Pairwise selection of features and prototypes

Elżbieta Pękalska, Artsiom Harol, Carmen Lai and Robert P.W. Duin

Information and Communication Theory group
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology, The Netherlands
{e.pekalska,a.harol,c.lai,r.p.w.duin}@ewi.tudelft.nl

Summary. Learning from given patterns is realized by learning from their appropriate representations. This is usually practiced either by defining a set of features or by measuring proximities between pairs of objects. Both approaches are problem dependent and aim at the construction of some representation space, where discrimination functions can be defined.

In most situations, some feature reduction or prototype selection is mandatory. In this paper, a pairwise selection for creating a suitable representation space is proposed. To determine an informative set of features (or prototypes), the correlations between feature pairs are taken into account. By this, some dependencies are detected, while overtraining is avoided as the criterion is evaluated in two-dimensional feature spaces. Several experiments show that for small sample size problems, the proposed algorithm can outperform traditional selection methods.

1 Introduction

The construction of a proper representation space is essential for designing successful learning procedures. Concerning both the computational efficiency and performance of a recognition system, one is usually interested in a space of a low dimensionality. Since the initial space may be large, some reduction methods are necessary to either detect or create informative features. An ideal technique is capable of reducing the dimensionality effectively, while preserving the class separability in the data. As some information is unavoidably lost in this process, it is, therefore, desirable to formulate a method that significantly reduces the dimensionality, but still preserves the information. In this paper, we focus on feature selection approaches.

Feature selection methods rely on a quantitative criterion that measures their performance. This criterion is used in some optimization process to determine a subset of informative features. Selection methods are usually divided into filters and wrappers [10]. Filters evaluate the relevance of features based on a feature capacity to discriminate between the classes. Wrappers employ a classification algorithm, used later to build the final classifier, to judge the

quality of a feature. Both approaches involve a combinatorial search through the constructed space of possible feature subsets. Usually, greedy procedures such as forward or backward eliminations are employed due to their computational attractiveness. More complex procedures such as floating searches and genetic algorithms can also be applied [5, 10, 14, 11].

Concerning the evaluation of the 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 their interdependencies into account. When features are correlated, these techniques are able to construct good feature subsets, while univariate techniques may fail. A disadvantage of multivariate approaches, however, is that they evaluate features in a multidimensional space, not only demanding a considerable computational effort, but also resulting in a loss of accuracy in case of a limited training set. Due to overfitting, feature subsets that do not ensure a good discrimination may be still judged as 'good'. The more features have to be selected, the worse this problem becomes.

In this paper, a pairwise feature selection procedure is investigated. Some ideas in this direction can be found in [2], where a particular pairwise selection algorithm was proposed for gene expression data. Since pairs of features are considered, second order dependencies are taken into account. Multidimensional spaces are now restricted to two dimensions, hence this method does not suffer from overfitting as other multivariate approaches do.

The problem of feature selection is similar to the selection of prototypes used to define a linear embedding of proximity data. In this case, a set of objects is represented by a dissimilarity matrix, where each entry describes a degree of commonality between pairs of objects. The chosen prototypes determine a vector space, in which all objects are represented as points and the corresponding dissimilarities are preserved as well as possible. Pairwise prototype selection is an appealing alternative to random, individual and multivariate selections [7, 12, 13], especially for low-dimensional embedded spaces.

2 Feature selection techniques

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

Three incremental selection methods are considered here. These are individual, forward and pairwise strategies. Assume that $F = \{f_1, f_2, \ldots, f_m\}$ is

a set of m features. Denote by $\tilde{F} \subset F$ a subset of the selected features. In each step, a feature or a pair of features is chosen according to some criterion J and added to $\tilde{F}$. Note that $\tilde{F} = \emptyset$ in the beginning.

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 be formally written as:

$$\begin{aligned} &\tilde{F} := \tilde{F} \cup f, \quad \text{where} \quad f : \max_{f_i \in F} J(f_i) \\ &F := F \backslash f \end{aligned} \tag{1}$$

In this procedure, features are ranked from the most to the least informative according to the criterion J and the most indicative features are finally selected.

Forward selection. Forward feature selection starts with the single most informative feature and adds next most informative features in a greedy fashion. Each step can be formalized as follows:

$$\begin{aligned} &\tilde{F} := \tilde{F} \cup f, \quad \text{where} \quad f : \max_{f_i \in F} J(F \cup f_i) \\ &F := F \backslash f \end{aligned} \tag{2}$$

Pairwise selection. The relevance of features is judged by evaluating pairs of features. In each step, the best feature pair is detected. Two approaches are here possible. Either both features are chosen from the current unselected feature set F or only one of them, as the other one comes from $\tilde{F}$. In each step, one has:

$$\begin{aligned} &\tilde{F} := \tilde{F} \cup \{f \cup f'\}, \quad \text{where} \quad \{f \cup f'\} : \max_{f_i \notin \tilde{F} \vee (f_j \neq f_i \wedge f_j \notin \tilde{F})} J(f_i \cup f_j) \\ &F := F \backslash \{f \cup f'\} \end{aligned} \tag{3}$$

Criterion. In our experiments, the inter-intra criterion is used [8]. It is applied in some representation space, where the between-scatter S_b and within-scatter S_w matrices are computed. S_w measures the average dispersion of a class sample around its mean, while S_b describes the scattering of the class means around the overall average. Given n samples, K classes and n_k samples per class, the inter-intra criterion is given as

$$J = \text{trace}(S_w^{-1} S_b), \tag{4}$$

where $S_w = \frac{1}{n} \sum_{k=1}^{K} n_k S_k$, $S_b = \frac{1}{n} \sum_{k=1}^{K} n_k (\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^T$, and $\mathbf{m}$ is the estimated overall mean and $\mathbf{m}_k$ and S_k are the estimated mean and covariance matrix of the k-th class, respectively. The higher value of the criterion, the more informative the corresponding feature. For a single feature and two-class problems, this criterion is equivalent to the Fisher criterion [5] $J_{FC} = \frac{|m_1 - m_2|}{\sqrt{s_1^2 + s_2^2}}$, where s_1 and s_2 are the class standard deviations.

3 Examples

The potential of pairwise feature selection is illustrated by three examples.

3.1 Artificial example

An artificial data set with some correlated feature pairs is generated to investigate the behavior of feature selection methods in a controlled environment. Assume that n samples and m features are given, where only q features, generated in correlated pairs, are informative. The samples for each correlated feature pair are drawn from a Gaussian distribution with the class means $\mu_1 = \begin{bmatrix} 0 & 0 \end{bmatrix}^T$ and $\mu_2 = \frac{\sqrt{2}}{2} \begin{bmatrix} r & -r \end{bmatrix}^T$ for some $r > 0$. The covariance matrix, identical for both classes, is $\Sigma_1 = \Sigma_2 = \begin{bmatrix} v+1 & v-1 \\ v-1 & v+1 \end{bmatrix}$. 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 the identity matrix. We set $m = 300$, $q = 20$, and, in order to have a class overlap, $r = 3$ and $v = \sqrt{40}$. Since we want to simulate a small sample size problem, we chose $n = 100$ for the training set, while the size of the test set is set to $n = 10000$.

For each selection method, a Fisher linear discriminant (FLD) [5] is trained on a training set with a growing number of features (starting from the best two features) and tested on an independent test set. All selection methods rely on the criterion (4). As a result, the classification error can be plotted versus the number of features used. The error is estimated based on 50 repetitions of the experiments with different generations of the training set.

Figure 1, left plot, shows the behavior of different feature selection methods as judged by the average classification error. The peaking phenomenon visible in the plot occurs when the number of samples is comparable to the number of features. This is due to the use of a pseudo-inverse instead of the usual inverse of the sample covariance matrix on which the FLD relies [15]. Some solutions to avoid this problem can be found e.g. in [16].

The univariate approach performs the worst and the pairwise selection performs the best. This is expected, since pairs of features are strongly correlated. Although the forward search reaches a higher accuracy than the univariate technique, it is limited by the greedy procedure it is based on.

Figure 1, right plot, shows the number of detected informative features versus the number of selected features. The results are the average of 50 experiments. The pairwise approach determines all informative features perfectly. In this small sample sizes problem, the forward search retrieves more uninformative features than the univariate approach.

3.2 Feature selection example

A feature selection example on a more general data set is here presented. The Waveform data, as described in [4], is chosen as it clearly shows what can be gained by the pairwise selection. This three-class problem is based on sampling triangle shaped waves and, thereby, it really needs a significant subset of the 21 original features in order to reach a proper class separation. There are 5000 objects in total, approximately equally distributed over the three classes.

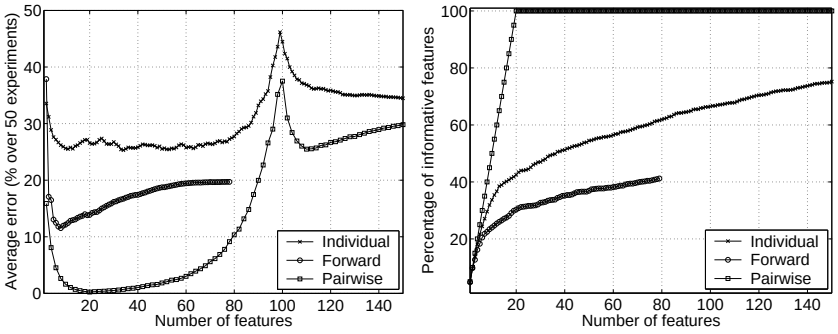

Fig. 1. Artificial data. Left: average classification error across 50 repetitions for the three feature selection procedures. Right: percentage of relevant features retrieved by the selection techniques.

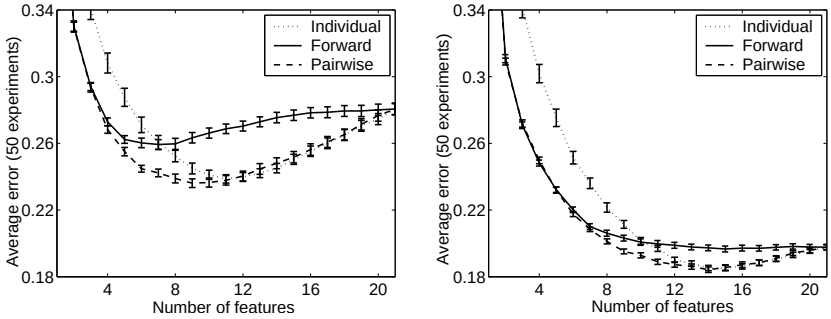

Fig. 2. Waveform data. Average classification error over 50 repetitions of the NLC for three feature selection procedures using 35 (left) and 100 (right) objects per class for feature selection and classifier training.

Figure 2, left plot, shows the average classification error over 50 experiments, in which 35 objects per class have been chosen at random. Using these objects, feature selections are performed based on the inter-intra criterion, formula (4). In the resulting feature spaces, a Bayes normal-density based linear classifier (NLC), assuming class normal distributions with equal covariance matrices [5], is trained on the same training set. The classifiers are tested on the remaining objects. The resulting error rates are averaged out and the standard deviations of the means are computed and shown in the plot. This is a clear example, where a pairwise selection behaves equal or better than the forward selection as well as the individual selection. For larger feature sizes, the forward procedure cannot compute the criterion values in a sufficiently accurate way, which leads to suboptimal feature subsets. The pairwise procedure shows a continuous improvement.

Figure 2, right plot, presents the results for 100 objects per class used for the feature selection and training. The same phenomena can be observed as

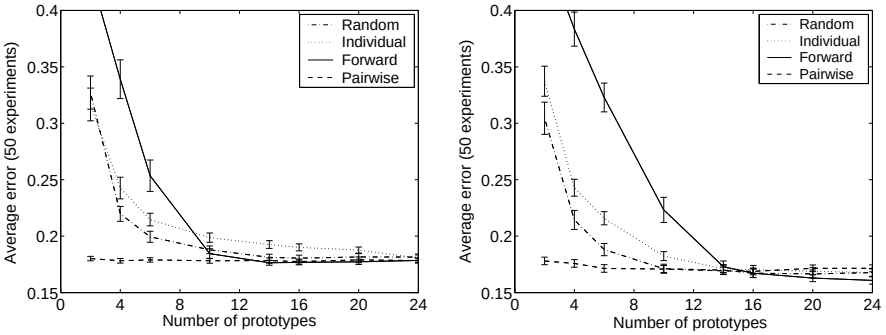

Fig. 3. Heart disease data. Average classification error over 50 repetitions based on 50 (left) and 150 (right) objects in total used for prototype selection, embedding and classifier training. The NLC is trained in embedded spaces.

for 35 objects per class, but less pronounced, as the forward procedure now suffers less from overtraining.

3.3 Prototype selection example

The heart disease data set [1] is considered for the prototype selection example. There are 303 cases, where 139 and 164 refer to ill and healthy patients, respectively. A subset of 13 attributes of mixed types is used to compute a Gower's distance representation [6], which is known to be Euclidean.

The data are split into a training set T and test set S with the same prior probabilities (either 50 or 150 objects in total for training). The distance matrix $D(T,T)$ is used for prototype selection, embedding and training the NLC. For testing, the distance matrix $D(S,T)$ is used. In the individual and forward selection methods, the prototypes are determined by using the inter-intra criterion (4) applied to $D(T,T)$. This means that a distance representation is interpreted in a vector space, where each dimension describes a distance to a particular object from T [12, 13]. Inspired by [17], the pairwise prototype selection is realized by evaluating the criterion in two-dimensional spaces determined by isometric embeddings of $D(\cdot,[p_i\ p_j])$, where p_i and p_j are different objects. The details can be found in [17, 7]. Having found a prototype set R, the classical scaling [3] is used for a linear isometric embedding of $D(R,R)$. The remaining $D(T\backslash R,R)$ are then projected to the embedded space and the NLC is trained [3, 13]. Testing is realized by projecting $D(S,T)$ to the same space and applying the trained NLC. The experiments are performed for growing prototype sets and repeated 50 times for various splits into the training and test sets. The average classification error is plotted as a function of the prototype set size; see Figure 3.

Note that the number of prototypes m defines the embedding and describes the dimensionality p of an embedded space ($m = p+1$). The prototypes should be significantly different (i.e. vectors of distances to them should differ) to preserve the most of the distance information in the data. For a small number of

prototypes, this holds for the pairwise selection and can be observed in Figure 3. The forward selection performs then the worst, since the embedding is defined by prototypes p_i and p_j which are characterized by correlated vectors of distances $D(\cdot, p_i)$ and $D(\cdot, p_j)$. In this case, this does not ensure yet that the resulting embedded space will be good for discrimination. The random selection is better here than the forward and individual selections as it tends to choose objects that differ with respect to distance information.

4 Discussion and conclusions

The need for dimension reduction holds in a similar way for traditional feature spaces as for embedded spaces defined on the dissimilarities to a set of prototype objects. In this paper, we presented a new procedure for dimension reduction by selection. There are several reasons to lower the dimensionality of a representation space in which classifiers have to be trained. First, less dimensions implies less computational effort to represent new objects to be classified: less features to be measured or less proximities to be computed. Secondly, in low-dimensional spaces the accuracy of trained classifiers is higher than in spaces with more dimensions. The trade-off is, however, that by removing dimensions (features) the class separability is deteriorated. So, feature selection should be done carefully.

An issue often neglected in previous studies on feature selection is that the accuracy of the criterion itself, like the classifier, also may suffer from small training set sizes. Procedures like backward elimination, branch and bound, forward selection [9] and floating search [14] evaluate the criteria in a multi-dimensional space. The estimation of the criterion values suffers from noise. Many criteria used for judging class separability are biased for small sample sizes: classes seem to be better separable than they are, in fact. Even when corrections for such a bias are made, e.g. by using the F-statistics, there is still a bias caused by the selection mechanism itself. This is for high-dimensional spaces more severe than for low-dimensional spaces as the variance in the criterion estimate is larger in the former case.

The individual evaluation and ranking of features suffers the least from this problem. It is, however, entirely unable to take into account the dependency between features in estimating the separability. The pairwise selection procedure studied in this paper makes some trade-off. Feature spaces are judged just in various combinations of feature pairs. So, whenever the dependency between two features is of importance, it is detected and can be used. This procedure is expected to be almost always better than individual ranking, except for very small sample sizes or for very large feature sets, as in these cases also pairwise evaluation will cause an overtraining. The proposed procedure may be better than the multivariate techniques when more than just a few features are needed and the training set size is small. For large training sets multivariate approaches do not suffer from overtraining and may detect higher

order useful dependencies between features. If the problem can be solved by a small set of features, multivariate techniques may find them as well.

In conclusion, the pairwise procedure for the selection of features or prototypes may be a useful strategy in case of small sample size problems. Some examples are presented to support this claim.

References

1. Blake CL and Merz CJ (1998) UCI Repository of machine learning databases, http://www.ics.uci.edu/ mlearn/MLRepository.html.
2. Bo T and Jonassen I (2002) New feature subset selection procedures for classification of expression profiles, Genome biology 3.
3. Borg I and Groenen P (1997) Modern Multidimensional Scaling. Springer-Verlag.
4. Breiman L, Friedman JH, Olshen RA and Stone CJ (1984) Classification and regression trees, Wadsworth, California.
5. Duda RO, Hart PE and Stork DG (2001) Pattern Classification 2nd. edition, John Wiley & Sons.
6. Gower JC, A general coefficient of similarity and some of its properties, Biometrics vol. 27, 25-33, 1971.
7. Harol A, Pękalska E and Duin RPW (2005), Pairwise prototype selection on distance data, submitted.
8. van der Heijden F, Duin RPW, de Ridder D and Tax DMJ (2004) Classification, Parameter Estimation and State Estimation. An Engineering Approach using Matlab. John Wiley & Sons Ltd.
9. Jain AK, Duin RPW and Mao J (2000) Statistical Pattern Recognition: A Review. IEEE Trans on PAMI 22:4-37.
10. Kohavi R and John GH (1997) Wrappers for feature subset selection. Artificial Intelligence 97:273-324.
11. Li L, Weinberg CR, Darden TA and Pedersen LG (2001) Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics 17:1131-1142.
12. Pękalska E, Duin RPW and Paclík P (2005), Prototype Selection for Dissimilarity-based Classifiers. accepted to Pattern Recognition.
13. Pękalska E (2005) Dissimilarity representations in pattern recognition. Concepts, theory and applications. PhD thesis. Delft University of Technology.
14. Pudil P, Novovicova J, and Kittler J (1994) Floating search methods in feature selection. Pattern Recognition Letters 15:1119-1125.
15. Raudys S, and Duin RPW (1998) On expected classification error of the Fisher linear classifier with pseudo-inverse covariance matrix. Pattern Recognition Letters 19:385-392.
16. Skurichina M (2001), Stabilizing weak classifiers. PhD thesis. Delft University of Technology.
17. Somorjai RL, Dolenko B, Demko A, Mandelzweig M, Nikulin AE, Baumgartner R, Pizzi NJ (2004) Mapping high-dimensional data onto a relative distance plane an exact method for visualizing and characterizing high-dimensional patterns, Journal of Biomedical Informatics 37:366-379.