Pattern Anal Applic (2007) 10:55–68
DOI 10.1007/s10044-006-0050-x

THEORETICAL ADVANCES

Pairwise feature evaluation for constructing reduced representations

Artsiom Harol · Carmen Lai · Elżbieta Pękalska · Robert P. W. Duin

Received: 28 November 2005 / Accepted: 3 August 2006 / Published online: 19 October 2006

Abstract Feature selection methods are often used to determine a small set of informative features that guarantee good classification results. Such procedures usually consist of two components: a separability criterion and a selection strategy. The most basic choices for the latter are individual ranking, forward search and backward search. Many intermediate methods such as floating search are also available. The forward as well as backward selection may cause lossy evaluation of the criterion and/or overtraining of the final classifier in case of high-dimensional spaces and small sample size problems. Backward selection may also become computationally prohibitive. Individual ranking, on the other hand, suffers as it neglects dependencies between features. A new strategy based on a pairwise evaluation has recently been proposed by Bo and Jonassen (Genome Biol 3, 2002) and Pękalska et al. (International Conference on Computer Recognition Systems, Poland, pp 271–278, 2005). Since it considers interactions between features, but always restricted to two-dimensional spaces, it may circumvent the small sample size problem. In this paper, we evaluate this idea in a more general framework for the selection of features as well as prototypes. Our finding is that such a pairwise selection may improve over traditional procedures and we present some artificial and real-world examples to support this claim. Additionally, we have also discovered that the set of problems for which the pairwise selection may be effective is small.

Keywords Feature selection · Prototype selection · Pairwise feature evaluation · Pattern classification

1 Introduction

The construction of a proper vector space is essential in order to represent the data well and to design a successful statistical learning procedure. Concerning both computational efficiency and performance of a recognition system, one is usually interested in a space of low dimensionality. Since an initial space may be large, some reduction techniques are necessary for the optimization of an informative feature set, either by selecting or by merging features. An ideal technique is capable of reducing the dimensionality effectively, while preserving the separability between classes in the data. As some information is unavoidably lost in such a process, it is desirable to formulate a method that significantly reduces the dimensionality, but still preserves the separability. In this paper, we focus on selection approaches in the context of classification.

Feature selection methods rely on a quantitative criterion that measures their performance. This

A. Harol (✉) · C. Lai · E. Pękalska ·
R. P. W. Duin
Information and Communication Theory group,
Faculty of Electrical Engineering,
Mathematics and Computer Science,
Delft University of Technology, Delft, The Netherlands
e-mail: a.harol@ewi.tudelft.nl

C. Lai
e-mail: c.lai@ewi.tudelft.nl

R. P. W. Duin
e-mail: r.p.w.duin@ewi.tudelft.nl

E. Pękalska
School of Computer Science, University of Manchester,
Manchester, UK
e-mail: e.pekalska@ewi.tudelft.nl

criterion is used in some optimization process to determine a subset of informative features. Depending on how the suitability of features is judged, selection methods are divided into filters and wrappers [17, 19]. Filters evaluate the relevance of features based on a feature capacity to discriminate between classes. Wrappers employ a predetermined classification algorithm to judge the quality of a feature set. Advantages of filter and wrapper approaches are problem dependent. Filters rely on global data characteristics and are usually quite fast. Wrappers train a classifier appropriate for the given problem. As a result, they may find better features, but may also suffer more easily from overtraining.

Both approaches involve a combinatorial search over a constructed representation space of possible feature subsets. Usually, greedy procedures such as forward or backward eliminations are employed due to their simplicity and computational attractiveness. More complex procedures such as floating searches and genetic algorithms can also be applied [10, 14, 19, 20, 27], as well as other hybrid methods [8, 31].

Concerning the evaluation of a criterion, selection techniques are either univariate or multivariate. Univariate approaches are simple and fast. Multivariate approaches evaluate the relevance of features in a group, taking the interdependencies into account. When features are correlated, these techniques are able to construct good feature subsets, while univariate techniques may fail.

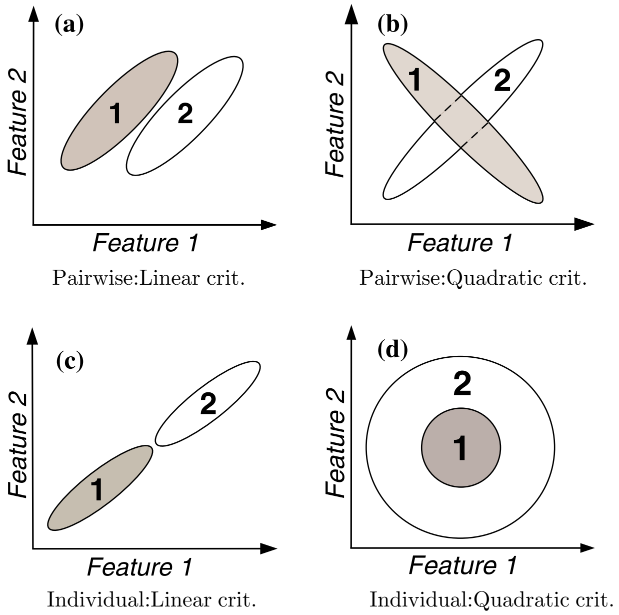
Unfortunately, there are two disadvantages of multivariate approaches. First, they evaluate features in a multi-dimensional space, not only demanding a considerable computational effort, but also resulting in a loss of accuracy in the case of a limited training set. Due to overfitting, feature subsets that do not ensure a good discrimination may still be judged as informative by a chosen criterion. The larger the number of selected features, the more pronounced this problem becomes. Second, large sets of features may have several groups of nested features that cannot be determined in a greedy forward selection. It was shown in [7, 27] that only the exhaustive search technique should be applied in order to reach the optimal subset of features. Although feature evaluation procedures involving the branch and bound algorithm for the optimization of a criterion may avoid the evaluation of all combinations, the most commonly used criterion functions do not satisfy the necessary conditions for this approach.

An attempt to preserve the advantages of univariate approaches without selecting highly correlated features has recently been proposed in [13, 32]. However, these heuristic algorithms still cannot find pairwise dependencies that might be present in the data. The main focus of these works is the computational issue. Therefore, before evaluating the correlations between features, initial univariate ranking is performed first, by which pairwise dependencies are missed.

As an alternative, the pairwise feature evaluation procedure was studied in [3] for the selection of genes in micro-array data and also by us in [24]. Since pairs of features are considered, second order dependencies are taken into account. On the other hand, since multi-dimensional spaces are now restricted to two-dimensional spaces, this method does not suffer from overfitting as other multivariate approaches do.

Figure 1 illustrates four types of feature subsets equally good for a classification problem when judged in pairs. The feature pairs (a) and (b) can only be found by a pairwise procedure based either on a linear criterion for the pair (a) or a quadratic criterion for the pair (b), while feature pairs (c) and (d) can be found by an individual ranking using either a linear criterion for the pair (c) or a quadratic criterion for the pair (d). Note that a single feature in the subplots (c) and (d) is sufficient for a good discrimination.

The problem of prototype selection can be seen similar to the problem of feature selection when prototypes are used to build representation spaces. Recent



Pairwise:Linear crit.

Pairwise:Quadratic crit.

Individual:Linear crit.

Individual:Quadratic crit.

Fig. 1 Examples of feature subsets which make equally good pairs for a classification problem. The relevance of pairs (**a**) and (**b**) can only be found by a pairwise procedure (with a linear and quadratic criterion, respectively). Feature pairs (**c**) and (**d**) can be found by an individual ranking (with a linear and quadratic criterion, respectively); it may not be necessary to find both features

research efforts [21, 23, 25, 26,] show that proximity representations, defined by a set of proximities computed to the given prototypes, are a good alternative to feature-based representations. Moreover, decision functions are not restricted to the nearest neighbor rule, as many other classifiers can be applied. A chosen m-element prototype set constructs a new m-dimensional representation space, in which each object is represented by its dissimilarities (proximities) to that set. Hence, every dimension is described by a dissimilarity to a particular prototype. As a result, one may follow a traditional feature-based approach and introduce a discrimination function there. For this reason we include some examples of prototype selection used to build proximity representations.

In this paper, the pairwise selection strategy is evaluated for both features and prototypes. In Sect. 2, basic feature selection methods are briefly described and the pairwise method is introduced. Since the idea seems very intuitive, we could easily construct an artificial example. On the other hand, we also tried to find real-world data examples. However, these examples were not easily found and we analyzed the causes. Our results are presented in Sects. 3 and 4. Section 5 discusses our findings.

2 Feature selection for classification

In a classification problem, feature selection techniques try to determine a small subset of features which are sufficient for a good discrimination. Usually, a type of a combinatorial search, in a forward manner (an incremental addition of features starting from a single one), a backward manner (an incremental removal of features starting from the entire set) or a floating manner is employed to find this feature subset. The optimization relies on a specified criterion (also used in the final classification), which is often related to a class separability and the way the relevance of a feature to be either added or removed is evaluated.

In a probabilistic framework, one assumes that real-world objects are represented as vectors $\mathbf{x}$ in a suitable vector space $\mathcal{X}$, e.g. $\mathcal{X} = \mathbb{R}^m$. The classification task relies on finding an unknown functional dependency ψ, a classifier, between $\mathbf{x}$ and the labels $y \in \mathcal{Y}$. Vectors $\mathbf{x}$ are assumed to be iid, drawn independently from a fixed, but unknown probability distribution $p(\mathbf{x})$. The function ψ is given as a fixed conditional density $p(y|\mathbf{x})$, which is also unknown. In practice, ψ is often parameterized by some $\boldsymbol{\alpha}$. It is found to be optimal according to some loss function Θ, measuring the discrepancy between the true and estimated values. The classification problem is then formulated as minimization of the true error $\mathcal{E}(\psi) = \int_{\mathcal{X}\times\mathcal{Y}} \Theta(y, \psi(\mathbf{x}, \boldsymbol{\alpha}))p(\mathbf{x}, y)\mathrm{d}\mathbf{x}\mathrm{d}y$, given a finite iid sample, i.e. the training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$.

Let $\boldsymbol{\tau} \in \{0, 1\}^m, m = \dim(\mathcal{X})$, denote a binary vector that will act as a feature selector. In a feature selection task one wants to find a transformation of the original data $\mathbf{x} \mapsto (\mathbf{x} * \boldsymbol{\tau})$, where $\mathbf{x} * \boldsymbol{\tau}$ denotes the Hadamard (element-wise) product of two vectors, and a set of parameters $\boldsymbol{\alpha}$ of the function ψ such that the functional

$$\Phi(\boldsymbol{\tau}, \boldsymbol{\alpha}) = \int_{\mathcal{X}\times\mathcal{Y}} \Theta(y, \psi((\mathbf{x} * \boldsymbol{\tau}), \boldsymbol{\alpha}))p(\mathbf{x}, y)\mathrm{d}\mathbf{x}\mathrm{d}y \tag{1}$$

is minimal. As a result, $\mathbf{x} * \boldsymbol{\tau}$ refers to a subset of feature values. Here, we assume that $||\boldsymbol{\tau}||_1 = \sum_i \tau_i = n$, where $n \lll m$. Since the joint probability $p(\mathbf{x},y) = p(\mathbf{x})$ $p(y|\mathbf{x})$ is unknown, one often minimizes the empirical error, given the finite training set:

$$\Phi_{\text{emp}}(\boldsymbol{\tau}, \boldsymbol{\alpha}) = \frac{1}{N}\sum_{i=1}^{N} \Theta(y_i, \psi((\mathbf{x}_i * \boldsymbol{\tau}), \boldsymbol{\alpha})) \tag{2}$$

or its regularized version $\Phi_{\text{reg}}(\boldsymbol{\tau}, \boldsymbol{\alpha}) = \Phi_{\text{emp}}(\boldsymbol{\tau}, \boldsymbol{\alpha}) + \gamma(\boldsymbol{\alpha})$, where γ is some penalty functional [28].

2.1 Selection methodologies

Three incremental wrapper-based selection methods are considered in this paper. These are individual, forward and pairwise strategies. We assume that an initial set F of m features $F = \{f_1, f_2, \ldots, f_m\}$ is given. The set F is a set of indices for the vector $\mathbf{x}$. Let $\tilde{F}, \tilde{F} \subset F$, denote a subset of selected features. Starting from an empty set, $\tilde{F} = \emptyset$, a single feature or a pair of features is chosen in each step according to the criterion based on the functional (2) and added to the set $\tilde{F}$. This step is repeated until $\tilde{F}$ consists of n predefined features.

In any case, one needs to go through the set F by fixing the values of $\boldsymbol{\tau}$ and approximating $\Phi(\boldsymbol{\tau}, \boldsymbol{\alpha})$ by

$$\Phi_{\text{emp}}(\boldsymbol{\tau}^*, \boldsymbol{\alpha}^*) = \operatorname*{argmin}_{\boldsymbol{\tau}, \boldsymbol{\alpha}} \Phi_{\text{emp}}(\boldsymbol{\tau}, \boldsymbol{\alpha}), \tag{3}$$

where $\boldsymbol{\tau}^*$ denotes a subset F_{eval} of the initial features to be evaluated by some criterion function and $\boldsymbol{\alpha}^*$ represents the parameters of the given ψ. For every particular subset of features F_{eval} we will use the following criterion function:

$$J(F_{\text{eval}}) = \exp(-\Phi_{\text{emp}}(\boldsymbol{\tau}^*, \boldsymbol{\alpha}^*)). \tag{4}$$

It returns a particular value of goodness for currently evaluated features as specified by τ^*. The better the classification (the smaller the value of the loss function Φ_{emp}), the larger the value of J.

2.1.1 *Individual (univariate) selection*

In this approach, the informativeness of each feature is evaluated individually according to the criterion J. In each step, a single best feature is chosen. This can formally be written as:

$$\begin{aligned}&\tilde{F} := \tilde{F} \cup \{f\}, \quad \text{where } f = \operatorname*{argmax}_{f_i \in F} J(f_i), \tilde{F} \cap \{f\} = \emptyset\\&F := F \setminus \{f\}\end{aligned} \tag{5}$$

In this procedure features are ranked from the most to the least relevant according to the values of J. In the end, the most indicative features can be finally selected.

2.1.2 *Forward selection*

Forward feature selection starts with the single most informative feature and continues to add next most informative features in a greedy fashion. The relevance of a feature is evaluated in the context of the already selected features by determining the criterion J in a feature space of growing dimensionality. Hence, $F_{\text{eval}} = \tilde{F} \cup \{f\}_i$ for some feature f_i. This step can be summarized as follows:

$$\begin{aligned}&\tilde{F} := \tilde{F} \cup \{f\}, \ \text{where}\, f = \operatorname*{argmax}_{f_i \in F} J(\tilde{F} \cup \{f\}_i), \ \ \tilde{F} \cap \{f\} = \emptyset.\\&F := F \setminus \{f\}\end{aligned} \tag{6}$$

2.1.3 *Pairwise selection*

The relevance of features is judged by evaluating all different pairs of features. Hence, $F_{\text{eval}} = \{f_i \cup f_i\}$. In each step, the best unselected feature pair is detected and added to the final subset $\tilde{F}$.

$$\begin{aligned}&\tilde{F} := \tilde{F} \cup \{f \cup f'\},\\&\text{where } \{f \cup f'\} = \operatorname*{argmax}_{\{f_i, f_j\} \in F} J(\{f_i \cup f_j\}), \ \ {}^{\sim}\tilde{F} \cap \{f \cup f'\} = \emptyset.\\&F := F \setminus \{f \cup f'\}\end{aligned} \tag{7}$$

Other variants of the pairwise approach are also possible. For example, one may first rank features according to the pairwise procedure as described above, and then add a pair to the subset $\tilde{F}$ such that not only this pair is good, but also joint performances with the already selected features are maximum. To speed up the selection process [3], one may also rank features on the basis of a univariate criterion. This serves as an order list to select the first feature in a pair. The other feature is added such that the joint criterion is maximized. In this paper, we will use the variant summarized in Eq. 7.

2.2 Class discrimination performance

To select our features, one needs to find the minimum of the functional (2), indicating how well a single feature or a pair of features contributes to the separation of the classes. Assume c classes, $\omega_1,\ldots, \omega_c$, whose labels were before encoded by y. Here, we will restrict ourselves to the linear (NLC) and quadratic (NQC) classifiers assuming normal distributions of the classes:

$$p(\mathbf{x}^*|\omega_i) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_i|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}^* - \boldsymbol{\mu}_i)^{\mathrm{T}} \Sigma_i^{-1} (\mathbf{x}^* - \boldsymbol{\mu}_i)\right\}, \tag{8}$$

where $\mathbf{x}^* = \mathbf{x} * \boldsymbol{\tau}^*$ is considered as an n-dimensional vector of feature values as selected by τ^* (the remaining zero values are neglected), while $\boldsymbol{\mu}_i$ and Σ_i are the mean vector and the covariance matrix for the class ω_i [10, 12, 16]. Here, $|\Sigma_i|$ denotes the determinant of Σ_i.

From the Bayes rule

$$\psi_i(\mathbf{x}^*) = p(\omega_i|\mathbf{x}^*) = \frac{p(\mathbf{x}^*|\omega_i)p(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x}^*|\omega_j)p(\omega_j)}, \tag{9}$$

and the fact that the above denominator is independent of a particular ω_i, we have

$$\psi_i(\mathbf{x}^*) \sim \tilde{\psi}_i(\mathbf{x}^*) = p(\mathbf{x}^*|\omega_i)p(\omega_i). \tag{10}$$

The minimum of $\Theta(\cdot,\cdot)$, chosen as the indicator function, is achieved in the classification when the vector $\mathbf{x}$ is assigned to the class ω_i which is more probable than any other ω_j, i.e.

$$\tilde{\psi}_i(\mathbf{x}^*) = \ln p(\mathbf{x}^*|\omega_i) + \ln p(\omega_i) > \tilde{\psi}_j(\mathbf{x}^*), \quad \forall j \neq i \tag{11}$$

Taking into account (8), we get the NQC

$$\tilde{\psi}_i(\mathbf{x}^*) = \mathbf{x}^{*\mathrm{T}} \mathbf{W}_i \mathbf{x}^* + \mathbf{w}_i^T \mathbf{x}^* + w_{i0} \tag{12}$$

where

$$\mathbf{W}_i = \frac{1}{2}\Sigma_i^{-1}$$
$$\mathbf{w}_i = \Sigma_i^{-1}\boldsymbol{\mu}_i$$
$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^T\Sigma_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\Sigma_i| + \ln p(\omega_i)$$

In case when the class mean vectors cluster, i.e. $(|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j| \to 0,\ \forall i \neq j)$, for example as depicted in Fig. 1d, the quadratic function (12) still takes into account the discriminatory information that is present in differences between the class covariance matrices. Thus, subsets of features that are not well separated by the use of mean vectors, may give a better accuracy when their correlations are judged.

In our experiments we also assumed equal covariance matrices $\Sigma_\mathrm{i} = \Sigma,\ \forall i$. In such a case, the quadratic function (12) becomes linear, which is the NLC:

$$\tilde{\psi}_i(\mathbf{x}^*) \sim \tilde{\psi}_i(\mathbf{x}^*) = \mathbf{w}_i^T\mathbf{x}^* + w_{i0} \tag{13}$$

where

$$\mathbf{w}_i = \Sigma^{-1}\boldsymbol{\mu}_i$$
$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^T\Sigma^{-1}\boldsymbol{\mu}_i + \ln p(\omega_i)$$

2.3 Regularization of singular covariances

In case the estimated covariance matrices are singular (which may happen in high-dimensional spaces), they can be regularized to ensure that the inverse operation is possible. This is done by using a regularized version instead, $\Sigma_{\lambda,\theta} = (1-\lambda-\theta)\cdot\Sigma + \lambda\cdot\mathrm{diag}(\mathrm{diag}(\Sigma)) + \frac{\theta}{n}\cdot\mathrm{trace}(\Sigma)\cdot I$, where I is the identity matrix. Note that $\lambda \in [0,1]$ and $\theta \in [0,1]$ are related to variances, so they can be determined more easily. θ takes care that none of the variances becomes zero. In practice, these parameters are set to 0.01 or less.

3 Feature selection experiments

The potential benefits and limitations of the pairwise feature selection are illustrated by several artificial and real-world examples. First, a brief description of the data sets is given, then the experimental results are presented.

3.1 Data sets

3.1.1 Artificial example

We have generated artificial data that has a set of informative features among a number of noisy ones. Two features are assumed to be informative if considered in a pair; see Fig. 1a. The correlation between features is chosen such that an individual selection strategy is not capable of finding this meaningful subset of features.

Assume that s samples and m features are given such that only q features, generated in correlated pairs, are informative. The samples for each correlated feature pair are drawn from a Gaussian distribution with the following class means $\mu_1 = [0\ 0]^T$ and $\mu_2 = \frac{\sqrt{2}}{2}[r\ 0]^T$ for some parameter $r >\ 0$. The covariance matrix, identical for both classes, is given as $\Sigma_1 = \Sigma_2 = \begin{bmatrix} v+1 & v-1 \\ v-1 & v+1 \end{bmatrix}$ for some value of v. The remaining $(m - q)$ features are uninformative, i.e. the two classes are drawn from a spherical Gaussian distribution $\mathcal{N}(\mathbf{0}, \frac{v}{\sqrt{2}}I)$, where I is an identity matrix. Here, we set $k = 100$, $m = 300$ and $q = 20$. In order to have a class overlap, we set $r = 3$ and $v = \sqrt{40}$. Since we want to simulate a small sample size problem, we chose $k = 100$ samples for the training set, while the test set $(s - k)$ consists of 10,000 examples.

3.1.2 Waveform

The Waveform data [4] is a three-class problem. It is based on a sampling of triangle shaped waves and has 21 features. There are 5,000 objects in total, approximately equally distributed over three classes. In order to simulate a small sample size problem, we randomly selected 35 samples for the training set and used the remaining samples as an independent test set.

3.1.3 Colon

The Colon data set [1] is a microarray gene expression data set measured on high-density oligonucleotide Affymetrix arrays. The data set is composed of 40 normal (healthy) samples and 22 tumor samples in a 1,908-dimensional feature space described by genes.

3.1.4 Texture

To create this set we took two images, scanned with 150 dpi, from the Brodatz album [5]. These are textures of the reptile skin and cork. Figure 2 presents 1.7″ × 1.7″ parts of the two images.

The linear resolution was reduced by a factor of 8 to obtain texture elements of a manageable size. The resulting images had a size of 170 × 136 pixels. They were normalized to have equal means and contrasts. As a result, the distributions of pixel intensities have

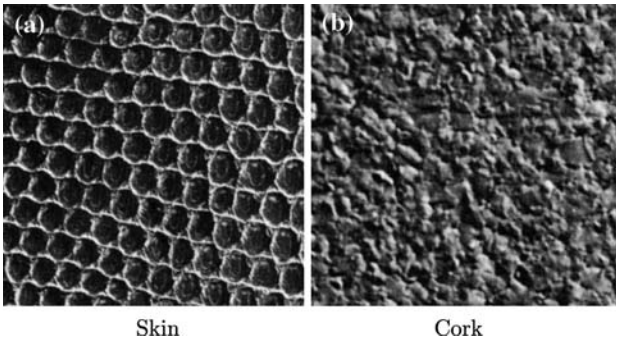


Fig. 2 Texture data set. Examples of images of the reptile skin and cork

equal means and standard deviations for these two images. Figure 3 shows enlarged parts of these images of the size 32 × 32 pixels. Thousand windows of 8 × 8 pixels are selected from each image at random, resulting in a two-class dataset of 2,000 objects and 64 features (pixel intensities). As these features have the same means and identical standard deviations (over the means as well as over the classes), there is no linear separability between these two classes, for none of the feature nor for any combination of them.

3.2 Results

3.2.1 *Experimental setup*

Individual, forward and pairwise feature selection techniques are evaluated here [11]. The NLC is used as an output performance measure in the criterion J, for each of the selection techniques. The NLC is trained on a training set with a growing number of features and tested on an independent test set. As a result, the averaged classification error can be estimated as a function of the number of chosen features. This procedure is repeated 50 times for the Artificial, Waveform and Texture data sets with randomly generated training and tests sets. The final results are averaged.

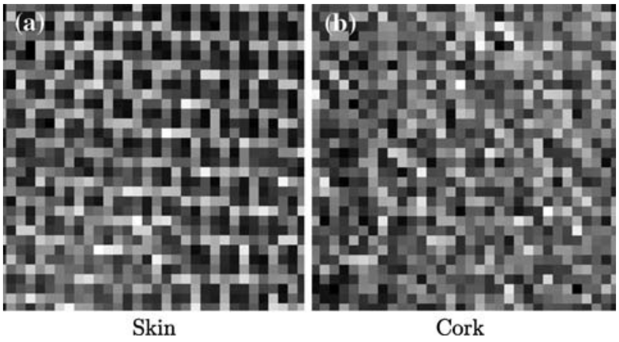


Fig. 3 Normalized Texture data set. The distributions of the pixel intensities have equal means and standard deviations for the two images

For the Artificial, Waveform and Texture data sets, a significantly large independent test set can be generated. However, the Colon data contains just a small number of objects in a very high-dimensional vector space. So, there is no large independent test set available. Therefore, we performed tenfold cross-validation to estimate the classification error as suggested by Kohavi [18].

3.2.2 *Results*

Figure 4 shows the behavior of three different feature selection methods for the Artificial data as judged by the average classification error. Although the standard deviations are omitted for visual clarity, the differences between the methods are statistically significant. The pairwise selection procedure performs significantly better than the individual selection. This holds since pairs of features are constructed to yield strong correlations, that may only be captured in a pairwise manner. On the other hand, a forward selection does not achieve a good performance since features are added one by one. By missing the notion of pairwise dependencies, the forward procedure cannot find the existing correlations between features. Due to noise, it starts even with different features than the pairwise procedure does.

The averaged classification error for the Waveform data set is depicted in Fig. 5. This is one more artificial example, where the pairwise selection behaves much better than the forward selection as well as the individual ranking. From beginning the pairwise approach shows a significant improvement, while other methods fail. For a larger feature size, the forward procedure cannot estimate the criterion values appropriately, leading to overtraining, as a result. Nevertheless, the pairwise procedure still shows a continuous improvement up to eight features. A further extension of the feature size also leads to overtraining.

Figure 6 illustrates an averaged cross-validation error of the Colon data set. Individual ranking performs well and the results are not significantly improved by more complex selection techniques. The experiment is consistent with the work of Bo et al. [3], where the performance of the individual search was comparable to the result of the pairwise search applied to the same data set evaluated in a leave-one-out cross-validation approach. This is again the effect of the curse of dimensionality due to the small number of samples as compared to a huge number of features. As just a

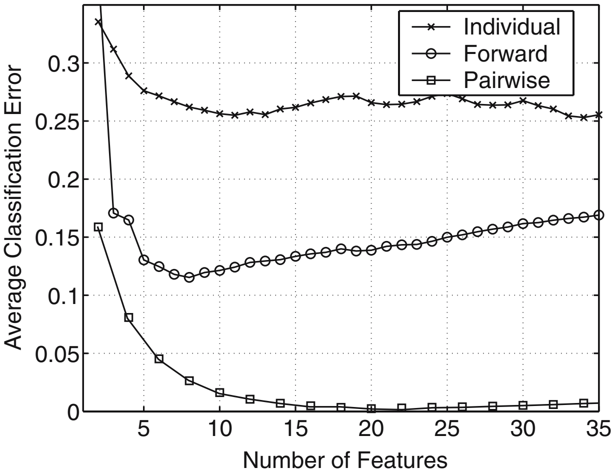


Fig. 4 Two-class Artificial data. Average classification error of the NQC as a function of $|\tilde{F}|$ found by three selection procedures. The estimation is based on 50 repetitions

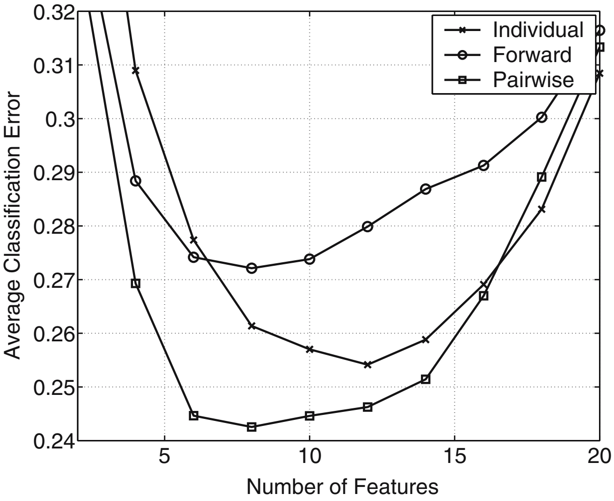


Fig. 5 Three-class Waveform data. Average classification error of the NLC as a function of $|\tilde{F}|$ found by three selection procedures. The estimation is based on 50 repetitions

single run of the tenfold cross-validation was used due to computation time constraints, error differences may not be significant.

Figure 7 depicts an averaged classification error of the three feature selection procedures for the Texture dataset. This is one more example of a real-world dataset where one may profit from the use of the pairwise selection procedure. As explained in the dataset description quadratic discrimination is needed here. Feature pairs are related as shown in Fig. 1b in a symbolic way. The classification error for the pairwise procedure is significantly lower than for other methods and shows a continuous improvement up to 50 features.

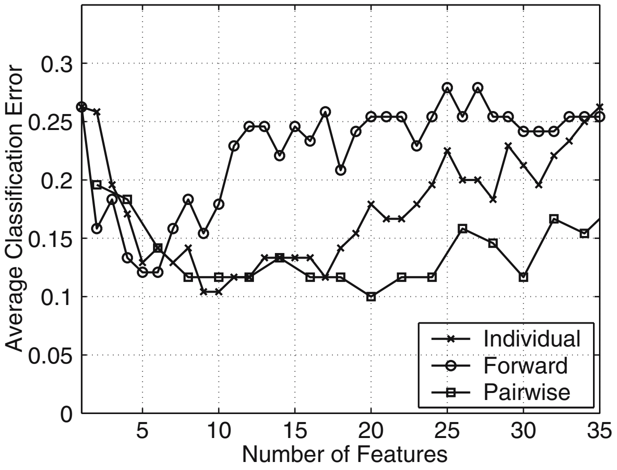


Fig. 6 Two-class Colon data. Tenfold cross-validation error of the NLC as a function of $|\tilde{F}|$ found by the three selection procedures

The differences in the performance of classifiers for various feature subsets are significant for all the data (as the standard deviations are very small), except for the Colon case.

4 Prototype selection experiments

Dissimilarity (or proximity) representations rely on pairwise object comparisons and are an alternative to feature-based descriptions. They are especially advantageous when discriminative features are difficult to obtain or when objects contain an inherent, identifiable structure such that suitable, e.g. edit-type, distances can be used for their comparisons. Such representations are universal, since all types of information, statistical, structural, hierarchical, relational, logical, or heterogeneous can be encoded by various proximity measures, and combined, if necessary [23]. Moreover, there already exists a plethora of practically designed, both metric and non-metric, measures used for all type of matching purposes; see e.g. [2, 6, 9, 23, 29].

More precisely, assume a representation set R of n prototypes, $R = \{p_1, p_2, ..., p_n\}$ and a dissimilarity measure d, computed or derived from the objects directly, or their initial representations. d has to be nonnegative and obey the reflexivity condition, $d(x,x) = 0$, but it may be non-metric. An object x is represented as a vector of dissimilarities computed between x and the prototypes from R, i.e. $D(x,R) = [d(x,p_1), d(x,p_2), ..., d(x,p_n)]^T$. Given a set T of N objects, such a representation becomes an $N \times n$ dissimilarity matrix $D(T,R)$.

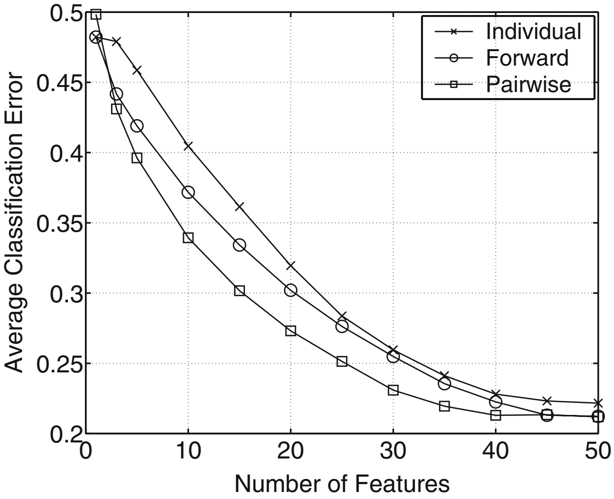


Fig. 7 Two-class Texture data. Average classification error of the NQC as a function of $|\tilde{F}|$ found by three selection procedures. The estimation is based on 50 repetitions

This dissimilarity representation $D(T,R)$ is addressed as a data-dependent mapping $D(\cdot, R): X \to \mathbb{R}^n$ from an initial representation X to the so-called dissimilarity space [23, 25], equipped with the traditional inner product and the Euclidean norm. In such a space, each dimension denotes a dissimilarity to a given prototype $p_i \in R$, i.e. $D(\cdot,p_i)$. Since dissimilarities are nonnegative, all the data examples are projected as vectors to a nonnegative orthotope of the dissimilarity space. In practice, this means that any traditional classifier used in vector spaces can be applied here as well. More details can be found in [25, 23].

Given a complete representation $D(T,T)$, our task now is to select a small set R out of T to guarantee a good tradeoff between the recognition accuracy and the computational complexity, when classifiers are built on $D(T,R)$. As such, the feature selection techniques discussed in the previous section become now prototype selection methods.

4.1 Data sets

Four dissimilarity data sets are used in our study. They are described below.

4.1.1 NIST38

The NIST38 digits data set [30] describes a set of scanned digits, originally provided as 128 × 128 binary images. There are ten classes in total, each represented by 200 examples. The images are first smoothed with a Gaussian kernel with $\sigma = 8$ pixels and then the pixel-based Euclidean distances between such blurred images are derived [25, 26]. Smoothing is done to make the resulting distance measure somewhat robust (invariant) against tilting or shifting of the single digits. In our experiments only the digits 3 and 8 are used. Each class is represented by 500 examples.

4.1.2 Digit38

The data describe the NIST digits [30]. First a similarity measure, based on deformable template matching as defined in [15], is derived. Let $S = (s_{ij})$ denote the similarities. The off-diagonal symmetric dissimilarities $D = (d_{ij})$ are computed as: $d_{ij} = (s_{ii} + s_{jj} - s_{ij} - s_{ji})^{1/2}$ for $i \neq j$, since the data are slightly asymmetric. The resulting measure is significantly non-metric. Here, only the digits 3 and 8 are used, represented by 200 examples per class.

4.1.3 Polygon

The Polygon data consists of randomly generated polygons: convex quadrilaterals (four-sided polygons) and both convex and non-convex heptagons (seven-sided polygons) [23, 26]. The polygons are first scaled appropriately and then the non-metric modified Hausdorff distances are derived. The modified Hausdorff distance is defined between two sets (here, polygon corners) A and B as $d_{\mathrm{MH}}(A,B) = \max\ \{d^{\triangleright}_{\mathrm{avr}}(A,B), d^{\triangleright}_{\mathrm{avr}}(B,A)\}$, where $d^{\triangleright}_{\mathrm{avr}}(A,B) = \frac{1}{|A|}\sum_{a\in A} \min_{b\in B} d(a,b)$ is a directed distance and $d(a,b)$ is the Euclidean metric [9]. Our investigations rely on 1,000 examples, equally distributed over two classes.

4.1.4 RoadSign

The data set consists of gray level images of circular road signs scaled to 32 × 32 pixel raster. There are 300 road sign images (highly multi-modal) and 300 non-road sign images acquired under general illumination [22]. Some image examples are presented in Fig. 12.

The latter images are identified by a sign detector using a circular template based on local edge orientations. Since a circular template was used to detect the boards, this a priori knowledge was used to remove the pixels in the background. The resulting data set contains 793 of original 1,024 dimensions (pixels). Normalized cross-correlation, considered as similarity, is computed between the images. Let s_{ij} denote the similarities. Then, the final dissimilarities are derived as $d_{ij} = (1 - s_{ij})^{1/2}$.

4.2 Results

4.2.1 *Experimental setup*

In our prototype selection experiments, all dissimilarity data are linearly scaled to [0,1]. Such a scaling does not affect the NLC and the NQC, as they are scaling independent. We also assumed uniform prior probabilities. Each data set is randomly split into a training set T and a test set S. In all cases, T consists of 50 examples per class. This is done in order to investigate small sample size problems. The remaining test sets have the following cardinalities: 900, 300, 900 and 500 for the NIST38, Digit38, Polygon and RoadSign data, respectively. An incrementally growing prototype set R is selected out of T by inspecting the dissimilarity matrix $D(T,T)$. Four procedures are used for the selection of a prototype set. These are random selection, individual ranking, forward search and the pairwise strategy. The prototypes R are chosen as features $\tilde{F}$ in the corresponding dissimilarity spaces. The classification accuracy of the NQC is used to define the separability criterion J, since the final classifier is chosen to be quadratic, as well.

The final NQC is regularized here, since in small sample size problems the covariance matrices are nearly singular. The regularization parameters are fixed as $\lambda = 0.001$ and $\theta = 0.001$. As a result, the final performance of the RNQC (regularized NQC) may be somewhat worse than of the NQC when regularization is unnecessary for a very small number of prototypes. Having found a prototype set R, the NQC is trained on $D(T,R)$ and tested on $D(S,R)$. This is repeated 30 times and the average classification error is plotted as a function of a number of prototypes.

In two cases, for the Digit38 and Polygon data, we also include wrappers based on the NLC. These are clear examples where the linear classifier performs better than the quadratic one.

4.2.2 *Results*

Similar phenomena as in the previous section can be observed in Figs. 8, 9, 10, 11, 12, 13, 14. Again the forward selection is overtrained for a large number of prototypes. The pairwise approach selects prototypes p_i and p_j which are characterized by correlated vectors of distances $D(\cdot, p_i)$ and $D(\cdot,p_j)$ in a two-dimensional space such that the (linear or quadratic) separability is maximum. We can observe this phenomenon in Fig. 15, in which the best two prototypes (determined by all selection methods) construct two-dimensional dissimilarity spaces. The feature pair defined by the pairwise selection clearly discriminates between two classes, which is less possible in other approaches. Concerning the selection process, the pairwise strategy does not suffer from overtraining. However, it may still be sensitive to the performance of a classifier, which is ultimately constructed in a high-dimensional space.

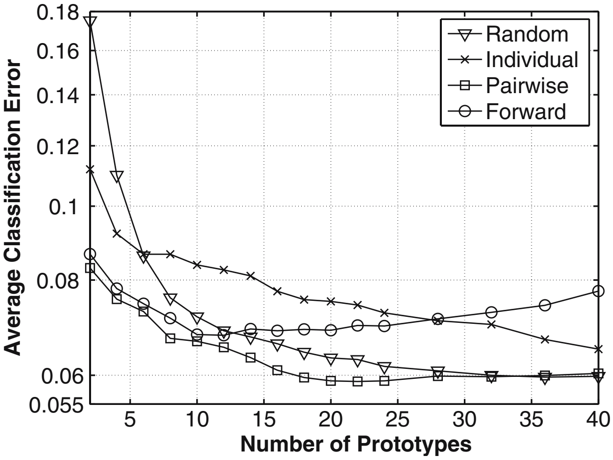


Fig. 8 NIST38 data. Average classification errors of the RNQC as a function of |R| found by the four selection procedures. The separability criterion J relies on the classification performance of the NQC. The standard deviations of the average errors are less than 0.005, except for the random selection of two features, and they are less than 0.0036 on average. The estimation is based on 30 repetitions. The y-axis has a logarithmic scale

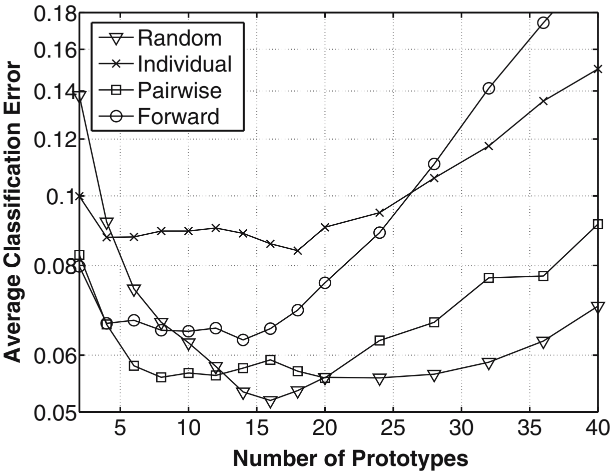


Fig. 9 Digit38 data. Average classification errors of the RNQC as a function of |R| found by the four selection procedures. The separability criterion J relies on the classification performance of the NQC. The standard deviations of the average errors are less than 0.0095, except for the forward selection of 40 features, and they are less than 0.005 on average. The estimation is based on 30 repetitions. The y-axis has a logarithmic scale

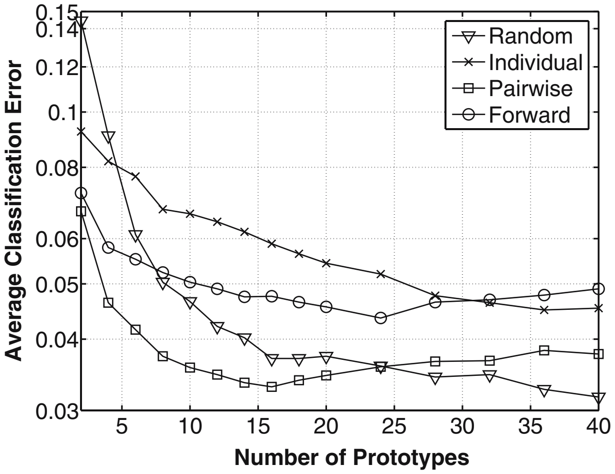


Fig. 10 Digit38 data. Average classification errors of the NLC as a function of $|R|$ found by the four selection procedures. The separability criterion J relies on the classification performance of the NLC. The standard deviations of the average errors are less than 0.0045, except for the random selection for two features, and they are less than 0.0035 on average. The estimation is based on 30 repetitions. The y-axis has a logarithmic scale

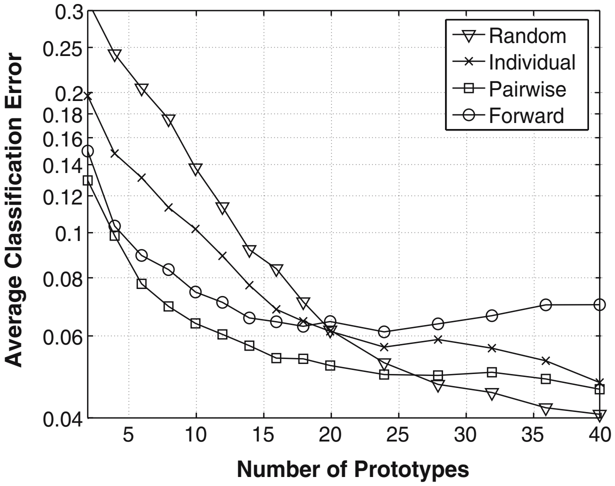


Fig. 11 Polygon data. Average classification errors of the RNQC as a function of $|R|$ found by the four selection procedures. The separability criterion J relies on the classification performance of the NQC. The standard deviations of the average errors are less than 0.01, except for the random selection for two features, and they are less than 0.006 on average. The estimation is based on 30 repetitions. The y-axis has a logarithmic scale

As observed in the plots, the univariate procedure often fails, especially for a small number of prototypes. This is simply due to its inability to find informative features in data, as they cannot be captured by just having a bulk of good single features, even if they are relevant. The random selection is often bad for small prototype sets, however, its potential grows with a growing number of prototypes. It may outperform the other selection approaches provided that the prototype set is sufficiently large [21, 23]. Having a sufficiently large size, randomly chosen prototypes tend to describe data characteristic well. As they are likely less correlated than the sets chosen by systematic selection procedures, they may also suffer little from overtraining in case of large sets. This effect can be clearly observed for the Digit38 data, Figs. 9 and 10, and for the Polygon data, Fig. 11.

In all presented cases the pairwise selection performs better than the individual ranking and better or similar than the forward selection. The differences are statistically significant for individual selection and small prototype sets, as well as, for forward selection and large prototype sets, as can be judged from the standard deviations of the averaged errors reported in figures.

An interesting example is the RoadSign data, for which the pairwise and forward selection strategies yield similar results up to 16 prototypes, and then the forward selection makes the NQC overtrain. On the other hand, also pairwise and random strategies are not significantly different for ten or more prototypes, yielding ultimately the same performance of the NQC. A possible explanation of this fact is that in this real-world example, the information about class separability is spread in different ways over many dissimilarity vectors $D(\cdot,p_i)$ such that in pairs they provide a good separability. Basically, many prototypes do have a discrimination power which is complementary to other prototypes, so when added to the current prototype set, they still contribute.

5 Discussion and conclusions

The purpose of our study is to evaluate the idea of pairwise feature selection in the classification context. It may be applied to the selection of prototypes as well. In general, it is beneficial to perform such a selection not by evaluating individual features, but their combinations. The reason is that good combinations may exist for features that are individually not informative. The results of this phenomenon can be observed in all our experiments. There are two reasons for restricting the search of a good feature set to pairwise evaluations.

First, if feature combinations have to be evaluated in a multi-dimensional setting, with a dimensionality eventually as large as the final feature set (in forward selection procedures), or even much higher (in backward selection procedures), then the criterion has to be

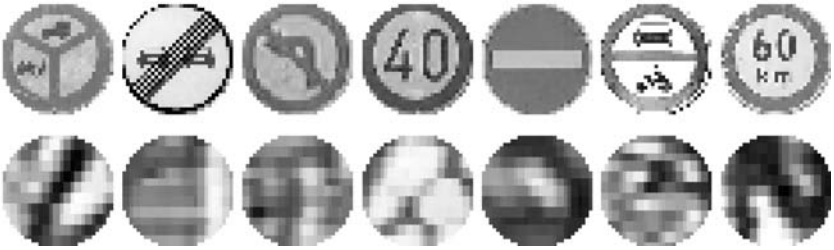


Fig. 12 RoadSigns data set. Examples of images of the road signs and non-signs

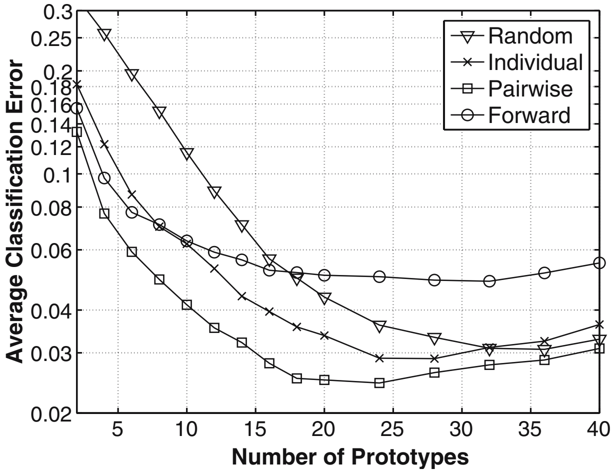


Fig. 13 Polygon data. Average classification errors of the NLC as a function of |*R*| found by the four selection procedures. The separability criterion *J* relies on the classification performance of the NLC. The standard deviations of the average errors are less than 0.008, except for the random selection for two and four features, and they are less than 0.005 on average. The estimation is based on 30 repetitions. The *y*-axis has a logarithmic scale

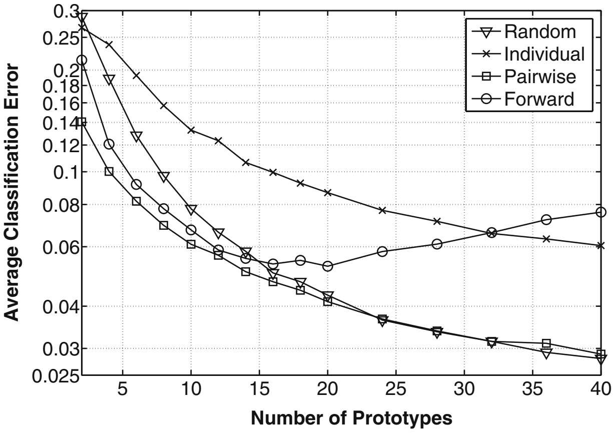


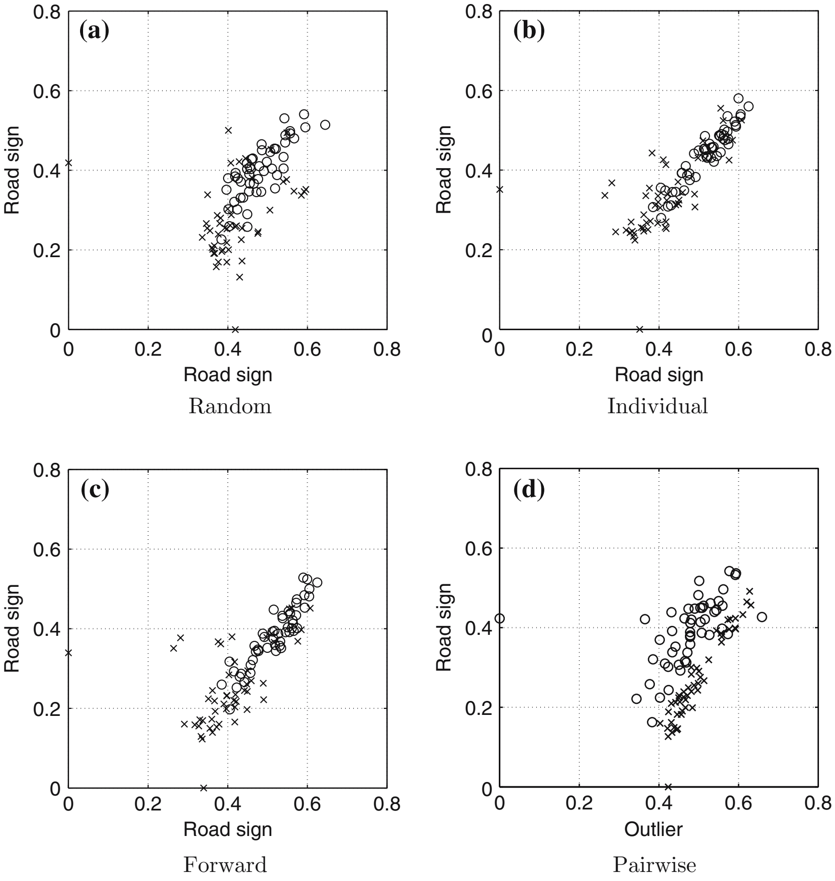
Fig. 14 RoadSign data. Average classification errors of the RNQC as a function of |*R*| found by the four selection procedures. The separability criterion *J* relies on is the classification performance of the NLC. The standard deviations of the average errors are less than 0.0085 except for the random selection for two features and they are less than 0.007 on average. The estimation is based on 30 repetitions. The *y*-axis has a logarithmic scale

computed in a high-dimensional space. This may result in overtraining and a selection of suboptimal feature sets due to noise. Pairwise searches evaluate the criterion just in two dimensions, resulting in more reliable selections. It is clearly visible in Figs. 4, 5, 6, 7, 8, 9, 10, 11, 13 and 14 that in the case of high-dimensional representation spaces, the forward selection performs worse than the pairwise selection. One of the reasons may be the selection of bad features due to overtraining of a high-dimensional criterion. Additional explanation is also given below. Still, all selection procedures may suffer from classifier overtraining, e.g. in Figs. 4, 5, 6, 8, 9, 10, 11, 13 and 14.

The second reason for a good performance of the pairwise selection is its ability to select sets of unrelated good pairs for which combinations between pairs are still bad. The Artificial example in Fig. 4 shows this clearly, as it is constructed to support this claim. The Texture problem, Fig. 7, shows that this phenomenon may be observed in real world data for the quadratic classifier. Also in other examples this may be the case where the pairwise procedures perform better than the best combination, found by the forward selection. For higher dimensionalities, however, also the above discussed explanation may hold.

We have to admit that, in general, it appeared to be very difficult to find good examples strongly in favor of pairwise selection. The above arguments definitely hold, since the corresponding artificial problems may be generated [7]. However, real world examples that behave as such appeared to be rare. In particular, we doubt whether there are many real world examples with unrelated nested feature sets in which such a single one will be found by the forward selection procedure. This procedure starts with the best overall individual features and then searches for good combinations with this one and with the following ones that are selected. Other sets of well performing features will not be found only if they all show just a marginally increased performance in combination with the initially selected feature set. That this happens appears to be unlikely. Note that the existence of such nested set of features was one of the arguments for constructing the floating search procedures [27].

Fig. 15 RoadSign data. Dissimilarity spaces defined by the first two prototypes as chosen by four selection methods based on the NQC. The first dimension is a dissimilarity to the first chosen prototype, while the second dimension is a dissimilarity to the second prototype. That is why, two single object lie exactly on the axes as they have zero distances to themselves. Road signs are marked by 'x', while outliers (no road signs) are marked by 'o'. *Axis labels* indicate to which class the prototypes belong



In case of very large feature sets (thousands of features), a complete pairwise selection procedure is prohibitive, as it demands the evaluation of all feature pairs. Bo et al. [3] experimented with a full search of all feature pairs between just the individually best performing features and all others. Such pairs, however, will be found anyway, as a good performing feature has at least a similar performance with any other feature. A better way may be based on an individual selection in combination with an uncorrelation criterion as proposed by [13].

Finally, we conclude that the pairwise selection of features and prototypes is a good, additional tool in the feature selection toolbox. It is especially worthwhile for problems with large, but not very large sets of initial features. In a limited set of problems it may show a better performance than procedures based on the forward selection.

Acknowledgments This work is supported by the Dutch Organization for Scientific Research (NWO) and the Dutch Cancer Institute (NKI). The authors thank Prof. Anil Jain and Dr. Douglas Zongker for providing the Digit dissimilarity data and Dr. Pavel Paclík for providing the RoadSign dissimilarity data.

References

1. Alon U, Barkai N, Notterman D, Gish K, Ybarra S, Mack D, Levine A (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 96(12):6745–6750
2. Bennett CH, Gacs P, Li M, Vitányi PMB, Zurek W (1998) Information distance. IEEE Trans Inf Theory IT-44(4):1407–1423
3. Bo T, Jonassen I (2002) New feature subset selection procedures for classification of expression profiles. Genome Biol 3
4. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth, California
5. Brodatz P (1996) Textures: a photographic album for artists and designers. Dover, New York
6. Bunke H, Sanfeliu A (1990) Syntactic and structural pattern recognition theory and applications. World Scientific
7. Cover TM, van Campenhout JM (1977) On the possible ordering in the measurement selection problem. IEEE Trans Syst Man Cybern SMC-7(9):657–661
8. Das S (2001) Filters, wrappers and a boosting-based hybrid for feature selection. In: International Conference on Machine Learning, pp 74–81
9. Dubuisson MP, Jain AK (1994) Modified Hausdorff distance for object matching. In: International Conference on Pattern Recognition, vol 1, pp 566–568
10. Duda RO, Hart PE, Stork DG (2001) Pattern classification, 2nd edn. Wiley, New York

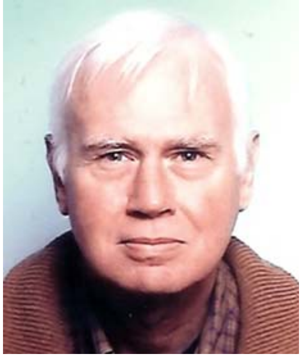
11. Duin RPW, Juszczak P, de Ridder D, Paclík P, Pękalska E, Tax DMJ (2004) PR-Tools, Pattern Recognition Tools. http://www.prtools.org
12. Fukunaga K (1990) Introduction to statistical pattern recognition, 2nd edn. Academic, INC
13. Hall M (2000) Correlation-based feature selection for machine learning. Ph.D Thesis, University of Waikato
14. Jain AK, Zongker D (1997) Feature selection—evaluation, application, and small sample performance. IEEE Trans Pattern Anal Mach Intell 19(2):153–158
15. Jain AK, Zongker D (1997) Representation and recognition of handwritten digits using deformable templates. IEEE Trans Pattern Anal Mach Intell 19(12):1386–1391
16. Jain AK, Duin RPW, Mao J (2000) Statistical pattern recognition: a review. IEEE Trans Pattern Anal Mach Intell 22:4–37
17. John GH, Kohavi R, Pfleger P (1994) Irrelevant features and the subset selection problem. In: Mahine learning: Proceedings of the Ninth International Conference. Morgan Kaufmann
18. Kohavi R (1995) The power of decision tables. In: Proceedings of the Eighth European Conference on Machine Learning ECML95, Lecture Notes in Artificial Intelligence, 914, pp 174–189. Springer, Berlin Heidelberg New York
19. Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell 97:273–324
20. Li L, Weinberg CR, Darden TA, Pedersen LG (2001) Gene selection for sample classification based on gene expression data: study of sensitivity to choice of paramethe GA/KNN method. Bioinformatics 17:1131–1142
21. Lozano M, Sotoca JM, Sanchez JS, Pla F, Pękalska E, Duin RPW (2006) Experimental study on prototype optimisation algorithms for dissimilarity based classifiers. Pattern Recognit 39(10):1827–1838
22. Paclík P, Novovičová J, Somol P, Pudil P (2000) Road sign classification using Laplace Kernel classifier. Pattern Recognit Lett 21(13–14):1165–1173
23. Pękalska E, Duin RPW (2005) The dissimilarity representation for pattern recognition. Foundations and applications. World Scientific, Singapore
24. Pękalska E, Harol A, Lai C, Duin RPW (2005) Pairwise selection of features and prototypes. In: International Conference on Computer Recognition Systems, Poland, pp 271–278
25. Pękalska E, Duin RPW, Paclík P (2002) A generalized Kernel approach to dissimilarity based classification. J Mach Learn Res 2(2):175–211
26. Pękalska E, Duin RPW, Paclík P (2006) Prototype selection for dissimilarity-based classifiers. Pattern Recognit 39(2):189–208
27. Pudil P, Novovicova J, Kittler J (1994) Floating search methods in feature selection. Pattern Recognit Lett 15:1119–1125
28. Vapnik V (1998) Statistical learning theory. Wiley, New York
29. Veltkamp RC, Hagedoorn M (2000) Shape similarity measures, properties, and constructions. Advances in visual information systems, pp 467–476
30. Wilson CL, Garris MD (1992) Handprinted character database 3. Technical Report, National Institute of Standards and Technology
31. Xing E, Jordan M, Karp R (2001) Feature selection for high-dimencional genomic microarray data. In: International Conference on Machine Learning, pp 601–608
32. Yu L, Liu H (2003) Feature selection for high-dimensional data: a fast correlation-based filter solution. In: International Conference on Machine Learning, Washington

Author Biographies

Artsiom Harol studied Radio-Physics and Electronics at Belarusian State University, Minsk, where he obtained his M.Sc. degree in 2002. Currently he is working towards his Ph.D at Delft University of Technology, The Netherlands. His interests are centered around probabilistic models from the alternative representations applied to pattern recognition and machine learning.

Carmen Lai received the M.Sc. degree in electronical engineering from Cagliari university, Italy, in 2002. Currently, she is a Ph.D student in pattern recognition applied in bioinformatics at the Delft University ofTechnology. She collaborates with the Dutch Cancer Institute on classification and analysis of high-throughput data.

Elżbieta Pękalska received an M.Sc. degree in computer science from Wrocław University, Poland. After carrying out a Ph.D research, in January 2005 she obtained a cum laude Ph.D degree for her work on dissimilarity representations in pattern recognition. She currently holds a post doc position at Delft University of Technology, being simultaneously a honorary visiting research associate at the University of Manchester, UK. Her research interests include representations in learning, statistical and structural learning methodologies, one-class classification problems, invariant measures, kernel methods and theory of Hilbert and Krein spaces. She is also looking into biomedical applications.



Robert P.W. Duin studied applied physics at Delft University of Technologyin the Netherlands. In 1978 he received the Ph.D degree for a thesison the accuracy of statistical pattern recognizers. In his research heincluded various aspects of the automatic interpretation of measurements,learning systems and classifiers. Between 1980 and 1990 he studiedand developed hardware architectures and software configurations forinteractive image analysis. After this period his interest was redirectedvia neural networks to pattern recognition.

At this moment he is an associate professor of the Faculty of Electrical Engineering, Mathematics and Computer Science of Delft University of Technology. His present research is in the design, evaluation and application of algorithms that learn from examples. This includes neural network classifiers, support vector machines and classifier combining strategies. Recently he started to investigate alternative object representations for classification and became thereby interested in dissimilarity based pattern recognition and in the possibilities to learn domain descriptions. Additionally he is interested in the relation between pattern recognition and consciousness.