New Approaches in Applied Statistics
Anuška Ferligoj and Andrej Mrvar (Editors)
Metodološki zvezki, 16, Ljubljana: FDV, 2000

# Metric Approach to Property Prediction

Damijana Keržič[1]

**Abstract**

This paper surveys dissimilarity functions and weighting properties methods focusing on prediction problem. We discuss some transformations of dissimilarity functions and weighting methods. We can improve the quality of prediction by using two steps in weighting procedures, which is shown in the experiments where we achieved higher prediction accuracy on average using weights in dissimilarity functions than not using them.

## 1 Introduction

Let $\mathcal{E}$ be a set of *units* and $\mathcal{L}$ a *learning set*, also set of examples, $\mathcal{L} \subset \mathcal{E}$. On the learning set $\mathcal{L}$ the property $y$ is known, $y : \mathcal{L} \to L_0$, where $L$ is a set of all possible property values.

The property prediction problem can be expressed as follows:

> *For the unit not in the learning set, $Z \in \mathcal{E} \setminus \mathcal{L}$, predict the unknown value $y(Z)$.*

In the property prediction problem we assume that the property values change smoothly over the similar units. The *metric approach* to the prediction of $y(Z), Z \in \mathcal{E} \setminus \mathcal{L}$, is based on the known values for the units from $\mathcal{L}$ that are in the neighborhood of $Z$ (with respect to a selected dissimilarity). For this purpose we have to measure how similar two units are or how different they are in "appropriate" way.

Each of the considered units is represented by a vector of values of $n$ measured properties, $X = (x_1, x_2, \ldots x_n)$, where is $x_i$ also called *variable*. In real problems variables can be measured in different scales: nominal, ordinal, interval, ratio, or absolute. Looking for an appropriate dissimilarity, we have to take care about the type of scales of variables too - we will discuss this problem in Section 4.

[1] University of Ljubljana, IMFM/TCS, Ljubljana, Slovenia.

Thanks to the reviewers, whose comments helped me to greatly improve this article.

We shall use the following notation:

| | |
|---|---|
| $X_i = (x_1^{(i)}, x_2^{(i)}, \dots x_n^{(i)})$ | unit, $i = 1 \dots m,\ m = \lvert\mathcal{L}\rvert$ |
| $L_1, \dots L_n,$ | description properties |
| $L_j = (x_j^{(1)}, x_j^{(2)}, \dots x_j^{(m)}), j = 1 \dots n$ | measured values of property $L_j$ on $\mathcal{L}$ |
| $x_j^{(i)} = L_j(X_i)$ | description (explanatory) variable |
| $L_0,\ L_0 = (y_1, y_2, \dots y_m)$ | prediction property known on $\mathcal{L}$ |
| $y_i = y(X_i) = L_0(X_i)$ | prediction variable (response) |
| $d(X_i, X_j) = d_{ij}$ | dissimilarity, distance |
| $\mathbf{D},\ \mathbf{D}_{ij} = d_{ij}$ | dissimilarity matrix |

In this paper we survey some facts about dissimilarities in Section 2. Than we discuss how to manage with different scale type variables using in the descriptions of units. In Section 4 we review some weighting possibilities in dissimilarity function. We continue with the prediction methods based on dissimilarities and at the end we present two experimental results.

## 2 Dissimilarities

In this Section we introduce some definitions and general facts about dissimilarity measures. A *dissimilarity* on $\mathcal{E}$ is a mapping $d : \mathcal{E} \times \mathcal{E} \longrightarrow \mathbb{R}$ that has the properties:

1. vanishes on the diagonal: $d(X, X) = 0,\ \forall X \in \mathcal{E}$,
2. nonnegative: $d(X, Y) \geq 0,\ \forall X, Y \in \mathcal{E}$,
3. symmetric: $d(X, Y) = d(Y, X),\ \forall X, Y \in \mathcal{E}$.

   The ordered pair $(\mathcal{E}, d)$ is a *dissimilarity space*. A dissimilarity $d$ is said to be *semi-distance* if and only if

4. *triangle inequality*: $d(X, Z) + d(Z, Y) \geq d(X, Y),\ \forall X, Y, Z \in \mathcal{E}$

   holds. When $d$ is semi-distance $(\mathcal{E}, d)$ is called a *semi-metric space*. If $d$ is also

5. *definite*: $d(X, Y) = 0 \Longrightarrow X = Y$,

then $(\mathcal{E}, d)$ is a *metric* space and $d$ is a *distance*.

Definition of *Euclideanicity* of metric space $(\mathcal{E}, d)$ is the following: There exists $\varphi : \mathcal{E} \longrightarrow \mathbb{R}^n$, for some $n$, such that $d(X_i, X_j) = \delta(\varphi(X_i), \varphi(X_j)),\ \forall X_i, X_j \in \mathcal{E}$, where $\delta$ is the Euclidean distance.

Several examples of dissimilarity measures for different scales of variables can be found in any book of data analysis, see for example Anderberg (1973).

Using different transformations we can adopt a dissimilarity to additional requirement. We search for such transformations that will improve the quality of

dissimilarity in the sense that better prediction results will be obtained. When applying these transformations to a dissimilarity we wish that the properties of $d$ are preserved. We mention here some ways for constructing transformations that preserve dissimilarity, see for example Batagelj and Bren (1995):

**Proposition 1:** (dissimilarity $\rightarrow$ dissimilarity)
*Let $d$ be a dissimilarity on $\mathcal{E}$ and let a mapping $f : \mathbb{R}_0^+ \longrightarrow \mathbb{R}_0^+$ has the property $f(0) = 0$, then $\delta(X, Y) := f(d(X, Y))$ is also dissimilarity on $\mathcal{E}$.*

**Proposition 2:** (distance $\rightarrow$ distance)
*Let $d$ be a distance on $\mathcal{E}$ and let a mapping $f : \mathbb{R}_0^+ \longrightarrow \mathbb{R}_0^+$ has the properties:*

1. $f(x) = 0 \Longleftrightarrow x = 0$
2. $x \leq y \Longrightarrow f(x) \leq f(y)$
3. $f(x + y) \leq f(x) + f(y)$

*then $\delta(X, Y) := f(d(X, Y))$ is also a distance.*

The following transformations satisfy the last theorem, and so they preserve the metric property: $f(x) = \alpha x, \quad \alpha > 0$; $f(x) = \log(1 + x)$, $x \geq 0$; $f(x) = \frac{x}{1+x}$, $x \geq 0$; $f(x) = \min\{a, x\}, \quad a > 0$; $f(x) = x^\alpha$, $0 \leq \alpha \leq 1$. It is easy to verify that all concave transformations preserve metricity.

We can also combine dissimilarities into a new dissimilarity.

**Proposition 3:**
*Let $d_1$ and $d_2$ be dissimilarities. Then*

1. $d = \alpha d_1 + \beta d_2 \;\; ; \alpha + \beta > 0,\ \alpha, \beta \geq 0$
2. $d = \sqrt[p]{(d_1^p + d_2^p)}\ ;\ p \in \mathbb{R}, p \geq 1$
3. $d = \max\{d_1, d_2\}$

*is also a dissimilarity. If $d_1$ and $d_2$ are (semi-)distances then $d$ is a (semi-)distance too.*

Using last theorem we can show for example:
if $d_i$, $i = 1..k$ are distances, then $d = \sqrt[p]{\sum_{i=1}^{k} \alpha_i {d_i}^p}$ is a distance, $p \in \mathbb{N}$, $\alpha_i \geq 0$, $\sum \alpha_i > 0$.

Another important transformations of dissimilarities are transformations which improving dissimilarity into a distance or even into a Euclidean distance. The following methods of transformations are usually used (Joly and Le Calvé, 1994). Let $d$ be a dissimilarity, then:

- a positive additive constant: $d(X_i, X_j) + c$ is distance for some $c > 0,\ i \neq j$;
- a power transformation: $d^{\alpha}(X_i, X_j)$ is distance for some $\alpha,\ 0 \leq \alpha \leq 1$.

Let us cite three results. The proofs can be found for example in Gower and Legendre (1986) (Theorem 1), Joly and Le Calvé (1986) (Theorem 2, Theorem 3), and also in oldest references, see Cailliez and Pages (1976), Hoang (1978), Schoenberg (1937).

**Theorem 1**
*If $d$ is a dissimilarity then $d(X_1, X_2) + c,\ X_1 \neq X_2$ is metric if and only if $c \geq \max\{|d(X_1, X_2) + d(Z, X_2) - d(X_1, X_2)|;\ X_1, X_2, Z \in \mathcal{E}\}$.*

**Theorem 2**
*If $d$ is a dissimilarity there exists a unique nonnegative number $p \in \mathbb{R}^+$, such that $d^{\alpha}(X_1, X_2)$ is a distance for all positive $\alpha$ smaller or equal than $p$, $0 \leq \alpha \leq p$, and $d^{\alpha}(X_1, X_2)$ is not a distance for all $\alpha$ greater than $p$, $\alpha > p$.*

The threshold value $p$ from the last theorem is called a *metric index.*
Another important question is when transformation result is a Euclidean distance. Here is the answer the reader could find in Gower and Legendre 1986 or in Schoenberg (1938).

**Theorem 3**
*A dissimilarity $d$ is Euclidean if and only if the matrix $(\mathbf{I} - \mathbf{es}')\mathbf{\Delta}(\mathbf{I} - \mathbf{se}')$ is positive semi-definite where $\mathbf{I}$ is a unit matrix, $\mathbf{e} = [1, 1, \ldots 1]$ and $\mathbf{\Delta}_{ij} = -\frac{1}{2}d_{ij}^2$, $\mathbf{s}'\mathbf{e} = 1$.*

# 3 Dissimilarity and different scale types of variables

The prediction based on dissimilarity between units is very sensitive to the definition of the selected dissimilarity. In real world problems descriptions of units usually combine variables of different scale types. One possible strategy calculating the dissimilarities between units in such cases is conversion of variables from one type to another to achieve homogeneity, higher or lower level in the hierarchy of scale types. After such conversion we can use one of possible dissimilarity functions appropriate

for chosen scale type. Of course conversion of variables means also loss or ignoring of information or on the other hand adding additional artificial relations between units.

In the second approach we take the description variables just as they are. For each type of variables we choose appropriate dissimilarity and combine all of them into new dissimilarity. If $d_k$ is the dissimilarity for property $L_k$, the new dissimilarity between units $X_i$ and $X_j$ is $d(X_i, X_j) := z(d_1, d_2, \dots d_n)$, where $z$ is any function constructing by consecutive application of steps described in Propositions 1,2,3.

Here is an example taken from the literature named *Heterogeneous Euclidean-Overlap Metric* (HEOM) (Wilson and Martinez, 1997).
Let the units be represented by the vectors of nominal and continuous (ratio or interval scale type) variables, $X_i = (x_1^{(i)}, x_2^{(i)}, \dots x_n^{(i)})$, $X_j = (x_1^{(j)}, x_2^{(j)}, \dots x_n^{(j)})$.

The dissimilarity function $d_k$, that returns a dissimilarity between two values of the same property $L_k$, is defined as:

$$d_k(x_k^{(i)}, x_k^{(j)}) = \begin{cases} 1 & x_k^{(i)} \text{ or } x_k^{(j)} \text{ unknown} \\ d_n(x_k^{(i)}, x_k^{(j)}) & \text{nominal variables} \\ d_c(x_k^{(i)}, x_k^{(j)}) & \text{continuous variables} \end{cases} \tag{3.1}$$

depending on scale of variable. For nominal type of variables the discrete distance is used:

$$d_n(x_k^{(i)}, x_k^{(j)}) = \begin{cases} 0 & x_k^{(i)} = x_k^{(j)} \\ 1 & \text{otherwise} \end{cases} \tag{3.2}$$

and for continuous variables normalized Euclidean distance is used:

$$d_c(x_k^{(i)}, x_k^{(j)}) = \frac{|x_k^{(i)} - x_k^{(j)}|}{|\max_{L_k} - \min_{L_k}|} \tag{3.3}$$

where $\max_{L_k}$ is the maximal value and $\min_{L_k}$ is the minimum value of the property $L_k$ in the learning set $\mathcal{L}$, $\max_{L_j} = \max\{x_j^{(i)}; X_i \in \mathcal{L}\}$. The dissimilarity between two units $X_i,\ X_j \in \mathcal{E}$ is calculated:

$$d_{HEOM}(X_i, X_j) = \sqrt{\sum_{k=1}^{n} d_k^2(x_k^{(i)}, x_k^{(j)})}. \tag{3.4}$$

There is also the problem of weighting the contribution in the dissimilarity function of the different variables. In the next Section we will look for some possibilities how to calculate weights of dissimilarities.

# 4 Weights in a dissimilarity

Property prediction methods based on a dissimilarity are essentially dependent on the dissimilarity functions used. Since some description properties are more important than others the contribution of variable in a dissimilarity should has different weighting factors. Using weights for properties we compute the dissimilarity between the units as $d(X_i, X_j) = z(w_1 d_1, w_2 d_2, \ldots w_n d_n)$. For example the *Gower dissimilarity* appropriate for mixture of variable types it has the form presented in Anderberg (1973):

$$d(X_i, X_j) = \frac{\sum_{k=1}^{n} w_k(X_i, X_j) d_k(x_k^{(i)}, x_k^{(j)})}{\sum_{k=1}^{n} w_k(X_i, X_j)} \tag{4.1}$$

or Minkowski distance: $d(X_i, X_j) = \left(\sum_{k=1}^{n} (w_k(X_i, X_j) d_k(x_k^{(i)}, x_k^{(j)}))^r\right)^{\frac{1}{r}}$,

where $w_k(X_i, X_j) \geq 0$ is the weight for property $L_k$, usually not dependent on units.

The property weighting procedure should assign low weight to the contribution of the property that provides little information in the prediction process and higher weight to the property that provides more reliable information. We can distinguish between two approaches to determine weights in a dissimilarity space:

- *global*: the weights are constant over the entire dissimilarity space;
- *local*: the weights may differ among the regions of the dissimilarity space. Two types of local weighting procedure are popular. In the first approach we assign a different weight to each value of a property, so the weights are identical for all units with the value. In the second property weights are a function of the units, $w = f(X_i)$, see Wettschereck, Aha, and Mohri (1997).

In this paper we will talk only about the weights defined globally. On the other hand we distinguish between *property selection* and *property weighting* methods. In property selection algorithms we assign binary weights, so the property is deleted or accepted. These algorithms reduce the dimensionality of the space and perform best when the description properties are either highly correlated with the prediction or completely irrelevant.

Weights can be determined on the basis of:

- *specific knowledge*: the relative importance of the description properties is known in advance;
- *optimization methods* on learning set to improve prediction accuracy;

- *information theory*: using entropy and information between description variables and prediction.

The last two methods are briefly discussed here.

## 4.1 Optimization methods

One of the possible ways to determine appropriate weights in the dissimilarity function is the optimization procedure. In this procedure we want to minimize the difference between the known (measured) $y$ and the predicted values $\hat{y}$ on a training set. Prediction of the unknown value is based only on the dissimilarities between units and the optimization method using only the informations of the predicted values and dissimilarities of the units in the learning set $\mathcal{L}$.

As a measure of accuracy of prediction we usually use:

- *variance* - the average squared error:

$$e = \frac{1}{|\mathcal{L}|} \sum_{X \in \mathcal{L}} |y(X) - \hat{y}(X; \mathcal{L} - \{X\})|^2 \tag{4.2}$$

- *mean absolute distance* - the average error of prediction:

$$e = \frac{1}{|\mathcal{L}|} \sum_{X \in \mathcal{L}} |y(X) - \hat{y}(X; \mathcal{L} - \{X\})| \tag{4.3}$$

where the expression $\hat{y}(X; \mathcal{L} - \{X\})$ denotes the prediction for the unit $X$ based only on the rest of the units in our learning set.

In the experiments, see paragraph 7, we used *leave-one-out* method on the learning set to minimize the prediction error. We used one of the direct search methods named the method of Hooke and Jeeves, which can be found in any book of the optimization, see for example Burday and Garside (1987). The initial weights in the procedure are all the same. On each step of the procedure we calculate the prediction for every unit in the learning set using weights in dissimilarity function, $\hat{y}(X; \mathcal{L} - \{X\}) = f(w_1 d_1, w_2 d_2, \ldots w_n d_n)$, and the prediction accuracy, $e$. If the accuracy reduced, we have new weights and we search the new one in the direction according to the algorithm of Hooke and Jeeves method.

## 4.2 Information theory and weights

This group of weights are motivated by the information theory and are most appropriate for categorical (nominal, ordinal scale types) variables, or using the discretization methods, they are also able to deal with continuous (interval, ratio scale

types) variables. We could separate the data into equal-lenght intervals or we could use one of the discretization algorithm, see for example Doughtery et al. (1995).

Let $p(U) = (p(u_i))_{i=1}^{k}$ is the probability distribution for the random discrete variable $U$, where $p(u_i)$ is the probability that $U$ takes on the $u_i$-th value. The *entropy* of the variable $U$ is defined as

$$H(U) = -\sum_{i=1}^{k} p(u_i) \log p(u_i) \tag{4.4}$$

where $\log = \log_2$ and $0 \log 0 = 0$. The entropy $H(U)$ is a measure of randomness of a random variable. The entropy is maximal when all $p(u_i)$ are equal (uniform distribution) and is greater or equal to 0: $0 \leq H(U) \leq \log k$.

*Information* of two variables $U$ and $V$ is a measure of common information shared between these two variables. Let $p(U) = (p(u_i))_{i=1}^{k}$ and $p(V) = (p(v_j))_{j=1}^{s}$ are probability distributions for variables $U$ and $V$ and $p(UV) = (p(u_i, v_j))_{i=1\,j=1}^{k\,s}$ probability distribution for the pair of variables $(U, V)$. The information of this two variables is defined as:

$$I(U, V) = \sum_{i=1}^{k} \sum_{j=1}^{s} p(u_i, v_j) \log \frac{p(u_i, v_j)}{p(u_i)p(v_j)} \tag{4.5}$$

The information is greater or equal to 0 and the maximal value is limitted with the entropies of the variables: $0 \leq I(U, V) \leq \min(H(U), H(V))$.

Here are some possibilities for the weighting factors in dissimilarity for the description property $L_i$, see Wettschereck et al. (1997):

- *Mutual information* (Shanon, 1948): $w_k = I(L_0, L_k)$.
  If $L_i$ provides no information about the $L_0$, the mutual information will be 0.
- *Information gain* (Quinlan, 1986): $w_k = H(L_0) - \sum_{i=1}^{n_k} p_i \sum_{j=1}^{n_0} -p_{ij} \log p_{ij}$.
- *Raiski coefficient* (Ustinov and Felinger, 1973): $R(L_i \to L_0) = \dfrac{I(L_i, L_0)}{H(L_0)}$.
  Raiski coefficient is a measure of "functional dependence" of the prediction property $L_0$ on description property $L_i$. If $R(L_i \to L_0) = 1$, then $L_0$ is a function of $L_i$.

# 5 Dissimilarity based prediction

We will present the following two methods based on the dissimilarity for prediction property:

- *k-nearest neighbors algorithm*: the prediction is based on the values of the $k$-nearest units in the learning set,

- *neighborhood subspace approximation*: we define subspaces, such as line and plane, in the (semi-)metric space induced by the selected dissimilarity. Then we look for the most suitable subspace where the unit with the unknown value is lying and use a (generalized) linear inter/extra-polation in this subspace for approximation of the unknown property.

## 5.1 Prediction based on $k$-nearest neighbors

For the unit $Z$ the unknown value $y(Z)$ is approximated with the values $y(X_i)$ of the $k$ units in learning set $\mathcal{L}$ that are closest to the unit $Z$ with respect to the chosen dissimilarity. $N(Z,k) := \{X_1, \ldots X_k\}$, $k$-nearest neighbors. There are two possibilities how to calculate the unknown value:

$$\hat{y}(Z) = \text{median}Y\{y(X) : X \in N(Z,k)\} \tag{5.1}$$

$$\hat{y}(Z) = \frac{\sum_{i=1}^{k} \alpha_i y(X_i)}{\sum_{i=1}^{k} \alpha_i}, \quad \alpha_i = h(d(Z,X_i)) \tag{5.2}$$

If $h \equiv c$, any constant, then the estimate for unknown property for the unit $Z$ is the mean value of the $y$ values for the $k$-nearest neighbors

$$\hat{y}(Z) = \frac{1}{k}\sum_{i=1}^{k} y(X_i) \tag{5.3}$$

On the other hand the function $h$ could be any function which is decreasing with the dissimilarity between unit $Z$ and the unit $X_i$ in the neighborhood, for example:

$$h(d(Z,X_i)) = \frac{1}{d(Z,X_i)^p}, p \in \mathbb{N}, \tag{5.4}$$

or a common Gaussian function:

$$h(d(Z,X_i)) = \exp\left(\frac{-d^2(Z,X_i)}{2^K}\right) \tag{5.5}$$

where parameter $K$ determines how quickly weights decline.

## 5.2 Prediction using linear inter/extra-polation along line in semimetric space

In the dissimilarity space $(\mathcal{E}, d)$ we define *a ray* from unit $X_1$ through unit $X_2$ as

$$[X_1, X_2\rangle := \{Z : \ |d(X_1, X_2) - d(X_1, Z)| \ = \ d(X_2, Z)\} \tag{5.6}$$

and *a line* as a union of two rays

$$\langle X_1, X_2\rangle := [X_1, X_2\rangle \cup [X_2, X_1\rangle \tag{5.7}$$

Suppose that for the unit $Z \in \mathcal{E} \setminus \mathcal{L}$ there exists a line such that unit $Z$ lies on it, $Z \in \langle X_1, X_2\rangle$ , $X_1, X_2 \in \mathcal{L}$, then we can use linear inter/extra-polation of the unknown property along this line to approximate $y(Z)$ :

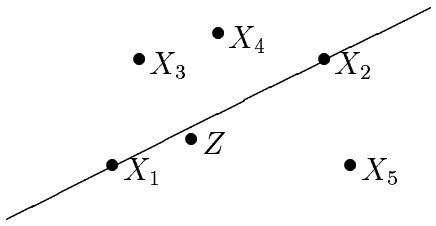
$$\hat{y}(Z; X_1, X_2) = \begin{cases} \dfrac{y(X_1)d(Z, X_2) + y(X_2)d(Z, X_1)}{d(X_1, X_2)} & \text{for } Z \in C_1 \\ \dfrac{y(X_2)d(Z, X_1) - y(X_1)d(Z, X_2)}{d(X_1, X_2)} & \text{for } Z \in C_2 \\ \dfrac{-y(X_2)d(Z, X_1) + y(X_1)d(Z, X_2)}{d(X_1, X_2)} & \text{for } Z \in C_3 \end{cases} \tag{5.8}$$

$$C_1 = \{\ Z \mid d(Z, X_1) \leq d(X_1, X_2) \text{ and } d(Z, X_2) \leq d(X_1, X_2)\} \tag{5.9}$$

$$C_2 = \{\ Z \mid Z \notin C_1 \text{ and } d(Z, X_1) \geq d(Z, X_2)\} \tag{5.10}$$

$$C_3 = \{\ Z \mid Z \notin C_1 \text{ and } d(Z, X_1) < d(Z, X_2)\} \tag{5.11}$$

Usually there is no such line, but we can often find lines to which the unit $Z$ is close and we choose one of these lines (usually first found) for property approximation. Line for condition $C_1$:



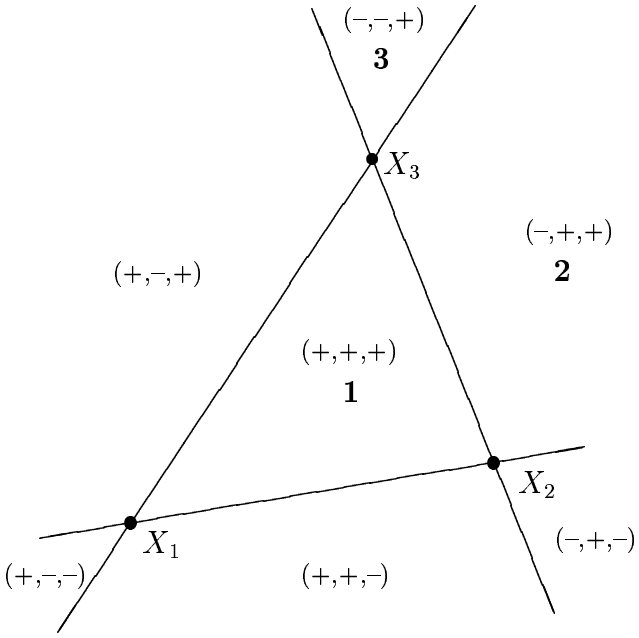

If there is no appropriate line, we can generalize the prediction method to the neighborhood subspaces of higher dimensions nearly containing the unit $Z$:

$$\hat{y}(Z; X_1, \ldots X_k) = \frac{1}{V(X_1, \ldots, X_k)} \sum_{i=1}^{k} \sigma_i y(X_i) V(X_1, \ldots, X_{i-1}, Z, X_{i+1}, \ldots, X_k) \tag{5.12}$$

$\sigma_i \in \{-1, 1\}$ and $V(X_1, \ldots, X_k)$ is the volume of the polyhedron spanned by the units $X_1, \ldots, X_k \in \mathcal{L}$.

The procedure is explained in details by Gower and Legendre (1986).

**Example**: Generalized linear inter/extra-polation in a plane




$$\hat{y}(Z; X_1, X_2, X_3) = \begin{cases} \dfrac{y(X_1)V(Z, X_2, X_3) + y(X_2)V(X_1, Z, X_3) + y(X_3)V(X_1, X_2, Z)}{V(X_1, X_2, X_3)} & Z \in \mathbf{1} \\ \dfrac{-y(X_1)V(Z, X_2, X_3) + y(X_2)V(X_1, Z, X_3) + y(X_3)V(X_1, X_2, Z)}{V(X_1, X_2, X_3)} & Z \in \mathbf{2} \\ \dfrac{-y(X_1)V(Z, X_2, X_3) - y(X_2)V(X_1, Z, X_3) + y(X_3)V(X_1, X_2, Z)}{V(X_1, X_2, X_3)} & Z \in \mathbf{3} \\ \ldots & \ldots \end{cases} \tag{5.13}$$

# 6 Experimental results

In this Section the results of the methods mentioned in previous Sections on two selected data sets are presented.

We used the learning/test set methodology to evaluate the prediction accuracy of the algorithms. Each data set was randomly divided into $k$ subsamples - $\mathcal{T}_i$ test

set, $\mathcal{E} \setminus \mathcal{T}_i$ learning set:

$$\mathcal{E} = \cup_{i=1}^{k} \mathcal{T}_i \qquad \mathcal{T}_i \cap \mathcal{T}_j = \emptyset;\ i \neq j \tag{6.1}$$

and the same learning and test sets were used for each methods.

We measure the prediction accuracy by the relative error

$$e = \Big(\sum_{i=1}^{k} \sum_{X \in \mathcal{T}_i} \frac{|y(X) - \hat{y}(X)|}{y(X)}\Big) * \frac{100}{|\mathcal{E}|}. \tag{6.2}$$

Here we mention another possibility of using Raiski coefficient. In our experiments we calculated the symmetric matrix of Raiski coefficients between pairs of description variables

$$R_{ij} = R(L_i \leftrightarrow L_j) = \frac{I(L_i, L_j)}{H(L_i L_j)};\ i, j = 1 \ldots n, \tag{6.3}$$

measuring the functional dependencies between them. If the $R_{ij}$ is almost one, let us say $\geq 0.9$, than we select only one of the description properties $L_i$ and $L_j$ for the prediction procedure. So we can reduce the dimensionality of our dissimilarity space. Afterwards we can use optimization methods to determine weights in the dissimilarity and with this reduce the computational time.

## 6.1 Life expectancy

First example is taken from the World Almanac and Book of Facts 1993, New York: Pharos Books and it could be found at the address
`gopher://jse.stat.ncsu.edu/00/jse/data/televisions.dat`. For each of the 38 selected countries we have four description variables: people per television, people per physician, female life expectancy and male life expectancy (according to 1990 population figures). Life expectancy was chosen for the prediction variable.

For the estimation dissimilarities between units we used the weighted Manhattan distance, $d(X_i, X_j) = \sum_{k=1}^{4} w_k |x_k^{(i)} - x_k^{(j)}|$. The prediction of the life expectancy were calculated with the interpolation lines in the neighborhood of six units. Data set were randomly divided into five disjunct sets. On each step we used one of them as test set and the others as learning set.

Prediction results are presented in Table 1. As we could see the predictions were calculated without weights in a dissimilarity and with weights calculated in two different processes. Here are the weights for the description variables calculated for one of the five learning set:

Raiski coefficient $R(L_i \rightarrow L_0),\ i = 1 \ldots 4$: 0.12, 0.30, 0.67, 0.73

Optimization methods (Hooke and Jeeves): 0.08, 0.09, 0.41, 0.42

**Table 1:** Relative error for life expectancy prediction.

| | Without weights | Weights | |
|---|---|---|---|
| | | Raiski | Optimization |
| Lines - relative error (%) | 5.54 | 4.00 | 1.23 |
| 5-NN - relative error (%) | 7.49 | 6.52 | 1.24 |

## 6.2 QSAR

The property prediction problem is one of the basic problems of QSAR (quantitative structure-activity relationship) studies. The interest in quantification of the similarity between chemical structures arises from the expectation that molecules with similar structures also have similar physicochemical properties and biological activities. The basic assumption is that the molecules with similar structures have also similar biological activities. So it can be assumed that the property values change smoothly over structurally similar compounds and this assumption leads us to test the metric approach in the property prediction.

For the description of the molecules the graph-theoretical methods have been largely applied in the QSAR. In this approach we use structural indices based on molecular graphs, what express in numerical form the structure of molecular graph, see for example Balaban et al. (1983). Usually compounds share the same molecular skeleton on which different substituents are bonded. With the prediction procedure we want to restrict the potential candidates for further analyzes.

Data base for the application of the QSAR can be found at
`http://www.awod.com/netsci/Issues/Jan96/feature1.html`
or in *Journal of Medicinal Chemistry*, 1988, **31** (11). There is 46 compounds. We calculated 24 graph-theoretic descriptors and used 8 descriptions of atoms:
– the minimal path from attaching place of substituent to one chosen atom on the basic compound;
– for each atom type (there are only 7 different types: Br, Cl, F, C, N, O, H), in substituent we describe how many times it appears in substituent.

The prediction variable here is biological activity.

For dissimilarity we used combination of two distances, Euclidean and Manhattan,

$$d(X_i, X_j) = \sqrt{\sum_{k=1}^{24} w_k (x_k^{(i)} - x_k^{(j)})^2} + \sum_{k=1}^{8} w_k |x_k^{(i)} - x_k^{(j)}| \qquad (6.4)$$

The Euclidean distance was used for the variables getting from graph-theoretical descriptors and the Mahhattan distance for the variables describing the number of different atoms.

In this experiment we succeeded to reduce the number of description variables with the procedure described at the beginning of this Section. We calculated the Raiski coefficient between pairs of variables and than we rejected one description variable in each pair where Raiski coefficient was $\geq 0.9$. After this procedure we got sixteen description variables instead of thirty two at the beginning. The results of prediction procedures under different conditions are represented in Table 2. In this experiment we used 4-nearest neighbors algorithm.

**Table 2:** Relative error for QSAR.

| | All variables | | Selected variables | |
|---|---|---|---|---|
| | Without weights | Raiski | Without weights | Optimization |
| Relative error (%) | 19.13 | 17.91 | 18.63 | 15.46 |

# 7 Conclusion

In the paper we reviewed several weighting methods in property prediction algorithms based on dissimilarities with intention to improve the accuracy of predictions. We represented some of our experimental results showing that a weighting steps in methods could improve accuracy of the predictions, especially in combination with selecting and weighting scheme.

Another important question which needs further study is the appropriateness of a given prediction model for a given data set, and on the other hand, combination of dissimilarities into new dissimilarity using in prediction model.

# References

[1] Anderberg, M.R. (1973): *Cluster Analysis for Applications.* New York: Academic Press.

[2] Balaban, A.T., Motoc, I., Bonchev, D., and Mekenyan, O. (1983): *Topological Indices for Structure-Activity Correlations.* Springer, Berlin, Heidelberg.

[3] Batagelj, V. and Bren, M.(1995): Comparing Similarity Measures. *Journal of Classification,* **12**, 73-90.

[4] Batagelj, V. (1989): Similarity Measures Between Structured Objects. In A. Graovac (Ed.), *Proceedings of International Course and Conference on the Interfaces between Mathematics, Chemistry and Computer Science, Dubrovnik 1988.* Studies in Physical and Theoretical Chemistry, **63**, 25-40. Amsterdam: Elsevier/Noth-Holland.

[5] Burday, B.D. and Garside, G.R. (1987): *Optimisation Methods in Pascal.* Edward Arnold.

[6] Cailliez, F. and Pages, J.P. (1976): *Introduction á l'analyse des données.* S.M.A.S.H. University of Paris VI.

[7] Doughtery, J., Kohavi, R., and Sahami M. (1995): Supervised and Unsupervised Discretization of Continuous Features. In Arman Prieditis and Stuart Russell (Eds.), *Machine Learning: Proceedings of 12th International Conference.* San Francisco: Morgan Kaufmann Publishers.

[8] Gower, J.C. (1971): A General Coefficient of Similarity and Some of Its Property. *Biometrics,*, **27**, 857-871.

[9] Gower, J.C. and Legendre, P. (1986): Metric and Euclidean Properties of Dissimilarity Coefficients. *Journal of Classification,* **3**, 5-48.

[10] Hoang, M. Thu (1978): A Short Algorithm to Transform Dissimilarities into Distance. *Proceedings of the computer science and statistics.* Institute of Statistics, North Carolina State University.

[11] Joly, S. and Le Calve, G. (1986). Etude des puissances d'une distance. *Statistique et Analyse des Données,* **11**, 30-50.

[12] Joly, S. and Le Calve, G. (1994): Similarity functions. In B. Van Cutsem (Ed.), *Classification and Dissimilarity Analysis, Lecture Notes in Statistics.* New York: Springer-Verlag.

[13] Keržič, D., Jerman Blažič, B., and Batagelj, V. (1994): Comparison of three Different Approaches to the Property Prediction Problem. *J. Chem. Inf. Comput. Sci.,* **34**, 391-394.

[14] Krippendorff, K. (1986): Information Theory - Structural Models for Qualitative Data. *Sage University Papers, Series: Quantitative Applications in the Social Sciences, 07-062.* Beverly Hills: Sage Pub.

[15] Quinlan, J.R. (1986): Introduction of decision trees. *Machine Learning,* **1**, 81-106.

[16] Schoenberg, I.J. (1937): On certain metric spaces arising from Euclidean spaces by a change of metric and their embedding in Hilbert spaces. *Ann. of Math.*, **38**, 787-793.

[17] Schoenberg, I.J. (1938): Metric spaces and positive definite functions. *Trans. Amer. Math. Soc.*, **44**, 522-536.

[18] Shanon, C.E. (1948): A mathematical theory of communication. *Bell Systems Technology Journal,* **27**, 379-423.

[19] Ustinov, V.A. and Felinger, A.F. (1973): *Istoriko-socyal'nye issledovanja, EVM i matematika.* Mysl', Moskva.

[20] Wettschereck, D., Aha, D.W., and Mohri, T. (1997): A Review and Empirical Evaluation of Feature Weighting Methods for Lazy Learning Algorithms. *Artificial Intelligence Review,* **11**, 273-314.

[21] Wilson, D.R. and Martinez, T.R. (1997): Improved Heterogeneus Distance Functions. *Journal of Artificial Intelligence Research*, **6**, 1-34.