = \Big\{i | \widehat{f}(x_i), \widehat{f}(w_i^{(j)})$ are reliable and $(x_{ij} = \gamma_{jl}$ or $x_{ij} = \gamma_{jm})\Big\}$. For binary variables we can notate

$$\alpha_j = \frac{1}{|J|}\sum_{j \in J}\Big(\widehat{f}(x_i) - \widehat{f}(w_i^{(j)})\Big)^2. \tag{4}$$

We propose to use an approximate optimal distance measure that is the approximated expected squared error loss

$$d(x_0, x_i) = \alpha_0 + \sum_{j=1}^{p} \alpha_j (x_{0j} - x_{ij})^2. \tag{5}$$

The coefficient α_0 is the error variance estimate $\widehat{\sigma}^2$. The notation (Eq. 5) is applicable for continuous and binary variables, but categorical variables can be taken into account as explained previously.

3 Distance between a single observation and a data set

We suggest that the distance of a single observation from a set of k observations, S_k, is measured on the basis of the expected squared error when the single observation is predicted based on S_k. This can be seen as the generalisation of the pairwise optimal distance measure. The true expected loss at x_0 is not known and has to be approximated. We predict $\mu_0 = E(y_0)$ with a distance-weighted linear combination of the y values measured in S_k, which results in measurement of the distance with the harmonic sum of pairwise distances.

Let $S_k = (x_1, x_2, \ldots, x_k)$ with the distances $d_1, d_2, \ldots, d_k$ from x_0, and let each distance be known $d_i = E(\mu_0 - y_i)^2$. Let us now estimate μ_0 with a weighted linear combination $\widehat{y}_0 = \omega_1 y_1 + \omega_2 y_2 + \cdots + \omega_k y_k$. Under the symmetry assumption $E(\mu_0 - y_i) = 0$, the minimum variance unbiased estimator gives weights proportional to the inverses of the variances and sums the weights to unity $\omega_j = \frac{1}{d_j}\Big/\sum_{i=1}^{k}\frac{1}{d_i}$. We use this distance-weighted estimator

$$\widehat{y}_0 = \Big(\sum_{i=1}^{k}\frac{1}{d_i}\,y_j\Big)\Big/\sum_{i=1}^{k}\frac{1}{d_i} \tag{6}$$

to predict y_0 based on S_k. We keep the estimator (Eq. 6) as a natural basis for the interpretation of our distance measure because the approach does not make any assumption about the form of the regression function. The expected squared loss of our estimator is the harmonic sum of pairwise distances d_i plus a bias term

$$\begin{aligned}
E(\widehat{y}_0-\mu_0)^2 &= E\Big(\sum_{i=1}^{k}(\omega_i y_i)-\mu_0\Big)^2 = E\Big(\sum_{i=1}^{k}\omega_i(y_i-\mu_0)\Big)^2\\
&= \sum_{i=1}^{k}\omega_i^2 E(y_i-\mu_0)^2 + 2\sum_{j=1}^{k}\sum_{i\neq j}E(y_i-\mu_0)E(y_j-\mu_0)\omega_i\omega_j\\
&= \Big(\frac{1}{\sum_{i=1}^{k}\frac{1}{d_i}}\Big)^2\Big[\sum_{i=1}^{k}d_i/d_i^2 + 2\sum\sum_{i\neq j}\frac{E(y_i-\mu_0)E(y_j-\mu_0)}{d_i d_j}\Big]\\
&= \frac{1}{\sum_{i=1}^{k}\frac{1}{d_i}} + \Big(\frac{1}{\sum_{i=1}^{k}\frac{1}{d_i}}\Big)^2 2\sum\sum_{i\neq j}\frac{E(y_i-\mu_0)E(y_j-\mu_0)}{d_i d_j}. \qquad (7)
\end{aligned}$$

We take the expectations (Eq. 7, 8, 9 and 10) over x_i, also, which means that x_i are assumed to be random points satisfying the condition $d(x_0, x_i) = d_i$. If the assumption $E_{Y,x_i|d_i}(\mu_0 - y_i) = 0\,\forall i$ holds, the bias term would be zero, and the expected squared error would be the harmonic sum of the pairwise distances. However, that is not a realistic assumption. Some query points x_0 may lie in a 'symmetric' position where the assumption holds. But some query points may lie at the bottom of a valley or on the top of a hill, where the expectation $E(y_i - \mu_0)$ is negative for all possible neighbours x_i.

Let us now us examine the bottom of a valley scenario in more detail. Because $E\ (y_i - \mu_0)^2 = d_i$ and Var $y_i = \sigma^2$, it holds that $E\ (y_i - \mu_0) = \sqrt{d_i - \sigma^2}$. Let $\bar{d}$ denote the average inverse distance $\frac{1}{k}\sum_{i=1}^{k} 1/d_i$. We can derive an upper bound for the bias term

$$\begin{aligned}
2\sum\sum_{i\neq j}\frac{E\ (y_i-\mu_0)E\ (y_j-\mu_0)}{d_i d_j} &= 2\sum\sum_{i\neq j}\frac{\sqrt{d_i-\sigma^2}\sqrt{d_j-\sigma^2}}{d_i d_j}\\
= \sum_{i=1}^{k}\Big[\frac{\sqrt{d_i-\sigma^2}}{d_i}\sum_{j\neq i}\frac{\sqrt{d_j-\sigma^2}}{d_j}\Big] &<= \sum_{i=1}^{k}\bar{d}\sqrt{\frac{1}{\bar{d}}-\sigma^2}\sum_{j\neq i}\bar{d}\sqrt{\frac{1}{\bar{d}}-\sigma^2}\\
= k(k-1)\bar{d}^2\Big(\frac{1}{\bar{d}}-\sigma^2\Big) &= (k-1)\sum_{i=1}^{k}\frac{1}{d_i} - \sigma^2\frac{k-1}{k}\Big[\sum_{i=1}^{k}\frac{1}{d_i}\Big]^2 (8)
\end{aligned}$$

The result follows from the Jensen inequality and concavity of the function $q(x) = x\sqrt{1/x - \sigma^2}$. Equality is achieved if the distances to all the neighbours are constant $1/d_i = \bar{d}\,\forall i$. When all the k neighbours are roughly equally distant, and x_0 lies at the bottom of a valley or on the top of a hill, the bias term can be approximated as a linear function of the harmonic sum $1/\sum_{i=1}^k \frac{1}{d_i}$ and k

$$\left(\frac{1}{\sum_{i=1}^k \frac{1}{d_i}}\right)^2 2\sum_{j=1}^k \sum_{i\neq j} \frac{E(y_i-\mu_0)E(y_j-\mu_0)}{d_i d_j} \approx \frac{k-1}{\sum_{i=1}^k \frac{1}{d_i}} - \sigma^2 \frac{k-1}{k}. \tag{9}$$

Simulation studies showed that this approximation holds well in practice: In all of the simulated data sets, the correlation between the harmonic sum and the bias term was over 0.99 when pairwise distances depending only on x were used (Eq. 5) and over 0.94 when the true distances $d_i = E(\mu_0 - y_i)^2$ were used and $k \leq 50$.

At all query points x_0, the true bias can be expressed in relation to the maximum bias with $E_{Y,x_i|x_0,d_i}(y_i - \mu_0) = c(x_0)\sqrt{d_i - \alpha_0}$. When x_0 lies in a symmetric position, $c(x_0) = 0$, at the bottom of the valley $c(x_0) = 1$, and on the top of the hill $c(x_0) = -1$. When we assume that $c(x_0)$ does not depend on the distance d_i and denote $E_{x_0}c(x_0)^2 = \delta^2$, the expected squared prediction error can be approximated with

$$\begin{aligned} E_{Y,x|d_1\ldots d_k}(\mu_0 - \widehat{y}_0)^2 &= E_{x_0}\left(\frac{1}{\sum_{i=1}^k \frac{1}{d_i}}\right)^2 2\sum\sum_{i\neq j} \frac{c(x_0)^2\sqrt{d_i-\sigma^2}\sqrt{d_j-\sigma^2}}{d_i d_j} \\ &+\frac{1}{\sum_{i=1}^k \frac{1}{d_i}} \approx \frac{1}{\sum_{i=1}^k \frac{1}{d_i}} + \delta^2(k-1)\frac{1}{\sum_{i=1}^k \frac{1}{d_i}} - \sigma^2\delta^2\frac{k-1}{k}. \end{aligned} \tag{10}$$

This is a linear transformation of the harmonic sum when k is kept fixed. Thus, the harmonic sum $1/\sum_{i=1}^k \frac{1}{d_i}$ can be used as a measure of the uncertainty about μ_0 when $y_1, \ldots, y_k$ and $d_1, \ldots, d_k$ are given. On the basis of simulated data, the approximation seems to work well in practice: The correlations between the approximation and the true expected loss were about 0.9.

We propose that the distance between a single observation x_0 and a set of observations $S_k = (x_1, x_2, \ldots, x_k)$ is measured with the harmonic sum of pairwise distances $d_i = d(x_i, x_0)$. When the pairwise distances correspond to the expected squared error $d_i \approx E(\mu_0 - y_i)^2$, our distance measure $d(x_0, S_k)$ approximates an increasing linear function of the expected squared prediction error $E(\mu_0 - \widehat{y}_0(S_k))^2$. We suggest that the distance between x_0 and S_k is measured with

$$\begin{aligned} d(x_0, S_k) &= \frac{1}{\sum_{i=1}^k \frac{1}{d_i}} \\ d_i &= \sum_{j=1}^p \alpha_0 + \alpha_j(x_{0j} - x_{ij})^2. \end{aligned} \tag{11}$$

4 Measuring the distance to a training data set

Our method could be used directly to measure the distance between a single observation and the training data set by letting $S_k = T$. However, when the training data set is large, it makes more sense to use only the k nearest observations. In the k-nearest-neighbour method, typically 5 to 100 neighbours are used to obtain the most accurate prediction. Thus, the observations lying far away from x_0 should not have an effect on the distance measure, because they do not affect the prediction. Let $d^{(k)}$ be the kth smallest distance $d(x_0, x_i)$. Our suggestion for the distance between the training data set and a single observation is

$$d(x_0, T) = d(x_0, S_k),\ S_k = \left\{x_i \in T \mid d(x_0, x_i) \leq d^{(k)}\right\}, \tag{12}$$

Our distance measure is problem-dependent. If we have the same inputs and several responses, the distance measure has to be defined separately for each response. The distance measure adapts itself to the regression function. The variables that do not affect the response do not affect the distance, either. The distance measure is invariant for linear transformations and approximately invariant for order-preserving transformations of the inputs. The distance measure also has a reasonable interpretation as the approximate measure of the expected loss function, which is an informative and novel way to measure the uncertainty about a new observation. The distance measure uses the squared error loss function, but can also be used for non-gaussian responses. If μ_0 were estimated with the unweighted k-nearest-neighbour method, the result would be the sum of single distances, just as proposed in [Angiulli and Pittuzi, 2005].

After the distance measure has been initialised by defining the α-coefficients, the major computational task is to find the k nearest training data observations. The computation of a single distance to the training data set requires about $N(p+2)+k^2$ operations. Initialision of the distance measure consists of fitting a prediction model and defining the α-coefficient for each explanatory variable, which is not a computational problem even in large data sets.

When prediction using some novel input values is needed, there rises the question of whether the model gives a reliable prediction or not. If the query point has enough training data instances nearby, the prediction can be kept reliable. If the query point is far away from the training data instances, the model will give a poor prediction with a high probability. The distance between the query point and the training data set gives information about the uncertainty of the prediction, see the example in Figure 1. The prediction accuracy of the model for validation data observations distant from the training data gives some information about the interpolation ability of model. In the example shown in Figure 1, the smoothed prediction accuracies of a linear regression model, a quadratic regression model and an additive spline model are plotted as functions of distance from the training data set.

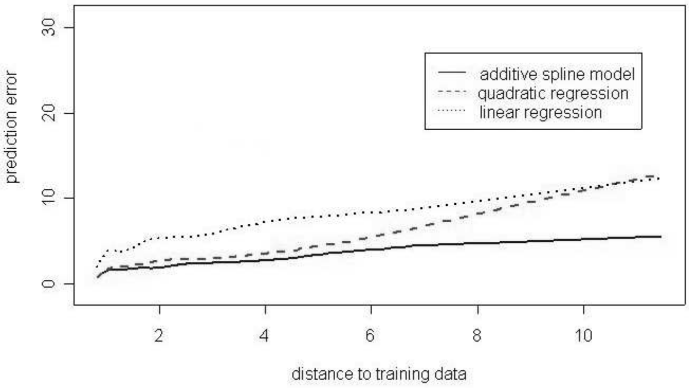

Fig. 1. Average prediction error (rMSE) in a simulated data set.

5 Performance in simulated data sets

The proposed distance measure reflects the expected squared error loss function $d(x_0, S_k) \approx c_1 + c_2 E(\mu_0 - \widehat{y}_0(S_k))^2$. We evaluated the correlation between the distance measure and the true expected loss using simulated data sets. The simulated data sets tried to represent a range of data sets which could arise from an industrial process of production. The observations occurred in clusters of different sizes, and the input variables were correlated. The true expected response was defined as a sum of 24 random effects of the form $\nu|b_0 + \beta_1 x_{(1)} + \beta_2 x_{(2)} + \cdots + \beta_{16} x_{(16)}|^b s$, where $b = e^{0.5 z_b}$, $z_b \sim N(0,1)$, making the typical effect rather linear, and the signum s turns the effect monotone with a probability of 0.7. Only 1, 2, 3 or 4 of β_i differs from 0, which means that interactions are restricted to the 4th order. The observed response was normally distributed around the true expected response. One simulated data set consisted of 10 000 observations and 16 input variables.

We simulated 20 data sets. We split all simulated data sets randomly into a learning data set and a validation data set. Out of the 2000 observations in the validation data, we calculated the distances to the learning data set using the proposed method. For each data set, we fitted an additive model with univariate thin plate regression splines as basis functions to define the α- coefficients of our distance measure. We defined the true pairwise distance as the true expectation $E(\mu_0 - y_i)^2$ and the true distance to the training data as the true expected squared prediction error

$$E\left(\mu_0 - \frac{\sum_{i=1}^{k} y_i / d(x_i, x_0)}{\sum_{i=1}^{k} 1/d(x_i, x_0)}\right)^2. \tag{13}$$

We examined the accuracy of our pairwise distance measure in approximating the expected loss $E(\mu_0 - y_i)^2 \approx \alpha_0 + \sum_{j=1}^{p} \alpha_j (x_{0j} - x_{ij})^2 = d(x_0, x_i)$ based on the correlations between the pairwise distance measure $d(x_i, x_j)$

and its theoretical reciprocal $(\mu_i - \mu_j)^2 + \sigma^2$. In the simulated data sets, the correlation varied between 0.19 and 0.81, the average correlation being 0.47. When neighbourhood size $k = 30$ was used, the correlation between the distance measure (Eq. 11) and the true expected squared error $E_Y(\mu_0 - \widehat{y}_0)^2$ ($\widehat{y}_0$ is defined in Eq. 6) varied between 0.41 and 0.66, the average correlation being 0.52. Thus, our distance measure $d(x_0, T)$ reflects relatively well its theoretical reciprocal, the expected squared error loss when x_0 is predicted based on T using distance-weighted k-nearest-neighbour. The deviation between the true expected squared error and our distance measure is mainly the consequence of the difficulty in approximating pairwise expected loss based only on x. If the true pairwise expected losses were known, the approximation would work much better: The correlation between the true expected loss $E(y_0 - \widehat{y}_0)^2$, $\widehat{y}_0 = \sum_{i=1}^{k} y_i / E(y_i - \mu_0)^2$ and the harmonic sum of the true pairwise distances $1/\sum_{i=1}^{k} 1/E(y_i - \mu_0)^2$ was typically about 0.93 and over 0.83 in all simulated data sets for $k \leq 200$. The size of the neighbourhood had a relatively small effect on the results, and all alternatives between $k = 5$ and $k = 500$ gave satisfactory correlations, and the best size of the neighbourhood varied greatly between the simulation runs. We suggest the use of $k = 30$, because that seemed to work best, and no larger neighbourhood was needed for k-nearest-neighbour prediction. Also, it seems intuitively reasonable that the distance to the training data can be defined based on the distances to the 30 nearest neighbours.

We compared our distance measure to the sum of pairwise distances of [Angiulli and Pittuzi, 2005]. Using $k = 30$, our distance measure was slightly better in 90 % of the simulation runs, and the average difference in correlation was 0.035. We also examined the effect of the method on defining α-coefficients for the distance measure. The average correlation between the pairwise distances based on a fitted additive model and on the true response surface was 0.92. The correlations between distances calculated based on two different learning methods were around 0.95, which means that the model fitting had only a small effect on the results. The method of [Mahamud and Hebert, 2003] for specifying α-coefficients gave poor results: The average pairwise correlation was only 0.30.

In the simulated data sets, the distance measure reflected the uncertainty about a new observation pretty well. We applied the distance measure to real industrial process data. We used a training data set having 90 000 observations, 26 continuous input variables and 6 binary input variables without any computational problems. In the test data set containing 60 000 observations, the average prediction error increased along with the distance from the training data (Figure 2). The correlations between the measured loss and the distance measure were between 0.25 and 0.5, depending on the response variable and the prediction model.

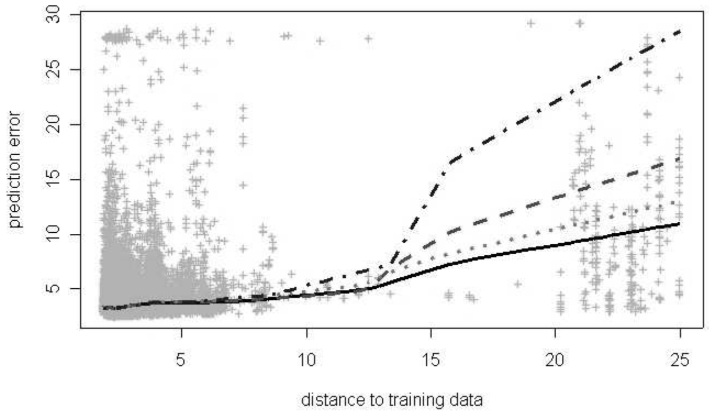

Fig. 2. Prediction error plotted against distance from the training data set in the real data set. The lines are the smoothed medians for four different prediction models.

6 Conclusion

We proposed a novel distance measure for the distance between a data set and a single observation. The distance measure can be interpreted to reflect the expected squared error loss when the single observation is predicted based on the data set using distance-weighted k-nearest-neighbour. Measurement of the distance from a data set has many potential applications, such as evaluation of the uncertainty of prediction and discovery of outliers.

References

[Angiulli and Pittuzi, 2005]F. Angiulli and C. Pittuzi. Outlier mining in large high-dimensional data sets. *IEEE Transactions on Knowledge and Data Engineering*, pages 203–215, 2005.

[Kaufman and Rousseeuw, 1990]L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley, New York, 1990.

[Mahamud and Hebert, 2003]S. Mahamud and M. Hebert. Minimum risk distance measure for object recognition. In *Proceedings of the ninth IEEE International Conference on Computer Vision (ICCV)*, pages 242–248, 2003.

[Markou and Singh, 2003]M. Markou and S. Singh. Novelty detection: a review. *Signal Processing*, pages 2481–2521, 2003.

[Wettschereck *et al.*, 1997]D. Wettschereck, D.W. Aha, and T. Mohri. Review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *Artificial Intelligence Review*, pages 273–314, 1997.