HETEROGENEOUS DISTANCE MEASURES AND NEAREST-NEIGHBOR CLASSIFICATION IN AN ECOLOGICAL SETTING

MATTHEW S. SPENCER, SAMANTHA C. BATES PRINS, AND MARGARET S. BECKOM

ABSTRACT. Deriving a suitable heterogeneous distance measure that mixes continuous and categorical attributes is a difficult problem with a variety of applications. We developed the Scaled Heterogeneous Euclidean Overlap Metric (SHEOM) and we adapted the Interpolated Value Difference Metric (IVDM) from Wilson and Martinez [2, 3]. Our adaptation of the IVDM utilizes output classes given by a continuous response variable. We applied both distance measures in a nearest-neighbor classification for an ecological assessment. Both of the distance measures we applied were shown to be improvements on the simpler Heterogeneous Euclidean Overlap Metric when used in this ecological assessment setting.

1. INTRODUCTION

A distance measure gives a numerical description of the similarity between two objects using some number of attributes measured on those objects. If all attributes are continuous then the well-known Euclidean distance function can be used as a distance measure. Let

$$\mathbf{x_i} = (x_{i,1}, x_{i,2}, \ldots, x_{i,R}) \text{ and } \mathbf{x_j} = (x_{j,1}, x_{j,2}, \ldots, x_{j,R})$$

be vectors of continuous data whose components are the measured values of R attributes or predictors on objects i and j, respectively. The Euclidean distance function defines the similarity between $\mathbf{x_i}$ and $\mathbf{x_j}$ as

$$E(\mathbf{x_i}, \mathbf{x_j}) = \sqrt{\sum_{r=1}^{R} (x_{i,r} - x_{j,r})^2}.$$

A computationally simplified version of the Euclidean distance function is the city-block or Manhattan distance function. This distance function is defined as

$$M(\mathbf{x_i}, \mathbf{x_j}) = \sum_{r=1}^{R} |x_{i,r} - x_{j,r}|.$$

In applications a subset of the attributes can overpower the remaining attributes in the distance measure. For example, in the ecological assessment setting described by Bates Prins and Smith [1], the similarity between stream i and stream j might be measured using the continuous attributes of latitude and catchment area. In this setting, an attribute is a predictor so we use that nomenclature when referring to this application. Suppose $x_{i,1}$ is the value of latitude at the ith stream with plausible values $x_{i,1} \in [0, 90]$ and $x_{j,2}$ is the value of catchment area at the jth stream with plausible values $x_{i,2} \in [1, 10]$. Then the values of $|x_{i,1} - x_{j,1}|$ will likely be larger than $|x_{i,2} - x_{j,2}|$ simply due to the greater range in latitude values. Thus, latitude will dominate the Euclidean or Manhattan distance measures. To deal with this problem in a general setting, distance functions are often normalized by dividing the contribution of each attribute by the range of possible or measured values of that predictor [2, 3]. This should force the contribution of each attribute to the distance measure to be in the interval $[0, 1]$.

A much more difficult problem arises when distance is measured using both continuous and categorical attributes. Finding a distance measure that incorporates both types of attributes, known as a heterogeneous distance measure, is our focus. The Heterogeneous Euclidean Overlap Metric (HEOM) from Wilson and Martinez [2, 3] is one example of a heterogeneous distance measure. Suppose we wish to find distances between some subset of n objects and that for each object we have measured the values of R predictors. Let $J = \{1, 2, \ldots, n\}$ be an index set for each of the n objects. For each $i, j \in J$, the HEOM defines the distance between the ith object and the jth object as

$$\text{HEOM}(\mathbf{x_i}, \mathbf{x_j}) = \sum_{r=1}^{R} d_r(x_{i,r}, x_{j,r}),$$

where

$$d_r(x_{i,r}, x_{j,r}) = \begin{cases} \frac{|x_{i,r} - x_{j,r}|}{\text{range}_r} & \text{if } r \text{ indexes a continuous attribute} \\ \delta_{i,j} & \text{if } r \text{ indexes a categorical attribute} \end{cases} \quad (1)$$

and $\delta_{i,j} = 1$ if $x_{i,r} \neq x_{j,r}$ and $\delta_{i,j} = 0$ if $x_{i,r} = x_{j,r}$. Here $d_r(x_{i,r}, x_{j,r})$ can be thought of as the contribution of the rth attribute to the overall distance and $\text{range}_r = \max_{j \in J}\{x_{j,r}\} - \min_{j \in J}\{x_{j,r}\}$. Notice that a continuous attribute's contribution to the HEOM distance is bounded above by 1.

Problems can arise when using the HEOM distance function with mixed-type attributes. Referring to Equation (1), if the values of a continuous attribute on objects i and j are equal, i.e., $x_{i,r} = x_{j,r}$, then $d_r(x_{i,r}, x_{j,r}) = 0$. Importantly, the same distance value results if r indexes a categorical

attribute. However, if objects i and j differ with respect to a continuous attribute and a categorical attribute then these two attributes contribute equal amounts to the HEOM distance only if the absolute difference in the continuous attribute is the largest difference observed in the data. In general, for a continuous attribute $|x_{i,r} - x_{j,r}| = \text{range}_r$ occurs for only a small number of pairs of objects. Thus, for a continuous attribute, non-matching values will almost always contribute less to the distance measure than non-matching values in a categorical attribute.

Both problems mentioned involve the issue of scaling; the first involves the scale of the attributes themselves, the second involves the scale of the attribute's contributions to the HEOM distance, and both arose in the ecological assessment setting described by Bates Prins and Smith [1]. In this setting, there are observed values of continuous and categorical predictors as well as six biological metrics, such as Ephemeroptera Richness, that are considered the responses. Each of the predictors and metrics/responses were observed at n reference streams located in the mid-Atlantic region. These reference streams are known to be minimally impaired from environmental stress. We let $J = \{1, 2, \ldots, n\}$ index the n reference streams. The data also consists of measurements of the continuous and categorical predictors and metrics/responses at some number of test streams that we wish to classify as either impaired from environmental stress or minimally impaired. In order to classify a particular test stream as impaired or minimally impaired, Bates Prins and Smith [1] found the k nearest-neighbor reference streams to a test stream using the HEOM distance measure and a subset of the available predictors. The value used for range_r was the range of predictor r measured on the n reference streams only. The k nearest-neighbor reference streams were then used to determine a scaled metric/response value, $\bar{y}_i^*$, at the test stream using the average and the standard deviation of the metric/response values observed at the k nearest-neighbor reference streams in Equation (9) of Bates Prins and Smith [1]. $\bar{y}_i^*$ was subsequently used to classify the test stream as either impaired or not by comparison to a t-distribution (see Bates Prins and Smith [1] for details).

For each metric/response, Bates Prins and Smith [1] chose the value of k and the particular subset of predictors to use in the nearest-neighbor method using a leave-one-out approach that minimized the mean squared error (MSE) of prediction. First, an initial value of k and an initial subset of the predictors are chosen. Then treating the first of the n reference streams as a test stream, the HEOM distance based on the chosen subset of predictors is calculated between stream 1 and each of the remaining $n - 1$ reference streams and the k closest neighbors of stream 1 are used to calculate the scaled metric value at stream 1. This process is repeated for each of the reference streams and the MSE is calculated using the n

resulting $\bar{y}_i^*$ values. This entire process is repeated for various values of k and different combinations of predictors. The value of k and the subset of predictors that result in the minimum value of MSE are then used to find the scaled metric/response value at a true test stream.

The presence of the first scaling issue in this application has already been illustrated using the predictors latitude and catchment area. The second scaling issue is also present because although the leave-one-out approach allows the $d_r(x_{i,r}, x_{j,r})$ for a continuous predictor to no longer be bounded above by 1, $d_r(x_{i,r}, x_{j,r}) > 1$ only when the predictor value at the left-out stream is equal to the minimum or maximum occurring in all n reference streams. That is, non-matching values in a continuous predictor will still almost always contribute less to the distance measure than non-matching values in a categorical predictor. The relative high contribution of categorical predictors to the HEOM distance resulted in these predictors seldom being chosen amongst the subset of predictors that minimized MSE.

This paper will offer two improved approaches to the Bates Prins and Smith [1] application of heterogeneous distance measures. First, we present a simple modification to the HEOM, which we refer to as the Scaled Heterogeneous Euclidean Overlap Metric (SHEOM) that addresses the scaling of the attributes directly. We then present an extension of the Interpolated Value Difference Metric (IVDM) of Wilson and Martinez [2, 3] for use with a continuous response. We conclude with a comparison of the benefits and complications presented by each method.

2. The Scaled Heterogeneous Euclidean Overlap Metric

The first and computationally simpler distance function we present is the Scaled Heterogeneous Euclidean Overlap Metric (SHEOM). The SHEOM distance function addresses the first scaling issue by scaling the continuous attribute values to have mean 0 and variance 1 prior to calculating the Manhattan distance. With $\mathbf{x_i}$ and $\mathbf{x_j}$ defined as in the introduction, the SHEOM first scales each continuous attribute using

$$x^*_{p,r} = \frac{x_{p,r} - \bar{x}_r}{s_r}, \tag{2}$$

where p indexes the object and $\bar{x}_r$ and s_r are the mean and standard deviation, respectively of predictor r measured on all but the pth object.

We then define the contribution of the rth predictor to the SHEOM distance measure using

$$d^*_r(x_{i,r}, x_{j,r}) = \begin{cases} \left|x^*_{i,r} - x^*_{j,r}\right| & \text{if } r \text{ indexes a continuous predictor} \\ \delta_{i,j} & \text{if } r \text{ indexes a categorical predictor,} \end{cases} \tag{3}$$

where again $\delta_{i,j}$ is as defined for the HEOM. Notice that $\left|x^*_{i,r} - x^*_{j,r}\right| = \frac{|x_{i,r}-x_{j,r}|}{s_r}$ so that SHEOM is equivalent to the HEOM with scaling performed by the standard deviation rather than the range. We then define the SHEOM as

$$\text{SHEOM}(\mathbf{x_i}, \mathbf{x_j}) = \sum_{r=1}^{R} d^*_r(x_{i,r}, x_{j,r}).$$

3. The Interpolated Value Difference Metric (IVDM)

The Interpolated Value Difference Metric (IVDM) was originally developed by Wilson and Martinez [2, 3] for an instance based learning algorithm. Central to the IVDM calculation are the use of input classes, the value of a categorical or discretized attribute, and the use of output classes, the value of a categorical response variable. The IVDM measures distance in terms of differences in the relative frequency of each input class and output class. We now extend the IVDM method to the setting in which the response variable is continuous.

Let $J = \{1, 2, \ldots, n\}$ be an index set for each of the n objects. These n objects take on the role of the training set referred to by Wilson and Martinez [2, 3]. Let i indicate a test object that we wish to find the nearest-neighbors of. p will either indicate the test object or a member of the set of n training objects. Recall that $\mathbf{x_p} = (x_{p,1}, x_{p,2}, \ldots, x_{p,R})$ is a vector of values of the R attributes observed on p. We will assume without loss of generality that the first R_G of these attributes are categorical and the remaining $R - R_G$ are continuous. y_p will indicate the value of the response on the pth object. In the ecological assessment setting i indicates a test stream that we wish to classify as impaired or not, p indicates either the test stream or a reference stream (in the case of the leave-one-out approach), J indexes the n reference streams (or $n - 1$ in the case of the leave-one-out approach), and y_p is the value of the metric at the pth stream.

We now give the steps involved in determining the IVDM distances.

(1) Discretize the response variable into C output classes. This is accomplished by dividing the range of measurements of the response based on the training objects into C subintervals of equal width w. These subintervals are labeled $1, 2, \ldots, C$ with interval 1 representing the lowest response values. The subinterval to which a particular y_p belongs becomes the discretized value of y_p and its output class. Thus, the output class associated with the response

observed at test object p is defined as

$$\text{discretize}(y_p) = \begin{cases} C & \text{if } y_p \geq \max\limits_{j \in J}\{y_j\} \\ 1 & \text{if } y_p < \min\limits_{j \in J}\{y_j\} \\ \lfloor (y_p - \min_{j \in J}\{y_j\}) \diagup w \rfloor + 1 & \text{otherwise,} \end{cases}$$

where

$$w = \frac{1}{C}\left|\max_{j \in J}\{y_j\} - \min_{j \in J}\{y_j\}\right|.$$

We note that $\text{discretize}(y_p) \in \{1, 2, \ldots, C\}$, but some subset of the output classes may not be observed in any particular collection of objects. Also, the response on the pth object has output class C if it is larger than the response over all training objects regardless of how much larger it is. A similar idea follows when y_p is smaller than all other responses.

(2) For each $x_{p,r}$ where $r \in \{1, 2, \ldots, R_G\}$ and for each $c = 1, 2, \ldots, C$, we define $N_{x_{p,r},c}$ as the number of times attribute r has value equal to $x_{p,r}$ on the training objects that have discretized response value c. Define $N_{x_{p,r}}$ to be the total number of times attribute r has value equal to $x_{p,r}$ on all training objects. The conditional probability that of output class c, given that the attribute value on object p is $x_{p,r}$ is

$$P_{x_{p,r},c} = \frac{N_{x_{p,r},c}}{N_{x_{p,r}}}.$$

If $x_{p,r}$ never appears within the training objects we set $P_{x_{p,r},c} = 0$. $P_{x_{p,r},c}$ is the relative frequency that will be used in the IVDM calculation.

(3) Continuous attributes are discretized using the same approach used for discretizing the continuous response. For each $r \in \{R_G + 1, \ldots, R\}$ divide the range of values for attribute r observed on the training objects into S_r subintervals of equal width w_r. The subinterval to which a particular $x_{p,r}$ belongs becomes the discretized value of $x_{p,r}$ and its input class. Thus, the input class associated with the value of the rth continuous attribute observed at test object p is defined as

$$u_{p,r} = \text{discretize}(x_{p,r}) = \begin{cases} S_r & \text{if } x_{p,r} \geq \max\limits_{j \in J}\{x_{j,r}\} \\ 1 & \text{if } x_{p,r} < \min\limits_{j \in J}\{x_{j,r}\} \\ \lfloor (x_{p,r} - \min_{j \in J}\{x_{j,r}\}) \diagup w_r \rfloor + 1 & \text{otherwise,} \end{cases}$$

where

$$w_r = \frac{1}{S_r}\left|\max_{j\in J}\{y_j\} - \min_{j\in J}\{y_j\}\right|.$$

As with the response, $u_{p,r} \in \{1, 2, \ldots, S_r\}$, although for any particular collection of objects some subset of these values may not occur.

(4) We now determine the relative frequencies that will be assigned to the midpoints of each input class for each continuous attribute. For each $r \in \{R_G + 1, \ldots, R\}, u_{p,r} \in \{1, \ldots, S_r\}$, and each $c = 1, 2, \ldots, C$, the conditional probability of output class c given that the input class for the rth attribute on object p is $u_{p,r}$ is given by

$$P_{u_{p,r},c} = \frac{N_{u_{p,r},c}}{N_{u_{p,r}}}.$$

As with the categorical predictors, $N_{u_{p,r},c}$ is the number of times training objects with output class c have an attribute r value equal to $u_{p,r}$ and $N_{u_{p,r}}$ is the total number of times $u_{p,r}$ appears within the training objects for the rth attribute.

(5) We now find the midpoint of each input class for continuous attributes. For each $r \in \{R_G + 1, \ldots, R\}$ and $u_{p,r} \in \{1, \ldots, S_r\}$,

$$\text{mid}(u_{p,r}) = \min_{j\in J}\{x_{j,r}\} + w_r \times (u_{p,r} - 0.5).$$

(6) We now assume a linear relationship between the relative frequencies determined in step (4) placed at each midpoint determined in step (5) to obtain the interpolated probability associated with each continuous attribute $r \in \{R_G + 1, \ldots, R\}$, output class $c = 1, 2, \ldots, C$, and object p using

$$P(x_{p,r})_c = \begin{cases} P_{u_{p,r},c} + \left(\frac{x_{p,r} - \text{mid}(u_{p,r})}{\text{mid}(u_{p,r}+1) - \text{mid}(u_{p,r})}\right) \times (P_{u_{p,r}+1,c} - P_{u_{p,r},c}) \\ \qquad \text{if } x_{p,r} \geq \text{mid}(u_{p,r}) \\ P_{u_{p,r}-1,c} + \left(\frac{x_{p,r} - \text{mid}(u_{p,r}-1)}{\text{mid}(u_{p,r}) - \text{mid}(u_{p,r}-1)}\right) \times (P_{u_{p,r},c} - P_{u_{p,r}-1,c}) \\ \qquad \text{if } x_{p,r} < \text{mid}(u_{p,r}), \end{cases}$$

where if $x_{p,r} < \min_{j\in J}\{\text{mid}(u_{j,r})\}$ we set $P_{u_{p,r}-1,c} = 0$, and if $x_{p,r} \geq \max_{j\in J}\{\text{mid}(u_{j,r})\}$ we set $P_{u_{j,r}+1,c} = 0$. If $x_{p,r} > \max_{j\in J}\{\text{mid}(u_{j,r})\} + w_r$

or $x_{p,r} < \min_{j \in J}\{\text{mid}(u_{j,r})\} - w_r$, that is, if the test object is more than $0.5w_r$ units larger (or smaller) with respect to the rth attribute than any object in the training set, we set $P(x_{p,r})_c = 0$.

(7) Finally, the distance between test object p and training object j as defined by the IVDM as

$$IVDM(\mathbf{x_p}, \mathbf{x_j}) = \sum_{r=1}^{R} ivdm(x_{p,r}, x_{j,r}),$$

where $ivdm(x_{p,r}, x_{j,r})$ is defined using the relative frequencies in step (2) and interpolated probabilities in step (6) as

$$ivdm(x_{p,r}, x_{j,r}) = \begin{cases} \sum_{c=1}^{C} \left| P_{x_{p,r},c} - P_{x_{j,r},c} \right|^2 & \text{if } r \text{ indexes a categorical predictor} \\ \sum_{c=1}^{C} \left| P(x_{p,r})_c - P(x_{j,r})_c \right|^2 & \text{if } r \text{ indexes a continuous predictor.} \end{cases}$$

Figure 1 shows an example of how the interpolated probability values would be calculated for a particular continuous attribute and response, $C = 4$ output classes, $S_r = 6$ input classes, and a test object that has attribute value 3.34 and response value of approximately 2. Discretization divides a scatter plot of attribute versus response into $S_r \times C = 24$ rectangular regions (top figure). The test object lies in the 4th vertical strip of $C = 4$ regions so it gets assigned input class, $u_{p,2} = 4$. Horizontal strips (consisting of $S_r = 6$ regions) indicate the output class so discretize$(y_p) = 1$ for this test object. The relative frequency, $P_{u_{p,r},c} = P_{4,1}$, associated with the region our test object lies in is calculated from step (4) using the number of objects in that region divided by the total number of objects in input class 4. Step (4) is repeated for all regions including those with the same output class as our test object resulting in $P_{1,1}, \ldots, P_{6,1}$. These 6 relative frequencies are associated with the midpoints calculated in step (5) and are plotted in the bottom figure. Because our test object lies above the midpoint of input class 4, the line through the relative frequencies for input classes 4 and 5 is used to find its interpolated probability, $P(x_{p,2})_1$.

The bottom plot in Figure 1 more clearly shows the effect of step (6) in that the interpolated probability is set to zero for attribute values that are more than $0.5w_r$ units larger (or smaller) than any object in the training set.

As noted by Wilson and Martinez [2, 3], there is subjectivity in the choice of S_r and C. We explain our choices in the results section.

FIGURE 1. The interpolated probability of the object that has $x_{p,2} = 3.34$ for output class $c = 1$ based on a continuous attribute ('Predictor 2') and response ('Response'). The top figure illustrates the discretization of the attribute and response using $C = 4$ and $S_2 = 6$ resulting in output class discretize$(y_p) = 1$ and input class $u_{p,2} = 4$ for the pth object. The midpoint associated with this object's input class is mid$(u_{p,2}) = 3.25$. The bottom figure shows (with points) the relative frequencies from step (4) placed at the midpoints of each of the six input classes obtained in step (5). The solid line indicates the interpolated probability, $P(x_{p,2})_1$ associated with the pth object.

4. Data

The data used in Bates Prins and Smith [1] included $n = 87$ reference streams. Six different biological metrics (responses) were used including the proportional abundance of tolerant taxa of aquatic macroinvertebrates (TOLRPIND), tolerant taxa richness (TOLRRICH), the proportional abundance of the three most abundant taxa (DOM3PIND), Ephemeroptera richness (EPHERICH), Plecoptera richness (PLECRICH), and total taxa richness (TOTLRICH). The continuous predictors included log-transformed catchment area (AREA), latitude (LAT), longitude (LON), and total rapid bioassessment protocol habitat score (RBP). The categorical predictor used was the Level III Ecoregion (ECO) with six levels. See Table 2 in Bates Prins and Smith [1] for numeric summaries of the data.

All analysis was done in R Version 2.4.1.

5. Results

We compared the performance of the HEOM, SHEOM, and IVDM distance functions in finding the k nearest-neighbor reference streams in the leave-one-out approach of Bates Prins and Smith [1]. We used two evaluation criteria: the mean squared error of prediction (MSE) as described briefly in the Introduction section and in detail in Bates Prins and Smith [1] and the percentage of reference streams correctly classified as not impaired. The latter was determined by treating each reference stream in turn as a test stream of unknown classification, using the distance function to find the stream's neighbors and the scaled metric/response value at that stream. Bates Prins and Smith [1] describe in detail how the scaled metric value is used to classify the stream as impaired or not impaired.

Wilson and Martinez [2, 3] suggested a heuristic approach to deciding S_r and C and we followed their convention. We initially set S_r and C to be the same as the number of levels in the categorical predictor, ECO, namely 6. To determine how sensitive the results might be to this choice we also used all combinations of $S_r, C \in \{6, 12, 18, 24\}$. The MSE of prediction was used to determine the best combination for each metric/response. Table 1 gives the values of S_r and C chosen for each metric; notice that the MSE chose values equal to or close to our initial choice.

Table 1 gives the optimal MSE values obtained by each distance function using each of the six metrics/responses and Figure 2 gives graphs of MSE as a function of k, the number of nearest neighbors for the chosen subset of predictors. Notice that the SHEOM and IVDM methods performed similarly in terms of MSE compared to the HEOM distance function for TOLRPIND and TOLRRICH and they performed better than HEOM for

EPHERICH. The IVDM obtained a noticeably lower MSE than the other two methods for PLECRICH, but obtained a noticeably higher MSE than the other two methods with DOM3PIND and TOTLRICH. The SHEOM method outperformed the HEOM for EPHERICH and TOTLRICH.

Table 1 also lists the percentage of reference streams correctly classified as "not impaired" in the RATE column. RATE should be 100% since all 87 reference steams are by design minimally impaired streams. The classification rates were similar across all three distance functions.

Finally, notice from Table 1 that the categorical predictor ECO was selected only by the IVDM and SHEOM distance functions and when ECO was selected by IVDM and SHEOM, the resulting MSE was lower than the MSE for the HEOM method in all but one case. This suggests that the IVDM and SHEOM more readily mix both continuous and categorical predictors.

6. Discussion

This paper offered two approaches to determining distances between mixed attributes and applied these to an ecological application described in Bates Prins and Smith [1]. First, we presented a simple modification to the HEOM, which we refer to as the Scaled Heterogeneous Euclidean Overlap Metric (SHEOM) that addresses the scaling of the attributes directly. We then presented an extension of the Interpolated Value Difference Metric (IVDM) of Wilson and Martinez [2, 3] for use with a continuous response. Both methods were shown to more readily mix continuous and categorical predictors in our chosen application than did the HEOM but had similar MSE and classification rates as the HEOM.

The IVDM and SHEOM can be used to make predictions of similar or better accuracy to the HEOM while more easily incorporating both continuous and categorical attributes. We believe the improvement of SHEOM over HEOM is due to the effect of scaling the continuous attributes to be mean 0 and variance 1 prior to calculating the distances. Recall that the contributions of attribute r to the total distance between objects i and j are $d_r(x_{i,r}, x_{j,r})$ and $d_r^*(x_{i,r}, x_{j,r})$ for the HEOM and SHEOM, respectively. The boxplots in Figure 3 represent the values of $d_r(x_{i,r}, x_{j,r})$ and $d_r^*(x_{i,r}, x_{j,r})$ for each of the continuous predictors considered in our application. The horizontal lines represent the contribution of a non-match in a categorical predictor (namely 1) and a match in a categorical predictor (namely 0) for both distance measures. With HEOM the contribution of a categorical predictor was almost always greater than the contribution of a continuous predictor. This meant that large distances were observed whenever the categorical predictor was selected within the nearest-neighbor algorithm. This was not the case with the SHEOM since the contributions

TABLE 1. The minimum mean squared error (MSE) obtained by each method together with the k, subset of predictors, S_r and C associated with that MSE value. Also included is RATE, the percentage of times a reference stream was correctly classified as unimpaired by a particular method. Higher values of RATE and lower values of MSE indicate the better distance function. The lowest values of MSE achieved for each metric are in bold. Predictors included in the nearest-neighbor distance calculations are listed in the fifth column.

EPHERICH					
Method	MSE	RATE	k	Predictors	S_r, C
IVDM	8.57	94	21	AREA,ECO	6 , 6
SHEOM	**8.43**	93	25	AREA,ECO	- , -
HEOM	9.08	95	15	AREA,LAT	- , -
PLECRICH					
Method	MSE	RATE	k	Predictors	S_r, C
IVDM	**3.74**	98	11	AREA,LAT	12 , 6
SHEOM	4.13	93	11	AREA,LAT	- , -
HEOM	4.16	93	14	AREA	- , -
TOTLRICH					
Method	MSE	RATE	k	Predictors	S_r, C
IVDM	134.03	92	20	AREA,LON	12 ,6
SHEOM	**122.77**	94	28	AREA,ECO	- , -
HEOM	127.70	95	21	AREA,LAT	- , -
TOLRRICH					
Method	MSE	RATE	k	Predictors	S_r, C
IVDM	2.49	97	25	AREA,LON	6 , 6
SHEOM	**2.45**	95	20	AREA	- , -
HEOM	**2.45**	95	20	AREA	- , -
TOLRPIND					
Method	MSE	RATE	k	Predictors	S_r, C
IVDM	0.00614	95	15	AREA	6 , 6
SHEOM	**0.00607**	96	21	AREA,ECO	- , -
HEOM	0.00609	94	20	AREA	- , -
DOM3PIND					
Method	MSE	RATE	k	Predictors	S_r, C
IVDM	0.0186	95	18	AREA,ECO	6 , 6
SHEOM	0.0163	96	18	AREA,LAT,LON,ECO	- , -
HEOM	**0.0161**	93	15	AREA,LAT	- , -

FIGURE 2. MSE of the six metrics at the 87 reference sites as a function of the number of nearest-neighbors for the optimal set of predictors. The lower the MSE, the better the prediction by a particular method. The solid line represents prediction with the IVDM method. The dashed and dotted lines represent prediction by the SHEOM and HEOM methods, respectively. The horizontal lines show the minimum MSE obtained by each method.

of the continuous predictors are centered just above 1 (indeed at approximately $\frac{2}{\sqrt{\pi}}$ as discussed below), allowing the contribution of a non-match in the categorial predictor to be greater than the contribution of a continuous predictor only about 50% of the time. This allows the SHEOM approach to result in use of the categorical predictor.

FIGURE 3. Boxplots of the individual contributions from each continuous predictor to the distance measure calculation for the HEOM (left) and SHEOM (right) that result from using the leave-one-out simulation described by Bates Prins and Smith [1]. Dotted horizontal lines indicate the values of $\delta_{i,j}$ in Equation (3) corresponding to a non-match and match in a categorical predictor. Circles represent outliers (values lying more than 1.5 times the interquartile range above the 75th percentile).

Assuming the scaled attribute values obtained in SHEOM using Equation (2) are independent and normally distributed, one expects

$$d_r^*(x_{i,r}, x_{j,r}) = \left| x_{i,r}^* - x_{j,r}^* \right|$$

in Equation (3) to have a mean of $\frac{2}{\sqrt{\pi}}$. As a result of this we investigated letting the contribution of a non-match in a categorical attribute in Equation (3) to be $\delta_{i,j} = \frac{2}{\sqrt{\pi}}$ rather than 1. No changes were observed in the relative performance of the SHEOM, HEOM, and IVDM methods in terms of mean squared error. The percentage of reference sites classified correctly by the SHEOM method either remained unchanged or improved to match or more closely match the HEOM method; although the methods differed by only a few percentage points to begin with. All in all, the changes were minimal so the output was not included.

The major benefit of the SHEOM distance measure is its computational simplicity relative to the IVDM. However, based on our investigation we recommend the IVDM method be used due to three factors: (1) by determining distances between relative frequencies, the contributions of all attributes are placed on the same scale regardless of their type; (2) the IVDM has been shown to be a good measure for a variety of data sets as discussed Wilson and Martinez (1997,2000) and the ecological application discussed by Bates Prins and Smith [1]; (3) multiple categorical attributes can be incorporated easily. In additional to its computational complexity relative to the HEOM or SHEOM methods, the IVDM has the disadvantage that the choice of the number of input classes and in our case also output classes is subjective. There appears to be little sensitivity to this based on our investigation. Somewhat surprisingly, Table 1 indicates that a smaller number of output classes was preferred for most metrics (we investigated 6–24 classes).

The Windowed Value Difference Metric (WVDM) of Wilson and Martinez [2] and the Density-Based Value Difference Metric (DBVDM) of Wojna [4] represent two potential improvements over the IVDM. The WVDM and DBVDM approaches calculate $P(x_{p,r})_c$ at every value of attribute r that occurs in the training objects rather than only at the midpoint of each input class; they represent a moving-window approach. The DBVDM differs from both IVDM and WVDM in that the number of observations within a window is fixed while the width of the window varies. The two methodology changes represented by WVDM and DBVDM, namely a moving window and a variable width window, would presumably improve the accuracy of the probability estimates in general and in particular for test objects that lie outside the range of the training objects. Both the WVDM and DBVDM are more computationally intensive than the IVDM.

Of particular interest to us was to find a distance measure which can both mix continuous and categorical attributes and calculate distances independent of the scale of those attributes when the response of interest was continuous. Although the IVDM distance measure we have discussed here

is not without problems, we have shown it to be a step in the direction of a heterogeneous distance measure suitable for a larger variety of applications.

7. Acknowledgement

The authors are grateful to the Jeffrey E. Tickle Scholarship for funding this work conducted during a summer research experience for undergraduates.

References

[1] S. C. Bates Prins and E. P. Smith, *Using biological metrics to score and evaluate sites: A nearest-neighbor reference condition approach*, Freshwater Biology, **50** (2006), 635–639.

[2] D. R. Wilson and T. R. Martinez, *Improved heterogeneous distance functions*, Journal of Artificial Intelligence Research, **6** (1997), 1–34.

[3] D. R. Wilson and T. R. Martinez, *An integrated instance-based learning algorithm*, Computer Intelligence, **16.1** (2000), 1–28.

[4] A. Wojna, *Analogy-based reasoning in classifier construction*, Lecture Notes in Computer Science, **3700** (2005), 277–374.

MSC2000: 62P12, 62H30

Department of Economics, University of North Carolina, Chapel Hill, NC 27599

E-mail address: spencems@email.unc.edu

Department of Mathematics and Statistics, James Madison University, Harrisonburg, VA 22807

E-mail address: prinssc@jmu.edu

U.S. Census Bureau

E-mail address: beckomms@jmu.edu