The alpha-procedure - a nonparametric invariant method for automatic classification of multi-dimensional objects

Tatjana Lange* Pavlo Mozharovskyi**

*Hochschule Merseburg, 06217 Merseburg, Germany

**Universität zu Köln, 50923 Köln, Germany

January 10, 2013

Abstract

A procedure, called α-procedure, for the efficient automatic classification of multivariate data is described. It is based on a geometric representation of two learning classes in a proper multi-dimensional rectifying feature space and the stepwise construction of a separating hyperplane in that space. The dimension of the space, i.e. the number of features that is necessary for a successful classification, is determined step by step using 2-dimensional repères (linear subspaces). In each step a repère and a feature are constructed in a way that they yield maximum discriminating power. Throughout the procedure the invariant, which is the object's affiliation with a class, is preserved.

1 Introduction

A basic task of pattern recognition consists in constructing a decision rule by which objects can be assigned to one of two given classes. The objects are characterized by a certain number of real-valued properties. The decision rule is based on a trainer's statement that states for a training sample of objects, whether they belong to class V_1 or class V_2. Many procedures are available to solve this task, among them binary regression, parametric discriminant analysis, and kernel methods like the SVM; see e.g. Hastie et al. (2009).

A large part of nonparametric approaches search for a separating (or rectifying) hyperplane dividing the two training classes in a sufficiently high-dimensional *feature space*. In doing so we face the

problem that the 'competencies' of measured properties (forming the axes of the original space) are unknown. Even more, we also do not know the correct scale of a property.

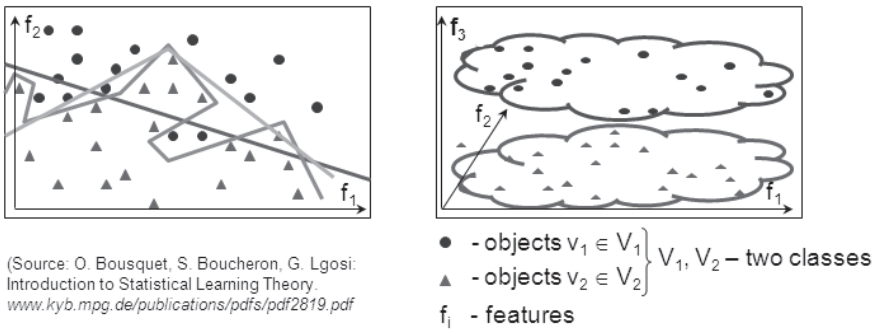

Figure 1: Instable solution for the separation of the classes (left). The selection of an informative property (right).

Very often these uncertainties lead to a situation where the complete separation of the patterns (or classes) of a training sample becomes difficult (Fig. 1, left). All these factors can cause a very complex separating surface in the *original* space which correctly divides the classes of the training sample but works rather poorly in case of new measured data. The selection of a more 'informative' property (Fig. 1, right) can give less intricate and thus more stable decisions.

The *α-procedure* (Vasil'ev (2003), Vasil'ev (2004), Vasil'ev (1969), Vasil'ev (1996) and Vasil'ev and Lange (1998)) uses the idea of a *general invariant* for stabilizing the selection of the separating plane. The invariant is the *belonging of an object to a certain class* of the training sample. The α-procedure - using repères - performs a step-by-step search of the direction of a straight line in a given repère that is as near as possible to the trainer's statement, i.e. separates best the training sample. It is completely nonparametric. The properties of the objects which are available for the recognition task are selected in a sequence one by one. With the most powerful properties a new space of 'transformed features' is constructed that is as near as possible to the trainer's statement.

2 The α-procedure

First, we perform some pre-selection, taking into further considerations only those properties p_q, $q = 1, \ldots, m$, whose values are completely separated or have some overlap as shown in Fig. 2. Next, we

define the *discrimination power* or *separating power* of a *single* property p_q as

Figure 2: Classification by a single property p_q, with $l = 15$, $c_q = 6$.

$$F(p_q) = \frac{c_q}{l}, \tag{1}$$

where l is the length of the training sample (= number of objects) and c_q is the number of correctly classified objects.

We set a minimum admissible discrimination power F_{min}, and at the first step select any property as a possible *feature* whose discrimination power exceeds the minimum admissible one:

$$F(p_q) > F_{min} \tag{2}$$

For the synthesis of the space, we select step-by-step those features that have best discrimination power. Each new feature shall increase the number of correctly classified objects. For this, we use the following definition of the *discrimination power* of a feature, selected at step k:

$$F(x_k) = \frac{\omega_k - \omega_{k-1}}{l} = \frac{\Delta\omega_k}{l}, \quad \omega_0 = 0, \tag{3}$$

where ω_{k-1} is the accumulated number of correctly classified objects before the k-th feature was selected and ω_k is the same after it was selected.

At Stage 1 we select a property having best discrimination power as a basis feature f_0 (= first axis) and represent the objects by their values on this axis; see Fig. 3.

At Stage 2 we add a second property p_k to the coordinate system and project the objects to the plane that is spanned by the axes f_0 and p_k. In this plane a ray originating from the origin is rotated up to the point where the *projections* of the objects onto this ray provide the best separation of the objects. The resulting ray, characterized by its rotation angle α, defines a possible new axis. We repeat this procedure for all remaining properties and select the property that gives the best separation of the objects on its rotated axis, which is denoted as $\widetilde{f}_1$. This axis is taken as the first *new feature*, and the respective plane as the first *repère*; see Fig. 4.

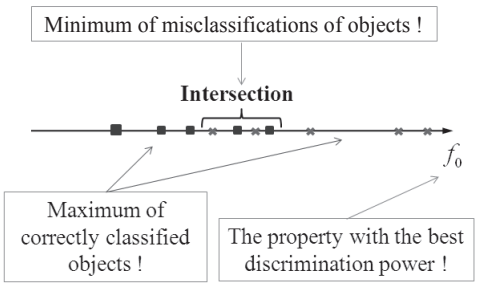

Figure 3: α-procedure, Stage 1.

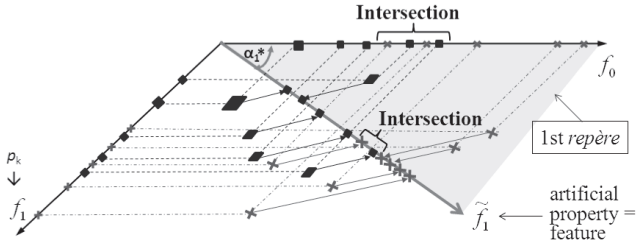

Figure 4: α-procedure, Stage 2.

At Stage 3 we regard another property p_j that has not been used so far and define the position of the objects in a *new plane* that is built by the axes $\widetilde{f}_1$ and p_j. Again we consider a ray in this plane and turn it around the origin *by the angle* α until the projections of the objects onto this axis give the best separation. We repeat this procedure for all remaining properties and select the best one, which, together with $\widetilde{f}_1$ forms the second repère (Fig. 5). In our simple example this feature already leads to a faultless separation of the objects.

If all properties have been used but no complete separation of all objects reached, a special stopping criterion as described in Vasil'ev (1996) is to be used.

3 Some formulae

As we see from the description of the idea, the procedure is the same at each step except for the first basic step defining f_0.

Let us assume that we have already selected $k-1$ features. We will use the symbol $\tilde{x}_{i,(k-1)}, i = 1, ..., l$, for the projections of the objects onto the feature f_{k-1} and ω_{k-1} as the number of already correctly classified objects (Fig. 6). For the *next step* k we have to compute

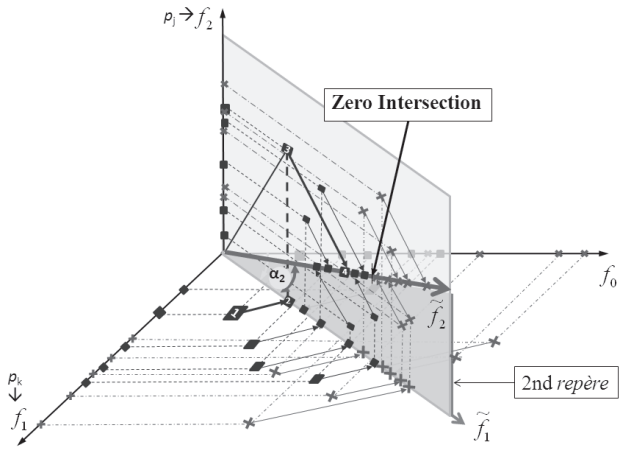

Figure 5: α-procedure, Stage 3.

the projection $\tilde{x}_{i,(k)} = \rho_i \cos(\beta_i + \alpha_q)$ for *all remaining* properties p_q, where $\tilde{x}_{i,(k-1)}$ is the value of feature $k-1$ for object i, x_{iq} is the value of property q for object i, $\rho_i = \sqrt{\tilde{x}^2_{i,(k-1)} + x^2_{iq}}$, $\beta_i = \arctan(x_{iq}/\tilde{x}_{i,(k-1)})$.

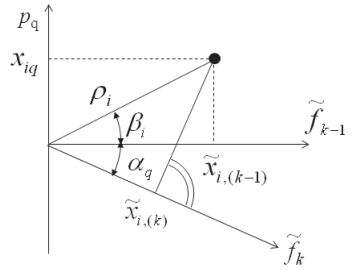

Figure 6: Calculating the value of feature $k-1$ for object i.

After n steps the normal of the separating hyperplane is given by

$$\left(\prod_{k=2}^{n} \cos\alpha_k^0, \sin\alpha_2^0 \prod_{k=3}^{n} \cos\alpha_k^0, ..., \sin\alpha_{n-1}^0 \cos\alpha_n^0, \sin\alpha_n^0 \right), \qquad (4)$$

where α_k^0 denotes the angle α that is best in step k, $k = 2, ..., n$. Due to the fact that (4) is stepwise calculated, the underlying features must be assigned *backwards* in practical classification. For example, the separation decision plane and the decomposition of its normal vector are shown in Fig. 7 and 8.

Note: If the separation of objects is not possible in the original space of properties, the space can be extended by building additional properties using products of the type $x_{iq}^s \cdot x_{ir}^t$ for all $q, r \in \{1, ..., m\}$ and i and some (usually small) exponents s and t. The solution is then searched in the extended space.

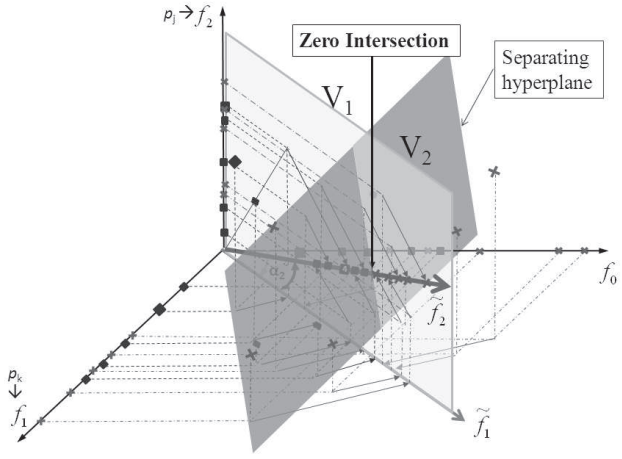

Figure 7: The separating decision plane.

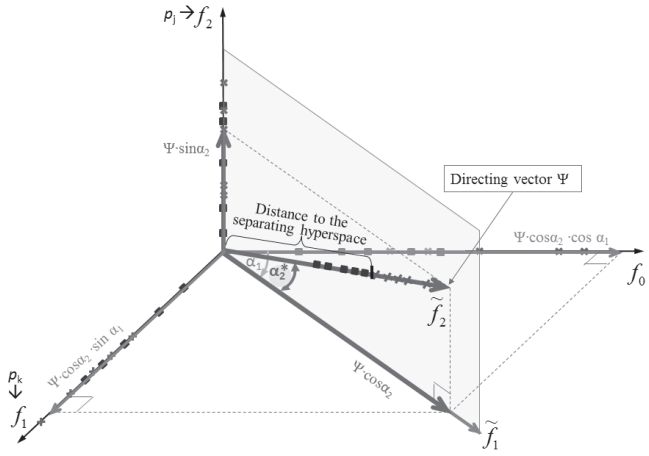

Figure 8: The separating decision plane with its defining vector.

4 Simulations and applications

To explore the specific potentials of the α-procedure we apply it to simulated data. Besides the application to the original data, the α-procedure can also be applied to properly transformed data; in particular, it has been successfully used to classify data on the basis of their so called DD-plot (DDα-classifier), see Lange et al. (2012a), Lange et al. (2012b). The α-procedure (applied in the original space (α-pr.(1)) and the extended space using polynomials of degree 2 (α-pr.(2)) and 3 (α-pr.(3))) is contrasted with the following nine classifiers: linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), k-nearest neighbors classification (KNN), maximum depth classification based on Mahalanobis (MM), simplicial (MS), and halfspace (MH) depth, and DD-classification with the same depths (DM, DS and DH, correspondingly; see Li et al. (2012) for details), and to the DDα-classifier.

Six simulation alternatives are used; each time a sample of 400 objects (200 from each class) is used as a training sample and 1000 objects (500 from each class) to evaluate the classifier's performance (= classification error). First, normal location (two classes originate from $N(\left[\begin{smallmatrix}0\\0\end{smallmatrix}\right], \left[\begin{smallmatrix}1&1\\1&4\end{smallmatrix}\right])$ and $N(\left[\begin{smallmatrix}1\\1\end{smallmatrix}\right], \left[\begin{smallmatrix}1&1\\1&4\end{smallmatrix}\right])$, see Fig. 9, left) and normal location-scale (the second class has covariance $\left[\begin{smallmatrix}4&4\\4&16\end{smallmatrix}\right]$, see Fig. 9, middle) alternatives are tried.

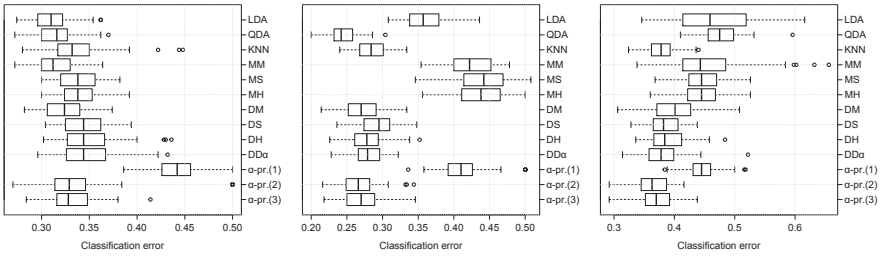

Figure 9: Boxplots of the classification error for normal location (left) and location-scale (middle), and normal contaminated location (right) alternatives over 100 takes.

To test the proposed machinery for robustness properties we challenge it using contaminated normal distribution, where the first class of the training sample of the normal location and location-scale alternatives considered above contains 10% objects originating from $N(\left[\begin{smallmatrix}10\\10\end{smallmatrix}\right], \left[\begin{smallmatrix}1&1\\1&4\end{smallmatrix}\right])$ (see Fig. 9, right and Fig. 10, left correspondingly). Other robustness aspects are demonstrated with a pair of Cauchy distributions forming a similar location-scale alternative, see Fig. 10, middle. Settings with exponential distributions ($(\mathrm{Exp}(1), \mathrm{Exp}(1))$ *vs.*

$(\mathrm{Exp}(1)+1, \mathrm{Exp}(1)+1)$, see Fig. 10, right) conclude the simulation study.

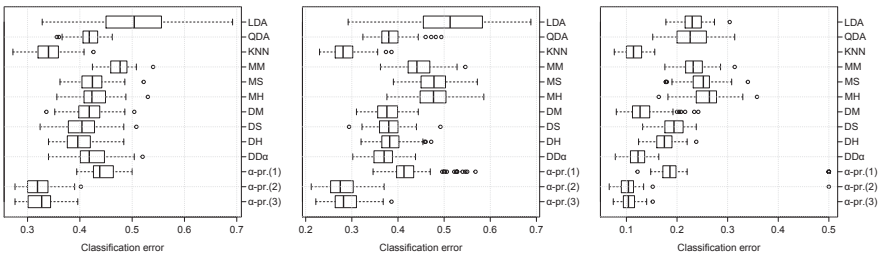

Figure 10: Boxplots of the classification error for normal contaminated location-scale (left), and Cauchy location-scale (middle), and exponential (right) alternatives over 100 takes.

The α-procedure performs fairly well for normal alternatives and shows remarkable performance for the robust alternatives considered. Though it works well on exponential settings as well, its goodness appears to depend on the size of the extended feature space. The last choice can be made by using either external information or cross-validation techniques.

5 Conclusion

The α-procedure calculates a separating hyperplane in a (possibly extended) feature space. In each step a two-dimensional subspace is constructed where, as the data points are naturally ordered, only a circular (that is, linear) search has to be performed. This makes the procedure very fast and stable. The classification task is simplified to a stepwise linear separation of planar points, while the complexity of the problem is coped with by the number of features constructed. The angle α of the plane at step $(k-1)$ defines a basic vector of the following *repère* at step k. Finally, the α-procedure is coordinate-free as its invariant is the belonging of an object to a certain class only.

References

[1] AIZERMAN, M.A., BRAVERMAN, E.M. and ROZONOER, L.I. (1970): *Method of potential functions in the theory of pattern recognition* (in Russian). Nauka, Moskow.

[2] HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J.H. (2009): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd Edition. Springer Verlag. New York.

[3] MOSLER, K. (2002): *Multivariate Dispersion, Central Regions and Depth: The Lift Zonoid Approach*. Springer Verlag, New York.

[4] MOSLER, K., LANGE, T. and BAZOVKIN, P. (2009): Computing zonoid trimmed regions of dimension $d > 2$. *Computational Statistics and Data Analysis, 53(7), 2000–2010.*

[5] LANGE, T., MOSLER, K. and MOZHAROVSKYI, P. (2012a): Fast nonparametric classification based on data depth. *Statistical Papers* (in print).

[6] LANGE, T., MOSLER, K. and MOZHAROVSKYI, P. (2012b): DDα classification of asymmetric and fat-tailed data. *This Volume.*

[7] LI, J., CUESTA-ALBERTOS J.A. and LIU R.Y. (2012): *DD*-classifier: Nonparametric classification procedure based on *DD*-plot. *Journal of the Americal Statistical Association, 107, 737–753.*

[8] NOVIKOFF, A. (1962): On convergence proofs for perceptrons. *Proceedings of the Symposium on Mathematical Theory of Automata. Polytechnic Institute of Brooklyn, 615–622.*

[9] VASIL'EV, V.I. (1969): *Recognition Systems - Reference Book* (in Russian). Naykova dumka, Kyiv.

[10] VASIL'EV, V.I. (1996): The theory of reduction in extrapolation problem (in Russian). *Problemy Upravleniya i Informatiki, 1-2, 239–251.*

[11] VASIL'EV, V.I. (2003): The reduction principle in problems of revealing regularities. Part I (in Russian). *Cybernetics and System Analysis, 39(5), 69–81.*

[12] VASIL'EV, V.I. (2004): The reduction principle in problems of revealing regularities. Part II (in Russian). *Cybernetics and System Analysis, 40(1), 9–22.*

[13] VASIL'EV, V.I. and LANGE, T. (1998): The duality principle in learning for pattern recognition (in Russian). *Kibernetika i Vytschislit'elnaya Technika, 121, 7–16.*

[14] VAPNIK, V. and CHERVONENKIS, A. (1974): *Theory of Pattern Recognition* (in Russian). Nauka, Moscow.