Discussiones Mathematicae
Probability and Statistics 25 (2005) 217–239

EFFECT OF CHOICE OF DISSIMILARITY MEASURE ON CLASSIFICATION EFFICIENCY WITH NEAREST NEIGHBOR METHOD

Tomasz Górecki

Faculty of Mathematics and Computer Science,
Adam Mickiewicz University,
Umultowska 87, 61–614 Poznań

e-mail: drizzt@amu.edu.pl

Abstract

In this paper we will precisely analyze the nearest neighbor method for different dissimilarity measures, classical and weighed, for which methods of distinguishing were worked out. We will propose looking for weights in the space of discriminant coordinates. Experimental results based on a number of real data sets are presented and analyzed to illustrate the benefits of the proposed methods. As classical dissimilarity measures we will use the Euclidean metric, Manhattan and post office metric. We gave the first two metrics weights and now these measures are not metrics because the triangle inequality does not hold. Howeover, it does not make them useless for the nearest neighbor classification method. Additionally, we will analyze different methods of tie-breaking.

Keywords and Phrases: nearest neighbor method, discriminant coordinates, dissimilarity measures, estimators of classification error.

2000 Mathematics Subject Classification: 62H30, 62J05.

1. Introduction

Suppose that a training sample $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_N)$ has been collected by sampling from a population P consisting of K subpopulations or classes $G_1, \ldots, G_K$. The ith observation in $\mathbf{z}$ is a pair denoted by $\mathbf{z}_i = (\mathbf{x}_i, y_i)$ where $\mathbf{x}_i$ is a p-dimensional feature vector and y_i is the label for recording class membership. The corresponding pair for an unclassified observation is denoted by $\mathbf{z}_0 = (\mathbf{x}_0, y_0)$. In this case, $\mathbf{x}_0$ is observed, whereas the class label y_0 is unobserved. The object of classification is to construct a classification rule for predicting the membership of an unclassified feature vector $\mathbf{x}_0 \in P$. An automated classifier can be viewed as a method of estimating the posterior probability of membership in G_k. The classification rule assigns $\mathbf{x}_0$ to the group with the largest posterior probability estimate. We denote the posterior probability of membership in G_k by

$$p_k(\mathbf{x}_0) = P(y_0 = k|\mathbf{x}_0).$$

One of the important non-parametric classifiers is a J-nearest neighbor classifier (J-NN classifier). The estimator of $p_k(\mathbf{x}_0)$ produced by the J-NN classifier is the sample proportion of the J-nearest neighbors belonging to G_k:

$$\hat{p}_k(\mathbf{x}_0) = \frac{1}{J}\sum_{i=1}^{N} I(\rho(\mathbf{x}_0, \mathbf{x}_i) \leq d_J(\mathbf{x}_0)) I(y_i = k), \quad \text{k=1,...,K}, \tag{1}$$

where $k = 1, \ldots, K$, $I(A)$ is the indicator function of the event A, $d_J(\mathbf{x}_0)$ is the J-th distance from the point $\mathbf{x}_0$ to the points $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and ρ is a given measure. Usually, ρ is an Euclidean metric

$$\rho^2(\mathbf{x}_0, \mathbf{x}_i) = \| \mathbf{x}_0 - \mathbf{x}_i \|^2 = (\mathbf{x}_0 - \mathbf{x}_i)'(\mathbf{x}_0 - \mathbf{x}_i). \tag{2}$$

If the features are correlated, we can use the Mahalanobis distance. If we have a tie among the largest estimates of group membership, various methods of tie-breaking can be used as described in Section 3.1.

The great effectiveness of the J-NN rule when the number of observations increases to infinity is well known [2]. However, in most real situations, the number of available observations is usually small, which often leads to dramatic degradations of the nearest neighbor method classification accuracy. The nearest neighbor rule is a suboptimal procedure. Its use will usually lead to an error rate greater than the minimum possible, the

Bayesian rate. However, with a very large number of samples (patterns), the error rate is never worse than twice the Bayes rate. On the other hand, unlike the Bayessian classifier, the nearest neighbor rule does not require estimation of the conditional probability density function for each class, so it is easier to implement. It was proved in [3] that even though we have rules, such as the J-nearest neighbor rule, that are universally consistent (i.e., they asymptotically provide optimal performance for any distribution), their finite sample performance can be extremely bad for some distributions. This explains the increasing interest in finding a measure of dissimilarity that helps improve the nearest neighbor method classification performance in small data sets. The Euclidean distance (2) is sensitive to change in scale. A good example of this sensitivity can be found in [10]. So to overcome this problem each variable could be scaled by dividing it by the range or standard deviation. This method will remove the dependence on the units of measurement, but it creates other problems described by [10]. We scaled variables differently, viz. we used weights to scale variables. Furthermore, we will assume that features are poorly correlated and as measures of distances we will use measures like (2) and weighed measures.

In Section 2, we will define dissimilarity measures and we will classify them comparatively, and give some equivalence relation among these measures. Also in this section, we will describe the proposed method of weights introduction. Section 3 is devoted to research on the efficiency of the proposed method on real data sets, and on methods for tie-breaking in the nearest neighbor method. The last section is devoted to reviewing on the methods that were used.

2. Measures of dissimilarity

Information on measures of dissimilarity can be found in [6] and [7]. On the basis of these monographs and [10] we can define some class of functions.

Definition 1. Function $\rho : X \times X \to \mathbb{R}$ is called a measure of dissimilarity if:

1. $\forall \mathbf{x}, \mathbf{y} \in X \;\; \rho(\mathbf{x}, \mathbf{y}) = \rho(\mathbf{y}, \mathbf{x})$,
2. $\forall \mathbf{x}, \mathbf{y} \in X \;; \mathbf{x} \neq \mathbf{y} \;\; \rho(\mathbf{x}, \mathbf{y}) > 0$,
3. $\forall \mathbf{x} \in X \rho(\mathbf{x}, \mathbf{x}) = 0$.

So it is a symmetrical non-negative function. From the definition of a metric we delete the triangle inequality. Apparently this inequality is not needed because we are interested only in ranking distances to point $\mathbf{x}_0$ and not in distances between all points. The following are the most common measures of dissimilarity between points $\mathbf{x}$ and $\mathbf{y}$:

- $$\rho_1(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{p} |x_i - y_i|,$$

  $$\bar{\rho}_2(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2},$$

- $$\tilde{\rho}_2(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{p} (x_i - y_i)^2,$$

- $$\rho_3(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{1 + |x_i - y_i|} \frac{1}{2^i},$$

- $$\rho_4(\mathbf{x},\mathbf{y}) = \begin{cases} \tilde{\rho}_2(0,\mathbf{y}) + \tilde{\rho}_2(\mathbf{x},0), & \text{for } \mathbf{x} \neq \mathbf{y}, \\ 0, & \text{for } \mathbf{x} = \mathbf{y}, \end{cases}$$

- $$\rho_5(\mathbf{x},\mathbf{y}) = \begin{cases} 0, & \text{for } \mathbf{x} = \mathbf{y}, \\ \bar{\rho}_2(\mathbf{x},\mathbf{y}), & \text{for } \mathbf{x} \neq \mathbf{y} \text{ and } \mathbf{x}\ ,\ \mathbf{y} \ \text{ are lying on a} \\ & \text{straight line with 0 point,} \\ \bar{\rho}_2(0,\mathbf{y}) + \bar{\rho}_2(\mathbf{x},0), & \text{otherwise,} \end{cases}$$

- $$\rho_6(\mathbf{x},\mathbf{y}) = (\mathbf{x} - \mathbf{y})'\mathbf{S}_y^{-1}(\mathbf{x} - \mathbf{y}),$$

- $\rho_7(\mathbf{x},\mathbf{y}) = \left\{ \sum_{i=1}^{p} |x_i - y_i|^{\alpha} \right\}^{\frac{1}{\alpha}},$

- $\rho_8(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{x_i + y_i}$ (for positive variables only),

- $\rho_9(\mathbf{x},\mathbf{y}) = 1 - \frac{2\sum_{i=1}^{p} \min(x_i, y_i)}{\sum_{i=1}^{p} (x_i - y_i)}$ (also for positive variables only).

Some of these measures of dissimilarity are metrics, e.g., ρ_1 (city-block metric or Manhattan metric), $\bar{\rho}_2$ (Euclidean metric), ρ_3 (Frechét metric), ρ_4 (post office metric), ρ_5 (subway metric). Measure ρ_6 is the so-called Mahalanobis distance, where $\mathbf{S}_y$ is the estimator of covariance matrix class of observation $\mathbf{y}$ (much information about this distance can be found in [4]. Minkowski metric (ρ_7) includes both the Euclidean and city-block metric. ρ_8 is Canberra metric and ρ_9 is Czekanowski coefficient.

Remark 1. Each metric is a measure of dissimilarity.

In most cases it turns out that not all features have the same importance in classification. It seems reasonable in this case to respect this fact giving weights to features which effect the classification. Also, if features are measured on different scales, giving weights should improve the performance of classification. We also want the functions thus changed to remain measures of dissimilarity. We proposed the following theorem:

Theorem 1. *If $d(x,y)$ is a metric in $\mathbb{R}$, then*

$$\tilde{d}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} w_i d(x_i, y_i)$$

is a metric in $\mathbb{R}^n$, whenever $w_i \geq 0\ \forall i$ and $\exists i\ w_i \neq 0$.

Proof. Let us check if all conditions in the definition of a metric are fulfilled.

1^o $\tilde{d}(\mathbf{x},\mathbf{x}) = \sum\limits_{i=1}^{n} w_i d(x_i, x_i) = 0,$

2^o $\tilde{d}(\mathbf{x},\mathbf{y}) = \sum\limits_{i=1}^{n} w_i d(x_i, y_i) = \sum\limits_{i=1}^{n} w_i d(y_i, x_i) = \tilde{d}(\mathbf{y},\mathbf{x}),$

3^o if $\mathbf{x} \neq \mathbf{y}$ then $\tilde{d}(\mathbf{x},\mathbf{y}) = \sum\limits_{i=1}^{n} w_i d(x_i, y_i) \geq \sum\limits_{i=1}^{n} \min\limits_{1\leq j\leq n} [w_j d(x_j, y_j)]$

$\geq n \min\limits_{1\leq j\leq n} [w_j d(x_j, y_j)] \geq 0,$ because $\forall 1 \leq j \leq n \; w_j d(x_j, y_j) \geq 0,$

4^o $\tilde{d}(\mathbf{x},\mathbf{z}) = \sum\limits_{i=1}^{n} w_i d(x_i, z_i) \leq \sum\limits_{i=1}^{n} w_i [d(x_i, y_i) + d(y_i, z_i)]$

$= \sum\limits_{i=1}^{n} w_i d(x_i, y_i) + \sum\limits_{i=1}^{n} w_i d(y_i, z_i) = \tilde{d}(\mathbf{x},\mathbf{y}) + \tilde{d}(\mathbf{y},\mathbf{z}).$

■

Corollary 1. *If $d(x, y)$ is a dissimilarity measure in $\mathbb{R}$, then*

$$\bar{d}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} w_i d(x_i, y_i)$$

is a dissimilarity measure in $\mathbb{R}^n$, where $w_i \geq 0 \; \forall i$ and $\exists i \; w_i \neq 0$.

In this way we can create varied measures of dissimilarity, but not all measures of dissimilarity can change the result of classification. Therefore, we will introduce a congruent relation between dissimilarity measures.

Definition 2. Given the dissimilarity measure ρ, a sequence $\{\mathbf{x}_k, \mathbf{x}_{a_2}, \ldots, \mathbf{x}_{a_n}\}$ with $a_i \in \{1, \ldots, n\}\backslash\{k\}$, such that, whatever i, j with $a_i < a_j$ implies $\rho(\mathbf{x}_k, \mathbf{x}_{a_i}) \leq \rho(\mathbf{x}_k, \mathbf{x}_{a_j})$ will be a system $D(\rho, \mathbf{x}_k)$ of points in the set $E = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ with respect to the dissimilarity measure ρ.

In these sequence of points dissimilarity ρ increases when we take the next point.

Definition 3. The dissimilarity measures ρ_1 and ρ_2 are congruent if

$$\forall n \geq 3 \; \forall \mathbf{x}_i \; i = \{1, \ldots, n\} \; D(\rho_1, \mathbf{x}_i) = D(\rho_2, \mathbf{x}_i).$$

Theorem 2. *The congruence relation φ between dissimilarity measures is an equivalence relation.*

Proof.

1^o Reflexivity

$$\varphi(\rho_1, \rho_1) = \varphi(\rho_1, \rho_1),$$

2^o Symmetry

$$\varphi(\rho_1, \rho_2) = \varphi(\rho_2, \rho_1),$$

3^o Transitivity

$$\varphi(\rho_1, \rho_2) \wedge \varphi(\rho_2, \rho_3) \Longrightarrow \varphi(\rho_1, \rho_3).$$

Let us fix i.

$$(D(\rho_1, \mathbf{x}_i) = D(\rho_2, \mathbf{x}_i)) \wedge (D(\rho_2, \mathbf{x}_i) = D(\rho_3, \mathbf{x}_i))$$

$$\Longrightarrow D(\rho_1, \mathbf{x}_i) = D(\rho_3, \mathbf{x}_i).$$ ■

As an equivalence relation, it divides the set of measures of dissimilarity into classes of equivalence. Only measures of dissimilarity belonging to different classes of abstraction can change classifications. To show that two measures of dissimilarity are not congruent, it is sufficient to find three points in $\mathbb{R}^2$ and prove that the first of these measures generates a system of points different from the other.

Example 1. We will show that the measures of dissimilarity

$$\rho_1(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i|$$

and

$$\rho_2(\mathbf{x},\mathbf{y}) = \frac{\sum\limits_{i=1}^{n} |x_i - y_i|}{1 + \sum\limits_{i=1}^{n} |x_i - y_i|}$$

are congruent. We have to show that $\forall \mathbf{x}\ D(\rho_1, \mathbf{x}) = D(\rho_2, \mathbf{x})$. Suppose that for some $\mathbf{y}, \mathbf{z}$

$$\rho_1(\mathbf{x},\mathbf{y}) \le \rho_1(\mathbf{x},\mathbf{z}).$$

Then

$$\rho_2(\mathbf{x},\mathbf{y}) = \frac{\rho_1(\mathbf{x},\mathbf{y})}{1+\rho_1(\mathbf{x},\mathbf{y})} \le \frac{\rho_1(\mathbf{x},\mathbf{z})}{1+\rho_1(\mathbf{x},\mathbf{z})} = \rho_2(\mathbf{x},\mathbf{z}),$$

because $\forall 0 < x < y$

$$\frac{x}{1+x} - \frac{y}{1+y} = \frac{x-y}{(1+x)(1+y)} < 0.$$

We have shown that for any $\mathbf{x}$ measures of dissimilarity $\rho_1,\ \rho_2$ produce the same system of points, i.e., they are congruent.

Example 2. We will show that the measures of dissimilarity

$$\rho_1(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i|$$

and

$$\rho_2(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} (x_i - y_i)^2$$

are not congruent.

Suppose that $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^2$ and to simplify calculations (without loss of generality) suppose that $\mathbf{x} = (0,0)$. Then the following condition has to be fulfilled

$$\begin{cases} \rho_1(\mathbf{x},\mathbf{y}) < \rho_1(\mathbf{x},\mathbf{z}) \\ \rho_2(\mathbf{x},\mathbf{y}) > \rho_2(\mathbf{x},\mathbf{z}) \end{cases} \quad \text{or} \quad \begin{cases} \rho_1(\mathbf{x},\mathbf{y}) > \rho_1(\mathbf{x},\mathbf{z}) \\ \rho_2(\mathbf{x},\mathbf{y}) < \rho_2(\mathbf{x},\mathbf{z}), \end{cases}$$

$$\begin{cases} |y_1| + |y_2| < |z_1| + |z_2| \\ y_1^2 + y_2^2 > z_1^2 + z_2^2 \end{cases} \quad \text{or} \quad \begin{cases} |y_1| + |y_2| > |z_1| + |z_2| \\ y_1^2 + y_2^2 < z_1^2 + z_2^2. \end{cases}$$

Now suppose that $\mathbf{z} = (3, 4)$. We have

$$\begin{cases} |y_1| + |y_2| < 7 \\ y_1^2 + y_2^2 > 25 \end{cases} \quad \text{or} \quad \begin{cases} |y_1| + |y_2| > 7 \\ y_1^2 + y_2^2 < 25. \end{cases}$$

A graphical solution of this inequalities system is presented in Figure 1.

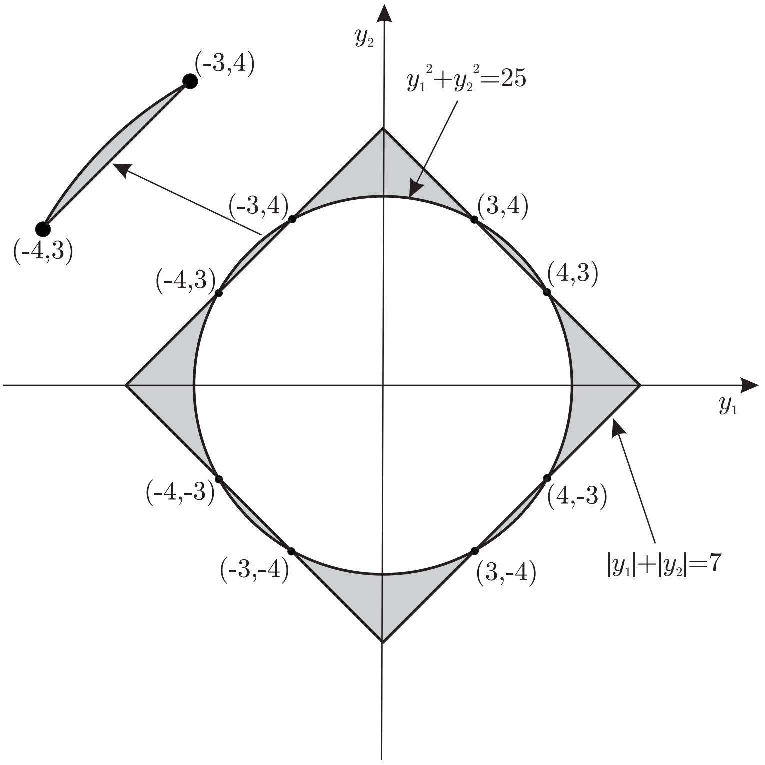

Figure 1. The shaded area is the set of plane points in which the measures of dissimilarity under consideration are different.

For example let us consider point $\mathbf{y} = (0, 6)$ for which we have:

$$\rho_1(\mathbf{x}, \mathbf{y}) = 6, \ \ \rho_1(\mathbf{x}, \mathbf{z}) = 7,$$

$$\rho_2(\mathbf{x}, \mathbf{y}) = 36, \ \ \rho_2(\mathbf{x}, \mathbf{z}) = 25,$$

$$D(\rho_1, \mathbf{x}) = \{\mathbf{x}, \mathbf{y}, \mathbf{z}\},$$

$$D(\rho_2, \mathbf{x}) = \{\mathbf{x}, \mathbf{z}, \mathbf{y}\}.$$

These measures of dissimilarity are not congruent for some sets of points since they produce different systems of points and they give distinct classifications through the nearest neighbor method.

Example 3. We will show that the measures of dissimilarity

$$\rho_1(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i|$$

and

$$\rho_2(\mathbf{x}, \mathbf{y}) = \begin{cases} 0, & \text{for } \mathbf{x} = \mathbf{y}, \\ \sqrt{x_1^2 + x_2^2} + \sqrt{y_1^2 + y_2^2}, & \text{for } \mathbf{x} \neq \mathbf{y} \end{cases}$$

are not congruent. Again suppose that $\mathbf{x} = (0, 0)$. One of the following conditions has to be fulfilled:

$$\begin{cases} |y_1| + |y_2| < |z_1| + |z_2| \\ \sqrt{y_1^2 + y_2^2} > \sqrt{z_1^2 + z_2^2} \end{cases} \quad \text{or} \quad \begin{cases} |y_1| + |y_2| > |z_1| + |z_2| \\ \sqrt{y_1^2 + y_2^2} < \sqrt{z_1^2 + z_2^2} \end{cases}$$

$$\begin{cases} |y_1| + |y_2| < |z_1| + |z_2| \\ y_1^2 + y_2^2 > z_1^2 + z_2^2 \end{cases} \quad \text{or} \quad \begin{cases} |y_1| + |y_2| > |z_1| + |z_2| \\ y_1^2 + y_2^2 < z_1^2 + z_2^2 \end{cases}$$

as in the situations in the previous example.

Example 4. We will show that the measures of dissimilarity

$$\rho_1(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} (x_i - y_i)^2$$

and

$$\rho_2(\mathbf{x}, \mathbf{y}) = \begin{cases} 0, & \text{for } \mathbf{x} = \mathbf{y}, \\ \sqrt{x_1^2 + x_2^2} + \sqrt{y_1^2 + y_2^2}, & \text{for } \mathbf{x} \neq \mathbf{y} \end{cases}$$

are not congruent. Suppose that $\mathbf{x} = (3, 3)$. One of the following conditions has to be fulfilled:

$$\begin{cases} (y_1 - 3)^2 + (y_2 - 3) < (z_1 - 3)^2 + (z_2 - 3)^2 \\ \sqrt{18} + \sqrt{y_1^2 + y_2^2} > \sqrt{z_1^2 + z_2^2} + \sqrt{18} \end{cases}$$

or

$$\begin{cases} (y_1 - 3)^2 + (y_2 - 3) > (z_1 - 3)^2 + (z_2 - 3)^2 \\ \sqrt{18} + \sqrt{y_1^2 + y_2^2} < \sqrt{z_1^2 + z_2^2} + \sqrt{18}. \end{cases}$$

Now suppose that $\mathbf{z} = (3, 4)$, we have

$$\begin{cases} (y_1 - 3)^2 + (y_2 - 3) < 1 \\ y_1^2 + y_2^2 > 25 \end{cases} \quad \text{or} \quad \begin{cases} (y_1 - 3)^2 + (y_2 - 3) > 1 \\ y_1^2 + y_2^2 < 25. \end{cases}$$

A graphical solution of this inequalities system is presented in Figure 2.

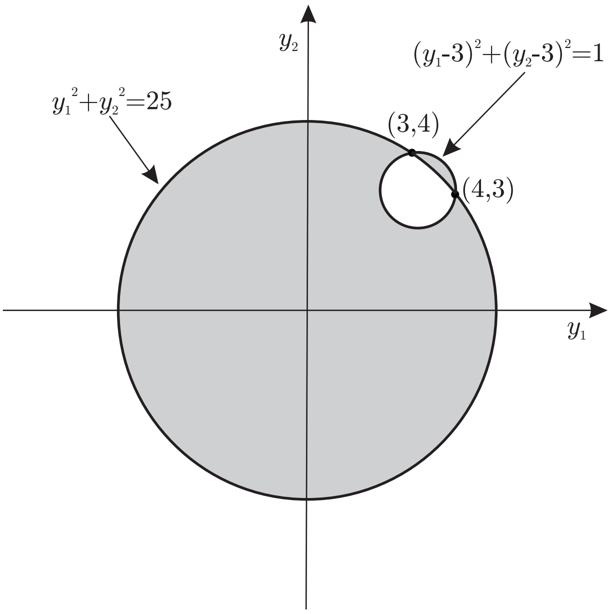

Figure 2. The shaded area is the set of plane points in which the measures of dissimilarity under consideration are different.

For example let us take $\mathbf{y} = (1, 3)$.

$$\rho_1(\mathbf{x},\mathbf{y}) = 4,\;\; \rho_1(\mathbf{x},\mathbf{z}) = 1,$$
$$\rho_2(\mathbf{x},\mathbf{y}) = \sqrt{18}+\sqrt{10},\;\; \rho_1(\mathbf{x},\mathbf{z}) = \sqrt{18}+\sqrt{25},$$
$$D(\rho_1,\mathbf{x}) = \{\mathbf{x},\mathbf{z},\mathbf{y}\},$$
$$D(\rho_2,\mathbf{x}) = \{\mathbf{x},\mathbf{y},\mathbf{z}\}.$$

Remark 2. All the foregoing considerations can be moved on $\mathbb{R}^n$, $n > 2$. It will suffice to put 0 as the remaining $n-2$ coordinates of points, to fix on 0, i.e., to consider only points on the plane.

Similarly, we can show that for some systems of weights, weighed measures of dissimilarity are not congruent. In a general case, the following theorem holds.

Theorem 3. *For any system of weights* $\{w_i : \ w_i \geq 0 \ , \ \exists \ i : w_i > 0\}$ *such that* $\exists i \neq j \ w_i \neq w_j$ *the measures of dissimilarity*

$$\bar{\rho}_1(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} w_i |x_i - y_i|$$

and

$$\bar{\rho}_2(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} w_i (x_i - y_i)^2$$

are not congruent.

Proof. It will suffice to limit ourselves to $\mathbb{R}^2$ and to three points. Additionally, without loss of generality, let us assume that $w_1 \geq w_2 \geq \ldots \geq w_n$, $\mathbf{x} = (0, 0)$ and $w_1 \neq w_2$. We want the following condition to be fulfilled:

$$\begin{cases} w_1(x_1 - y_1)^2 + w_2(x_2 - y_2)^2 < w_1(x_1 - z_1)^2 + w_2(x_2 - z_2)^2 \\ w_1|x_1 - y_1| + w_2|x_2 - y_2| > w_1|x_1 - z_1| + w_2|x_2 - z_2|. \end{cases}$$

Dividing both inequalities by w_1 we obtain (assuming $w_1 \neq 0$)

$$\begin{cases} (x_1 - y_1)^2 + w(x_2 - y_2)^2 < (x_1 - z_1)^2 + w(x_2 - z_2)^2 \\ |x_1 - y_1| + w|x_2 - y_2| > |x_1 - z_1| + w|x_2 - z_2|, \end{cases}$$

where $w = \frac{w_2}{w_1}$ and $0 < w < 1$

$$\begin{cases} y_1^2 + wy_2^2 < z_1^2 + wz_2^2 \\ |y_1| > |z_1| + w|z_2| - w|y_2|. \end{cases}$$

Now let us suppose that $y_1 \geq 0$. We start by replacing, in the second of the last expressions, inequality by equality and, next modifying y_1 so that again we have an equality.

$$\begin{cases} y_1^2 + wy_2^2 < z_1^2 + wz_2^2 \\ y_1 = |z_1| + w|z_2| - w|y_2|, \end{cases}$$

$$[|z_1| + w(|z_2| - |y_2|)]^2 + wy_2^2 < z_1^2 + wz_2^2.$$

Let us suppose that $y_2 = 0$. We have

$$(|z_1| + w|z_2|)^2 < z_1^2 + wz_2^2,$$

$$z_1^2 + 2w|z_1||z_2| + w^2z_2^2 < z_1^2 + wz_2^2,$$

$$2|z_1||z_2| + wz_2^2 < z_2^2.$$

Now let us suppose that $z_1 = 0$. We have

$$z_2^2w - z_2^2 < 0,$$

$$z_2^2(w-1) < 0, \forall z_2 \neq 0.$$

Therefore let us suppose that $z_2 = 2$. Then

$$y_1 = |z_1| + w|z_2| - w|y_2| = 2w.$$

Now we can modify y_1 so that the inequalities are fulfilled.

Let us assume

$$y_1^{'} = \varepsilon y_1,$$

where $\varepsilon > 1$. Hence we have

$$y_1^{'} > |z_1| + w|z_2| - w|y_2|.$$

Additionally, the following condition has to be fulfilled:

$$(y_1^{'})^2 + wy_2^2 < z_1^2 + wz_2^2,$$

$$(\varepsilon y_1)^2 + wy_2^2 < z_1^2 + wz_2^2,$$

$$\varepsilon^2 < \frac{z_1^2 + wz_2^2 - wy_2^2}{y_1^2},$$

$$\varepsilon < \sqrt{\frac{z_1^2 + wz_2^2 - wy_2^2}{y_1^2}},$$

$$\varepsilon < \sqrt{\frac{0 + 4w - 0}{4w^2}},$$

$$\varepsilon < \frac{\sqrt{w}}{w} = \varepsilon^*.$$

That is $\varepsilon \in (1, \varepsilon^*)$ and

$$\mathbf{x} = (0, 0),$$

$$\mathbf{y} = (2w\varepsilon, 0),$$

$$\mathbf{z} = (0, 2).$$

■

We see that for all weights we can find points for which dissimilarity measures give different systems of points. To illustrate the theorem we present the following example.

Example 5. We now present a few examples of systems of weights for weighed measures of dissimilarity and systems of points in these measures. We assume $\varepsilon = \varepsilon^* - 0.0001$. Results are presented in Table 1.

Table 1. Systems of weights and sequences of points for weighed measures of dissimilarity.

w_1, w_2	w	ε^*	y_1	$\bar{\rho}_2(\mathbf{x},\mathbf{y})$	$\bar{\rho}_2(\mathbf{x},\mathbf{z})$	$\bar{\rho}_1(\mathbf{x},\mathbf{y})$	$\bar{\rho}_1(\mathbf{x},\mathbf{z})$	$D(\bar{\rho}_2,\mathbf{x})$	$D(\bar{\rho}_1,\mathbf{x})$
$\frac{2}{3}, \frac{1}{3}$	0.50	1.41	1.41	1.3331	1.3333	0.94	0.67	$\mathbf{x},\mathbf{y},\mathbf{z}$	$\mathbf{x},\mathbf{z},\mathbf{y}$
$\frac{3}{4}, \frac{1}{4}$	0.33	1.73	1.15	0.99	1.00	0.87	0.50	$\mathbf{x},\mathbf{y},\mathbf{z}$	$\mathbf{x},\mathbf{z},\mathbf{y}$
$\frac{4}{5}, \frac{1}{5}$	0.25	2.00	1.00	0.79	0.80	0.80	0.40	$\mathbf{x},\mathbf{y},\mathbf{z}$	$\mathbf{x},\mathbf{z},\mathbf{y}$
$\frac{2}{3}, \frac{1}{2}$	0.75	1.57	1.73	1.99	2.00	1.15	1.00	$\mathbf{x},\mathbf{y},\mathbf{z}$	$\mathbf{x},\mathbf{z},\mathbf{y}$
7, 2	0.29	1.87	1.07	7.9991	8.00	7.48	4.00	$\mathbf{x},\mathbf{y},\mathbf{z}$	$\mathbf{x},\mathbf{z},\mathbf{y}$
50, 31	0.62	1.27	1.57	123.98	124.00	78.70	62.00	$\mathbf{x},\mathbf{y},\mathbf{z}$	$\mathbf{x},\mathbf{z},\mathbf{y}$

One of the ways to introduce weights is to replace our space with a space of discriminant coordinates as described in [10], which uses the term ”discriminant coordinates” since the term ”canonical variates” is reserved for the canonical correlation analysis. The measure of relative importance (influence on classification) of the ith discriminant coordinate is

$$\frac{\lambda_i}{\sum\limits_{j=1}^{s}\lambda_j}\quad i=1,\ldots,s, \tag{3}$$

where λ_i are equivalent eigenvalues. The ratios in expression (3) are used as weights in weighed measures of dissimilarity.

Another approach to finding weights is proposed in [9], namely: the following measure

$$\rho(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{p} w_{ci}(x_i - y_i)^2,$$

where c is the class containing $\mathbf{y}$. If all w_{ci} are equal to 1, this measure is an Euclidean distance but if the weights are the inverse of the variances in each dimension, the Mahalanobis distance is obtained. To find weights the authors of [9] used a fractional programming procedure. In general this measure does not fulfil Definition 1 because $\rho(\mathbf{x},\mathbf{y})$ may be different from $\rho(\mathbf{y},\mathbf{x})$ when $\mathbf{x}$ and $\mathbf{y}$ belong to different classes so the symmetry condition does not hold.

3. RESEARCH

3.1. Methods of tie-breaking

Using the nearest neighbor method we often are not certain into which class we should locate a new observation. This happens if for two (or more) classes posterior probabilities are the same, i.e.,

$$\exists j \ \max_{i \neq j} P(y_k = i|\mathbf{x}_k) = P(y_k = j|\mathbf{x}_k).$$

In this case tie-breaking may be used. Here we propose the following two methods:

- When we have a tie we add the next neighbor and if we still have a tie we add the next until a tie-break is obtained.
- We choose the class in which the nearest neighbor is located.

Theorem 4. *To solve a tie using the first strategy when the J-nearest neighbor is used for g classes, we need at most*

$$\begin{cases} \frac{J}{2}(g-2)+1, & \textit{for even } J, \\ \frac{J-1}{2}(g-2), & \textit{for odd } J \end{cases}$$

new neighbors.

Proof. Let us assume that we have a tie for the J-nearest neighbor method and g classes. The worst case is present if the maximum is achieved for two classes and these classes include all J neighbors (apart from odd J). In this case for even J we have to add at most $\frac{J}{2}(g-2)+1$ new neighbors and for odd J

$$\frac{J-1}{2}(g-2) = \frac{J}{2}(g-2) + 1 - \frac{g}{2}.$$

■

Theorem 5. *When we use the first method of tie-breaking we can be assured that this method will work for the J-nearest neighbor method if*

$$J \leq \left\lceil \frac{N}{g} \right\rceil * 2 - 1. \qquad \blacksquare$$

Proof. Again we have the worst situation when there are two classes with a maximum a posteriori probability. Then we have to add the next neighbors to break a tie and we cannot exceed N. Let J^* be the maximum J for which we can always break a tie with the first strategy. Then in each class (from g classes) it will be $\left\lceil \frac{N}{g} \right\rceil$ observations, i.e., in classes with the maximum a posterior likelihood it will be all in all $\left\lceil \frac{N}{g} \right\rceil * 2$ observations. Such a value of J^* does not break the tie if $\frac{N}{g}$ is a natural number. Assuming

$$J^* = \left\lceil \frac{N}{g} \right\rceil * 2 - 1$$

we can be assured that all ties will break.

Example 6. Let us consider the following case. Let $N = 50$ be the number of observations and let the number of classes be equal to 7. We will consider two cases:

a) $J = 6$,

b) $J = 7$.

Hence, to break the tie we need at most

a) $\frac{6}{2}(7-2) + 1 = 16$,

b) $\frac{7-1}{2}(7-2) = 15$

new observations. However, if we want to be certain that this strategy will work we have to fix J at most at

$$J = \left\lceil \frac{50}{7} \right\rceil * 2 - 1 = 15.$$

To compare these methods of tie-breaking we carried out some experiments. These experiments are presented later in Section 3.3.

3.2. Data sets

Several standard real datasets from [1] have been used (”glass”, ”ionosphere”, ”iris” and ”thyroid”). We used also the dataset ”beetles” from [7], the dataset ”school” from [5], the dataset ”irradiation” from [8]. The dataset ”turtles” is from the database of the Statistica 6.0 Pl program. Information about theses datasets are presented in Table 2.

Table 2. Information about real datasets.

Name	Number of features	Number of classes	Number of instances in classes	Number of all instaces
beetles	2	3	21,21,22	64
blood	3	4	20,20,20,20	80
crude-oil	5	3	7,11,38	56
fish	4	3	12,12,12	36
football	6	3	30,30,30	90
glass	9	6	70,76,17,139,29	214
ionosphere	34	2	225,126	351
iris	4	3	50,50,50	150
irradiation	3	4	6,14,15,10	45
school	2	3	31,28,26	85
thyroid	5	3	150,35,30	215
turtles	6	2	24,24	48

3.3. Comparison of methods

We carried out experiments to compare the behavior of the nearest neighbor method for various measures of dissimilarity. We also wanted to compare two strategies of tie-breaking. Experimental results are presented in

Table 3 and 4. Table 3 includes results of experiments with the bootstrap error estimation (50 bootstrap samples) averaging for $J = 1, 2, 3, 4, 5$ neighbors.

Table 3. Experimental results (in %) for bootstrap error estimator.

	jnn		jnnwa		jnntaxi		jnnwataxi		jnnpo	
beetles	6.70	6.99	**2.32**	2.52	7.00	7.05	3.11	3.14	36.42	36.79
blood	94.12	91.52	81.04	80.15	94.82	91.71	81.39	80.15	80.82	**78.93**
crude-oil	24.63	25.05	**14.30**	**14.30**	21.28	22.04	15.39	15.68	36.55	36.80
fish	47.87	45.30	**38.95**	40.74	50.39	46.64	39.79	40.60	76.02	75.02
football	42.27	41.89	39.24	**38.84**	41.64	41.75	39.38	39.43	47.68	47.75
glass	31.81	33.52	40.97	41.72	**30.09**	31.41	46.61	46.38	67.66	65.48
ionosphere	15.14	15.41	16.39	16.09	**10.94**	11.41	16.39	16.09	34.01	37.12
iris	4.37	4.35	3.37	**3.31**	4.96	4.92	3.38	3.34	24.18	24.71
irradiation	72.95	70.34	**67.41**	69.50	72.51	71.39	70.76	72.17	69.56	73.57
school	43.00	42.69	**7.03**	7.16	42.86	42.67	7.87	7.97	64.39	65.24
thyroid	6.93	7.39	5.33	5.27	6.07	5.43	5.21	**5.11**	26.70	27.27
turtles	19.28	20.01	11.66	**11.23**	18.95	18.95	11.66	**11.23**	37.39	39.67
mean error	34.09	33.68	**27.33**	27.53	33.46	32.95	28.42	28.44	50.11	50.70

Table 4 contains results with the 10-fold stratified cross-validation error estimation and for all datasets we have J for which the lowest error rate is achieving. The tables contain information about error rates (in %). If the discriminant space is one-dimensional, then for both strategies the weighed methods have the same accuracy (sets "ionosphere" and "turtles"). In the first and second column for all the methods there are results for the first and second tie-breaking strategy, respectively. Let us use the codes:

- jnn - method with square of the Euclidean metric,
- jnnwa - weighed method with the Euclidean metric,
- nntaxi - method with the Manhattan metric,
- jnnwataxi - weighed method with the Manhattan metric,
- jnnpo - method with the post office metric.

Table 4. Experimental results (in %) for the cross-validation error estimator.

	jnn		jnnwa		jnntaxi		jnnwataxi		jnnpo	
beetles	**1.35**	**1.35**	**1.35**	**1.35**	4.05	2.70	**1.35**	**1.35**	29.73	29.73
	3	3	4	2	2	3	1	1	1	1
blood	91.25	71.25	73.75	66.25	91.25	71.25	73.75	**63.75**	76.25	66.25
	20	22	7	10	28	14	7	29	8	31
crude-oil	16.07	16.07	**7.14**	**7.14**	14.29	12.50	8.93	8.93	26.79	26.79
	1	1	11	10	1	4	9	10	1	1
fish	36.11	33.33	30.56	**27.78**	38.89	33.33	**27.78**	**27.78**	66,67	63.89
	5	2	18	16	5	4	2	16	10	10
football	33.33	33.22	28.89	27.78	31.11	30.00	28.89	**25.56**	41.11	40.00
	10	28	6	6	32	32	23	4	7	6
glass	24.47	24.47	36.45	36.92	**23.83**	**23.83**	41.59	41.59	65.42	65.42
	1	1	2	1	1	1	1	1	19	18
ionosphere	12.82	12.82	12.82	12.82	**9.12**	**9.12**	12.82	12.82	25.93	25.93
	3	2	8	128	1	1	8	128	1	1
iris	2.00	2.00	**1.33**	**1.33**	2.67	2.67	**1.33**	**1.33**	8.00	8.00
	9	9	65	64	19	18	65	64	1	1
irradiation	62.22	55.56	55.56	**53.33**	60.00	60.00	60.00	**53.33**	60.00	62.22
	3	2	3	2	13	6	3	3	3	2
school	35.29	32.94	4.71	4.71	35.29	31.76	**3.53**	**3.53**	54.12	54.12
	5	2	4	4	5	7	1	1	1	1
thyroid	4.65	4.65	3.26	3.26	**2.79**	**2.79**	3.26	3.26	20.00	20.00
	1	1	6	6	1	1	8	8	39	38
turtles	12.50	14.58	**8.33**	**8.33**	10.42	12.50	**8.33**	**8.33**	31.25	31.25
	2	1	6	6	2	1	6	6	1	1
mean error	27.70	25.13	21.85	**20.92**	26.98	24.37	22.63	21.46	42.11	41.09

Experimental results with the bootstrap error estimator show that for all J weighed methods under consideration the error rate decreases for the Euclidean metric and also for the taxi metric. The "jnntaxi" method for all cases is slightly more accurate than the "jnn" method. However the "jnnwa" method wins with the "jnnwataxi" method for all cases. The "jnnpo" method compares dramatic regardless of J used and dataset chosen. For classical methods without the "jnnpo" the second strategy of tie-breaking seems more accurate than the first. However, for weighed methods both strategies are similar. Despite the reduction of mean error by weighed methods, for some datasets these methods increase the error rate (for dataset "glass" up to 10%) but on dataset "school" we have reduction of about 38%. On datasets "ionosphere" and "turtles" change of the performance is done only by changing from the original space to the discriminant space because it is one dimensional and hence weight is equal to 1 and the "jnnwa" and the "jnnwataxi" have the same performance.

Experimental results with cross-validation error estimates confirm the previous findings. Weighed methods win for 8 datasets and lose for 3 (we have one tie). We see that if the size of dataset is large we can have less accuracy with weighed methods than with the classic ones but this loss in accuracy is small (the greatest is for the "glass" dataset). It seems that the second strategy of tie-breaking is more accurate than the first.

4. Conclusion

Experimental results suggest that weighed methods can improve classification. It seems that the best results are achieved by the "jnnwa" method with the second strategy of tie-breaking. If we decide on classic methods we should choose the "jnntaxi" method.

The increase in performance of weighed methods can partly be a result of classification in different spaces and not arising from the introduction of weights. So it is interesting to see how classic methods work in discriminant space.

References

[1] C. Blake, E. Keogh and C. Merz, *UCI Repository of Machine Learning Databases*, http://www.ics.uci.edu/ mlearn/MLRepository.html, Univeristy of California, Irvine, Department of Information and Computer Sciences.

[2] T. Cover and P. Hart, *Nearest neighbor pattern classification*, IEEE Trans. Information Theory **13** (1) (1967), 21-27.

[3] L. Devroye, L. Györfi and G. Lugosi, *Probabilistic Theory of Pattern Recognition*, Springer New York 1996.

[4] R. Gnanadeskian, *Methods for Statistical Data Analysis of Multivariate Observations*, John Wiley & Sons London Second, New York 1997.

[5] R.A. Johnson and D.W. Wichern, *Applied Multivariate Statistical Analysis*, Prentice-Hall, New Jersey 1982.

[6] W.J. Krzanowski and F.H.C. Marriott, *Multivariate Analysis*, **1** Distributions, Ordination and Inference, Edward Arnold London 1994.

[7] W.J. Krzanowski and F.H.C. Marriott, *Multivariate Analysis*, **2** Classification, Covariance Structures and Repeated Measurements, London 1995.

[8] D.F. Morrison, *Multivariate statistical analysis*, PWN, Warszawa 1990.

[9] R. Paredes and E. Vidal, *A class-dependent weighted dissimilarity measure for nearest neighbor classification problems*, Pattern Recognition Letters **21** (2000) 1027–1036.

[10] G.A.F. Seber, *Multivariate Observations*, John Wiley & Sons, New York 1984.

Received 18 June 2004
Revised 4 August 2005