ELSEVIER

Pattern Recognition Letters 20 (1999) 1175–1181

Pattern Recognition Letters

www.elsevier.nl/locate/patrec

Relational discriminant analysis

Robert P.W. Duin *, Elżbieta Pękalska, Dick de Ridder

Department of Applied Physics, Pattern Recognition Group, Delft University of Technology, P.O. Box 5046, 2600 GA, Delft, The Netherlands

Abstract

Relational discriminant analysis is based on a proximity description of the data. Instead of features, the similarities to a subset of the objects in the training data are used for representation. In this paper we will show that this subset might be small and that its exact choice is of minor importance. Moreover, it is shown that linear or non-linear methods for feature extraction based on multi-dimensional scaling are not, or just hardly better than subsets. Selection drastically simplifies the problem of dimension reduction. Relational discriminant analysis may thus be a valuable pattern recognition tool for applications in which the choice of the features is uncertain. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Featureless pattern recognition; Support vector classifiers; Proximities; Dimension reduction; Non-linear mapping; Multi-dimensional scaling; Distance mapping

1. Introduction

In statistical pattern recognition, objects are traditionally represented by features. Recently (Duin et al., 1997) we argued that formally objects might also be represented by a proximity measure (distances, similarities) to a set of prototypes or support objects. This leads to a featureless approach to pattern recognition in which the application expert expresses the domain knowledge by defining the proximity measure instead of by defining a set of features. Finding classifiers between classes represented in this way is called relational discriminant analysis.

This approach has some resemblance with one of the early methods for pattern recognition called template matching. That method is also featureless and is also based on some proximity measure. What is discussed here goes much further as we will build discriminant functions on similarities and try to reduce the complexity of the representation. In this study the similarity to particular objects in the training set will take over the role of features of the traditional feature based methods.

The starting point of this analysis is an $m \times m$ matrix D between all m training objects and a corresponding set of m labels Λ. We can relate them uniquely by a weight vector w as

$$Dw = \Lambda, \quad \text{so } w = D^{-1}\Lambda, \tag{1}$$

if $\mathrm{rank}(D) = m$. For a new object x, represented as a set of similarities to the m training objects, the label can now be estimated as

$$\lambda_x = x^{\mathrm{T}} w. \tag{2}$$

The problem with this discriminant is that it may have a small generalization power as it has no

* Corresponding author.
E-mail address: duin@ph.tn.tudelft.nl (R.P.W. Duin)

0167-8655/99/$ - see front matter
PII: S0167-8655(99)00085-9

noise averaging possibilities. If a smaller *representation set* $\mathbb{R}$ is used, having $n \ll m$ objects, then the corresponding D_r has size $m \times n$ and the weight vector has a reduced size of n weights. Eq. (1) can now be based on a mean square error procedure (Fisher's Linear Discriminant), but also other discriminant functions may be used.

A dimension reduction is important for two reasons. First, the accuracy of a classifier improves if it is trained by m objects but represented in a space with dimensionality $n \ll m$, due to the curse of dimensionality (Jain and Chandrasekaran, 1987). Secondly, it reduces the complexity of measurements and computations in classifying new objects as they have just to be represented by their similarities to the objects in R only. In this way, the representation set replaces the traditional feature set.

In this paper we will show that in a practical application the *selection of objects* is (almost) as good as *feature extraction* by linear as well as non-linear methods for multi-dimensional scaling. Next we will analyze how critical the choice of the reduction R is. We will show that it is hard to improve the performance based on just a random selection. This makes relational discriminant analysis a very simple procedure.

2. Dimension reduction by the selection of objects

As stated, the reduction of the number of columns in D from m to n by selecting a representation set is almost identical to the feature selection problem. There is, however, one important difference between the similarities to a representation set and a feature set: features can differ largely; some features may even be unique. The similarities to the objects in the representation set, on the other hand, have a uniform interpretation. They can be very similar, as they arise from the same training set. Unless the training set is very small, each object has some neighbors that are rather similar. Consequently, not only their mutual similarities are large, but also their relations with the other training objects will be alike.

We will discuss the following possibilities to select a representation set from the training set.

2.1. Random selection

In this case we do not take any precautions that the representation set is really representative for the training set. It is an easy and fast method. Without affecting that, we distribute the random choices evenly over the classes. We already experimented with this method in (Duin, 1998).

2.2. Systematic selection

By using the k-centers method (Ypma and Duin, 1998), the objects are distributed evenly over the training set. This method selects the objects in such a way that the smallest similarity is maximized. In our procedure we do this class by class.

2.3. Feature selection

Recalling that the problem is similar to feature selection we choose for the forward selection, using the Mahalanobis distance as a criterion. So now the elements of D are used as feature values and its rows are interpreted as points in a feature space. Although we used class information in the first two methods, only in this method it is used in such a way that class separability is optimized.

3. Feature extraction by multi-dimensional scaling

Instead of selecting objects, a dimension reduction in the relational representation can be achieved by mapping the original set of objects on a feature space of a given reduced dimensionality. In order to preserve the original structure defined by a proximity matrix, we will demand that the distances in the new space reflect these proximities as well as possible. This technique is called multi-dimensional scaling (Borg and Groenen, 1997). Because this technique is based on proximity measurements, it is well applicable to our featureless approach.

The idea behind multi-dimensional scaling is simple. A configuration of points in a low-dimensional space is sought such that their proximities

reflect as well as possible the original ones. The classifiers between classes are then determined for the lower-dimensional representation instead of for the original set.

As the starting point, an $m \times m$ distance matrix D is considered. The multi-dimensional scaling technique maps the original set into a lower-dimensional space, imposing that all interpoint distances are preserved as well as possible. As a result, a faithful, lower-dimensional representation of the geometrical relations between the objects is obtained. We use the Euclidean metric. It is important to emphasize that the multi-dimensional scaling results, the extracted set of features, can be interpreted as linear or non-linear combinations of the, possibly virtual, original set of features on which D is based.

We investigated the following possibilities.

3.1. Classical scaling

This is a linear technique, similar to the Karhunen–Loève reduction method. Suppose that the coordinate matrix $X \in \mathbb{R}^{m\times k}$ is given. Let D^2 be the square distance matrix. Knowing that distances do not change under translations, it can be assumed that X has column means equal to 0.

Let us consider the matrix $B_D = XX^{\mathrm{T}}$. The matrix B_D can be expressed in another way (Borg and Groenen, 1997) as

$$B_D = -\frac{1}{2}JD^2J, \quad J = I - \frac{1}{m}\mathbf{1}\mathbf{1}^{\mathrm{T}}, \quad \mathbf{1}^{\mathrm{T}} = [1\ldots1] \in \mathbb{R}^{1\times m}. \tag{3}$$

The matrix B_D, as a scalar product of matrices $B_D = XX^{\mathrm{T}}$, is symmetric and has non-negative eigenvalues. Then the factorization of B_D by its eigenvector decomposition can be found:

$$B_D = Q\Lambda Q^{\mathrm{T}} = \left(Q\Lambda^{0.5}\right)\left(Q\Lambda^{0.5}\right)^{\mathrm{T}}, \tag{4}$$

where Q is an orthogonal matrix (which means that the eigenvectors are normalized) and $\Lambda^{0.5}$ is a diagonal matrix with the diagonal elements equal to the square roots of the eigenvalues taken in descending order. Hence, we have the equation

$$XX^{\mathrm{T}} = B_D = \left(Q\Lambda^{0.5}\right)\left(Q\Lambda^{0.5}\right)^{\mathrm{T}}. \tag{5}$$

It is not enough to show that $X = Q\Lambda^{0.5}$, but it can be proved (Borg and Groenen, 1997) that X and $Q\Lambda^{0.5}$ differ only by rotation. From this fact, X can be retrieved from Q and Λ by

$$X = Q\Lambda^{0.5}, \tag{6}$$

which is correct except for possible rotations in the projected space.

This analysis shows that having the square distance matrix D^2, a data matrix X can be retrieved under condition that the distances are preserved.

In the method of classical scaling, the square distance matrix D^2 (for the original data X in a k-dimensional space) is given. It is not necessary that we know X. A configuration of points in a d-dimensional space is sought, for which all the distances are possibly well preserved. Following the above reasoning, the matrix $B_D = -\frac{1}{2}JD^2J$ is computed, which allows to determine the data matrix $X \in \mathbb{R}^{m\times k}$ by using formula (3). This is an exact, k-dimensional result, except for a rotation. However, our interest is not in X, but in a lower-dimensional representation. Therefore, the matrix $Y \in \mathbb{R}^{m\times d}$ can be expressed as

$$Y_d = Q_d\Lambda_d^{0.5}, \tag{7}$$

where the matrix Λ_d is the same as the matrix Λ, but with the first d eigenvalues greater than zero, and Q_d stands for the first d columns of the matrix Q corresponding to these d eigenvalues. By incorporating only partial information, our result is not perfect. However, by taking the largest d eigenvalues and corresponding eigenvectors, we assure that for the d-dimensional configuration Y, the square distance matrix resembles the original matrix D^2 in a sense of the largest variance.

It can be proved (Borg and Groenen, 1997) that in case of the Euclidean distance, the classical scaling solution is the same as the solution given by the Karhunen–Loève projection. This implies that classical scaling is a linear projection technique.

3.2. Sammon mapping

The well-known Sammon mapping (Sammon, 1969) is a non-linear projection technique. It looks for such a representation in a lower-dimensional space that the distances between the objects are as close as possible to the corresponding distances in the original (high-dimensional) space. In order to judge whether one configuration of points is better than another, an error function (also called stress) is considered. It measures the difference between the present configuration and the original configuration via the distances. Let us define:

- δ_{ij} for $i,j = 1,2,\ldots,m$, the distance between two points in a k-dimensional space,
- d_{ij} for $i,j = 1,2,\ldots,m$, the distance between two points in a d-dimensional space.

The stress function is then given as follows:

$$E_S = \frac{1}{\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\delta_{ij}}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\frac{(\delta_{ij}-d_{ij})^2}{\delta_{ij}}, \tag{8}$$

and yields in fact a badness-of-fit measure for the entire representation. It is a sort of the normalized error, incorporating all the differences between the original δ_{ij} and mapped d_{ij} distances. The problem of finding the right representation in a lower-dimensional space is then an optimization problem. The configuration of points is sought for which the stress is minimum. In general, this is a complex problem due to the operation on $O(m^2)$ distances.

To perform Sammon mapping, one starts from an initial configuration (e.g., randomly chosen). The stress is then computed. Next, the points are adjusted so that the stress decreases. The whole process is reiterated until the map corresponding to the (local) minimum of the stress function is obtained. The optimum is found by using e.g. the Pseudo-Newton or Conjugate Gradients techniques.

3.3. Niemann mapping

The idea of Niemann mapping (Niemann, 1980) is similar to Sammon mapping but computationally more attractive. Having a k-dimensional configuration of points, represented by the distance matrix, the objective is to find its lower-dimensional representation, which preserves all the original distances. The distance between the points is defined to be the square one. The following stress function is used:

$$E_N = \frac{1}{\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\delta_{ij}^2}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\frac{(\delta_{ij}^2-d_{ij}^2)^2}{\delta_{ij}^2}, \tag{9}$$

where the meaning of symbols δ_{ij} and d_{ij} remains the same.

This stress function is presented by a similar formula as the Sammon stress. The difference is hidden in the distances: squared Euclidean instead of Euclidean. Again, the problem of finding the proper configuration in the low-dimensional space becomes a minimization problem. However, in this case, the optimization is conveniently done now by *a coordinate descent algorithm* (Niemann, 1980).

This method tries to preserve the squared distances and by doing this reflects more the global structure of the data. The classical scaling technique makes use of the squared distance matrix, imposing that these distances are reflected in the lower-dimensional representation. The classical scaling method is a linear approach. Consequently, also Niemann mapping, which operates on the squared distances, has a more linear aspect. Our experience confirms that fact.

In case of a high-dimensional data set, the task of finding the classifiers between the classes is often not easy. Therefore, it is a reasonable approach to map this data into a lower-dimensional space so that the inherent structure is preserved. (One problem that had to be solved is how new data should be mapped into the extracted feature spaces. We found a good and simple method called *Distance Mapping* which is reported elsewhere (Pekalska et al., 1999). Finding classifiers should then become more simple. For this reason, the multi-dimensional techniques are of interest here.

The use of them allows also for a fair comparison with relational discriminant analysis. In our featureless approach, first the distance matrix is considered. We study then two ways of dimension reduction. The first one relates to the selection of certain distances, creating a representation set, in order to reduce the problem size. Therefore, some information is disregarded. The second way of

dimension reduction, discussed in this section is to project the data on a feature space of lower dimensionality by a multi-dimensional scaling method. A lower-dimensional representation is found, but again some information is lost, as the distances are distorted as they approximate the original ones.

4. Experiments

We used a character database consisting of 200×10 handwritten numerals, each originally represented by 30×48 binary pixels. Out of this dataset, 5 different feature sets are derived: *pixel* (240 averages of 2×3 pixels), *face* (216 face distances), *Fourier* (76 Fourier shape descriptors), *Karhumen–Loève* (64 weights) and *Zernike* (47 rotational invariant moments plus 6 morphological features). So in total there are 649 features. See also (van Breukelen et al., 1997).

For each feature set a 2000×2000 Euclidean distance matrix was computed. At random 1000 (100×10) objects were selected for training and 1000 (100×10) objects for testing. This was repeated five items. Then we used the three selection methods described in Section 2 and the three methods for feature extraction described in Section 3 to select 5 and 10 dimensions for each of the five datasets. This implies that for each dataset and each class just one or even no object was selected in case of the random selection and the k-centers method. The object distances were combined into a single representation of 25 and 50 distances per object. The relational matrices D_r for the training set are thus 1000×25 and 1000×50, respectively.

Finally, three classifiers were trained on the relational representations (treating them as feature spaces): the Bayes classifier assuming normal densities with equal covariance matrices (a linear classifier), the Bayes classifier assuming normal densities with unequal covariance matrices (a quadratic classifier) and the 1-Nearest Neighbor rule (1-NN). It appeared that the scaled version of this rule performed much better, so we used it for the selection methods. For the extraction methods this makes no sense, as multi-dimensional scaling already tries to optimize the distances. All classifiers were tested and results were averaged over the five random experiments, see Tables 1 and 2.

Table 1
Averaged error for a representation set of 50 objects

	Linear classifier		Quadratic classifier		1-NN rule	
Dimension reduction	Train	Test	Train	Test	Train	Test
Random selection	0.013	0.019	0.000	0.026	0	0.029
K-centers method	0.009	0.019	0.000	0.026	0	0.027
Forward selection	0.011	0.017	0.000	0.026	0	0.027
Classical scaling	0.009	0.019	0.000	0.029	0	0.068
Sammon mapping	0.012	0.021	0.000	0.029	0	0.068
Niemann mapping	0.011	0.016	0.000	0.028	0	0.060

Table 2
Averaged error for a representation set of 25 objects

	Linear classifier		Quadratic classifier		1-NN rule	
Dimension reduction	Train	Test	Train	Test	Train	Test
Random selection	0.028	0.037	0.002	0.031	0	0.037
K-centers method	0.027	0.032	0.003	0.030	0	0.037
Forward selection	0.021	0.032	0.001	0.031	0	0.038
Classical scaling	0.019	0.023	0.001	0.027	0	0.106
Sammon mapping	0.025	0.032	0.002	0.028	0	0.088
Niemann mapping	0.021	0.025	0.001	0.029	0	0.088

5. Discussion

Table 1 shows that if all classes and all datasets are represented in the subset, a random selection yields almost the same performance as any of the more advanced selection methods and mapping techniques. The bad results for the NN classifier for the feature extraction methods may be explained by the fact that multi-dimensional scaling optimizes the set of distances globally, influencing the NN relations.

Note that the random subset selection needs a training effort of about 1 minute on a Sun Ultra-10 workstation for the training set of size 1000×649. It immediately reduces the set to 1000×50 distances in which all original features are represented. The systematic selection and the multi-dimensional techniques for feature extraction need up to a few hours and much more memory as they keep on handling 1000×1000 matrices.

A further dimension reduction from 50 to 25 (Table 2) shows globally a decreasing performance. So here we really loose information. This increase of error, however, is larger for the selection methods than for the mapping techniques. From this it can be concluded that feature extraction may be better able to preserve the class separability.

The selection methods are also much faster from the operational point of view. They demand the computation of just 25 or 50 distances instead of 1000 in classifying new objects.

Our final conclusion is that relational discriminant analysis based on random object selection is computationally efficient in both training and testing and may have a close to optimal performance.

For futher reading, see (Duin and de Ridder, 1997).

Discussion

Raghavan: When we were both here during the previous conference in the "Pattern Recognition in Practice" series we were talking about a paper by Lev Goldfarb, somewhat related, I think, to multi-dimensional scaling. (*Note of the editors*: *see the discussion in*: *R.P.W. Duin*, *D. de Ridder and D.M.J. Tax. Experiments with a Featureless Approach to Pattern Recognition*, *Pattern Recognition Letters*, *18*, *1997*, *pp. 1159–1166*.) The results of that paper allowed the type of distances to be more general than Euclidean distances. In your case, do the distances have to be Euclidean?

Duin: Not at all. I think that this is the nice thing of this relational representation. These distances are just given. I did not say that these distances have to be Euclidean.

Raghavan: No, but for the second group of techniques that you use, how well they can map, while preserving the distances, depends on whether the method itself can handle distances that are not Euclidean.

Duin: No, the given distance matrix may be any distance matrix. We map it in a new feature space and there we use the Euclidean distance. But there is no assumption about the given distances. Any similarity measure can be used for it.

Kuncheva: You used 1000 objects and you have over 600 features. I am wondering, have you tried to experiment with these features to see how some classifiers, like linear and quadratic discriminant classifiers, work?

Duin: I did that, but I do not have the figures available here. I think they are much worse because you have to compute covariance matrices in a 649-dimensional feature space based on just 1000 objects.

Kuncheva: What would happen if you would first reduce the feature space, to for instance 25 and then 50, and you would then use the same methods, including random selection?

Duin: I did not do that, because my goal was not to solve the feature space problem, but, starting from a 1000×1000 matrix, to compare various methods, given this relational representation.

Kanal: Are you saying that a bunch of objects is randomly selected to represent those 1000 objects?

Duin: Yes, the objects selected here are used for representing all objects, to be used as a basis for my representation. And I say that this basis for the representation is not very sensitive to the particular method according to which these objects are selected. It has something to do with the intrinsic dimensionality of the original problem. In my example, I select 50 objects if the intrinsic dimensionality is something like 20. Any set of 50 objects will represent the original set. And then all objects are mapped into the feature space built by this set of randomly selected objects. And thus, the training procedure of the classifier still uses all original objects. So it is a random selection of the representation basis, not of the training set.

Raghavan: What is the rationale for showing that the selected objects are in some way orthogonal to each other?

Duin: I treat the new feature space as a normal feature space, to compute classifiers. Some of the methods are sensitive to the orthogonality of the features, others are not.

References

Borg, I., Groenen, P., 1997. Modern Multi-dimensional Scaling. Springer, Berlin.

Duin, R.P.W., 1998. Relational discriminant analysis and its large sample size problem. In: Jain, A.K., Venkatesh, S., Lovell, B.C. (Eds.), ICPR'98, Proceedings of the Fourteenth International Conference on Pattern Recognition, Brisbane, 16–20 August 1998. IEEE Computer Society Press, Los Alamitos, pp. 445–449.

Duin, R.P.W., de Ridder, D., 1997. Neural network experiences between perceptrons and support vectors. In: Clark, A.F. (Ed.), Proceedings of the Eighth British Machine Vision Conference, Colchester, UK, 8–11 September 1997. University of Essex, Colchester, UK, pp. 590–599.

Duin, R.P.W., de Ridder, D., Tax, D.M.J., 1997. Experiments with a featureless approach to pattern recognition. Pattern Recognition Letters 18 (11–13), 1159–1166.

Jain, A.K., Chandrasekaran, B., 1987. Dimensionality and sample size consideration in pattern recognition practice. In: Krishnaiah, P.R., Kanal, L.N. (Eds.), Handbook of Statistics 2, North-Holland, Amsterdam, pp. 835–855.

Niemann, H., 1980. Linear and non-linear mappings of patterns. Pattern Recognition 12, 83–87.

Pekalska, E., de Ridder, D., Duin, R.P.W., Kraaijveld, M.A., 1999. A new method of generalizing Sammon mapping with application to algorithm speed-up. In: Boasson, M., Kaandorp, J.A., Tonino, J.F.M., Vosselman, M.G. (Eds.), ASCI'99, Proceedings of the Fifth Annual Conference of the Advanced School for Computing and Imaging, Heijen, NL, 15–17 June 1999. ASCI, Delft, pp. 221–228.

Sammon Jr., J.W., 1969. A nonlinear structure analysis mapping for data. IEEE Trans. Comp. C-18, 401–409.

van Breukelen, M., Duin, R.P.W., Tax, D.M.J., den Hartog, J.E., 1997. Combining classifiers for the recognition of handwritten digits. In: Pudil, P., Novovicova, J., Grim, J. (Eds.), Proceedings of the First International Workshop Statistical Techniques in Pattern Recognition, Prague, CR, 9–11 June 1997. pp. 13–18.

Ypma, A., Duin, R.P.W., 1998. Support objects for domain approximation. In: Niklasson, L., Boden, M., Ziemke, T. (Eds.), ICANN'98 Proceedings of the Eighth International Conference on Artificial Neural Networks, Skovde, Sweden, 2–4 September 1998. Springer, Berlin, pp. 719–724.