FEATURE SELECTION WITH NONPARAMETRIC STATISTICS

E. Franceschi[1]*, F. Odone*[1]*, F. Smeraldi*[2] *and A. Verri*[1]

[1] INFM - DISI, Università di Genova, Italy
[2] Computer Science, Queen Mary, University of London, UK

ABSTRACT

In this paper we discuss a general framework for feature selection based on nonparametric statistics. The three stage approach we propose is based on the assumption that the available data set is representative of a certain concept and aims at learning from the data the selection of a subset of descriptive features out of a large pool of measurements. The first stage requires the computation of a large number of image features. Simple significance tests and the maximum likelihood principle are at the basis of the second stage in which a saliency measure is used to reject the features which do not appear to be descriptive of the given data set. The third and final stage, by using the Spearman independence rank test, selects a maximal number of pairwise independent features. We report experiments on a face dataset (the MIT-CBCL database) which confirm the quality and the potential of the approach.

1. INTRODUCTION

The design of discriminative image representations through the skillful selection and disposition of image features is of paramount importance in pattern recognition, and a vast literature exists on specific problems such as, for instance, face or fingerprint processing. However, the need to deal automatically with generic patterns has motivated the development of several general–purpose feature selection methods (see [1, 2] and references therein, for example).

Automatic feature selection aims at extracting, out of a large pool of measurements, a reduced subset of *descriptive* features for the problem at hand. In pattern recognition, descriptiveness is often measured as discriminative power with respect to a "reject class" represented by an ad-hoc set of negative training examples [3, 4]. This requires casting into the two-class framework problems like object detection that would more naturally fit into the concept–learning scheme. In the limit case, descriptiveness becomes a by-product of the classifier training process, as with applications of Adaboost [3] and Support Vector Machines [5] (through the optimization of the α coefficients). However, feature selection arguably belongs to a more fundamental layer than classification, and can be founded on information theory [6], algebraic properties of the feature set (orthogonality, completeness [7]), or, as in our case, statistical independence.

In this paper, we present an all-purpose, nonparametric feature selection algorithm suitable for concept learning. We define a nonparametric measure of feature *descriptiveness* inspired by significance tests and the maximum likelihood principle. This notion of saliency holds under very general assumptions on the (unspecified) distribution of the negative test cases, and applies to uni-modal and multi-modal distributions as well.

By ranking the features in order of decreasing descriptiveness and performing a coarse thresholding, a large set of descriptive features is obtained. We then proceed to distill a subset of these by applying Spearman's independence rank test [8] to all feature pairs. A graph is constructed with a node for each feature and arcs joining each pair of independent features. A *maximal* descriptive set of independent features is then obtained as the union of the maximal cliques of the graph.

We report feature selection experiments over the facial images in the MIT-CBCL database [9], starting from a set of rank features (ranklets, [10]) that have an orientation selectivity pattern similar to Haar wavelets. Our results confirm the suitability of the selection process and the descriptiveness and stability of the selected features.

2. MEASURING FEATURE DESCRIPTIVENESS

In a concept learning problem, or whenever the "reject" class is not clearly specified, feature selection can only rely on very general assumptions. The maximum likelihood principle seems to be the most non-committal option — in general, we will want to accept a candidate object if the likelihood of the observed feature values is high under our model. Therefore, features with a peaked distribution seem to be preferable, as most instances of the object will correspond to high values of the likelihood. Seen in a different way, a sharply peaked distribution shows that the corresponding feature is consistent across several instances of the object, and thus captures some of its intrinsic characteristics.

However, since feature densities are in general multi-

0-7803-9134-9/05/$20.00 ©2005 IEEE

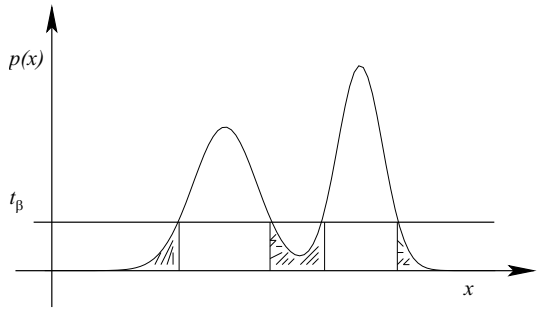

Fig. 1. The dashed areas of the distribution $p(x)$ of a feature contribute to the event B_t. For a fixed $\mathsf{B}_t = \beta$, t_β measures the descriptiveness of the feature.

modal, the sample variance is not a suitable measure of dispersion. For this reason, starting from the empirical marginal density $p(x)$ for a given feature, we perform a change of variables and define the probability density function $f(t)$ as

$$\int_0^t f(z)dz = \int_{-\infty}^{+\infty} p(x)U_0(t - p(x))dx$$

where $U_0(\cdot)$ is the unit step function. For a fixed $t \geq 0$, the integral on the l.h.s. is equal to the probability of the event

$$\mathsf{B}_t = p^{-1}([0, t]) = \beta, \tag{1}$$

i.e. the event that the likelihood will be smaller than t (see the dashed area in Fig. 1).

Our measure of feature descriptiveness is obtained as follows: let $0 < \beta \leq 1$ be fixed; for each feature i, solve Eq. 1 for $t = t_{i,\beta}$ and rank the features in order of $t_{i,\beta}$.

Note that β represents the contribution to the rate of false negatives that would result by discarding all instances of the object for which the likelihood of feature i is below $t_{i,\beta}$. As such, β is independent on the distribution of the negative class, its role being related to the significance level of a hypothesis test (see [11] for details).

After ranking the features, we proceed to select a subset of *descriptive* features by a coarse thresholding on $t_{i,\beta}$.

3. TESTING FOR FEATURE INDEPENDENCE

Out of the subset of descriptive features identified in the ranking phase, we then proceed to select a reduced descriptive subset based on statistical independence. The rationale is that once the feature set is pruned to a subset of independent variables, each feature in it contributes new information to the description of the object. At the same time, classification is simplified by the factoring of the multivariate density of the model into a product of univariate marginal distributions. An heuristic criterion for feature independence is used, for instance, in [12]; however, no quantitative measure is employed.

In our work, we make use of Spearman's independence rank test [8], which is an effective nonparametric way to estimate independence between two observables (in our case two features) that may have different measurement units.

Assume we are given n realizations of two random variables, R and S. Let $\pi_R(r_i)$ and $\pi_S(s_i)$ represent the rank of each observation among those of the respective variable. The Spearman's statistics $\mathcal{D}$ is defined as

$$\mathcal{D} = \sum_{i=1}^{n} (\pi_R(r_i) - \pi_S(s_i))^2.$$

The null distribution of $\mathcal{D}$ is obtained under the assumption that for independent variables all rankings occur with probability $1/n!$. For large n a normal approximation holds, with the tails corresponding to correlated or anticorrelated variables (i.e., equal or opposite rankings). Thus one runs a test against the independence hypothesis with significance α by checking whether $\mathcal{D}$ deviates from its average by more than some critical value d_α.

By performing Spearman's test over all pairs of descriptive features, we can thus quantitatively assess pairwise independence. We are then left with the problem of selecting a maximal subset of independent features. This is done by building a graph of which the single measurements represent the nodes. Two nodes are joined by an edge if the corresponding features *don't* reject the independence hypothesis at a fixed level of significance $0 < \alpha < 1$. Finally, maximally complete subgraphs - or *cliques* - are located in each graph. The nodes of the clique correspond to features pairwise independent with confidence greater than $1 - \alpha$.

Note that since all features not perfectly correlated can contribute some new information [13], the choice of α in practice expresses a balance between the dimensions of the resulting feature set and the descriptiveness of the final model.

4. EXPERIMENTS

In this section we report experiments on the MIT-CBCL face database. The set of measurements we start with are horizontal and vertical ranklets [10], rank features similar to wavelets invariant to image equalization (and thus well suited for the grey-level face images of very small size, 19×19 pixels, of the MIT-CBCL database). Other choices are clearly possible but ranklets, which appear to be well suited for face detection [10], seem to be sufficient to assess the potential of the proposed framework. The set we used consists of 2429 face images and 4548 non-face images (which we doubled in number by creating mirror images to enforce vertical symmetry). We started by computing a set of 10368 measurements obtained by varying the size support windows of the ranklets from 2×2 to 18×18 and shifting it all over the image.

Descriptiveness The solid line in Fig. 2 (left) shows the measurement distribution on the positive examples of the horizontal ranklet 5089 (ranked first according to the saliency procedure describe in Section 2). The distribution of the same measurements on the negative examples, which was nowhere used in the procedure, is displayed by the dotted line. An example of measurement ranked very low by the same procedure (horizontal ranklet 4018) is given through the solid line of Fig. 2 (right). The distribution of the same measurements on the negative examples, which again was nowhere used in the procedure, is displayed by the dotted line. From the qualitative viewpoint, visual inspection

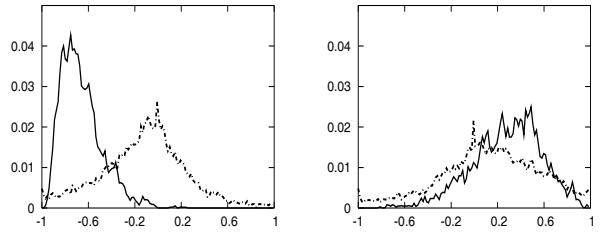

Fig. 2. Example of good and bad feature: *(left)* distribution of the horizontal ranklet 5089 on the positive examples (solid line) and on the negative examples (dotted line); *(right)* same as above for the horizontal ranklet 4018.

of a number of measurement distribution confirm that the adopted saliency measure makes sense. The range of values of the threshold $t_{i,\beta}$ in this case is quite small (typically 0.010 for top ranked features and 0.007 for low ranked features). We can approximately quantify the effectiveness of our measure of saliency by estimating the extent to which using negative examples would improve the selection. With reference to Section 2, for a fixed β (which controls the false negatives) we can select those features for which rejecting candidates with a likelihood below $t_{i,\beta}$ minimises the number of false positives over the training set. In the statistical test interpretation sketched in Section 2, this would correspond to maximising the power of the test against the alternative provided by the empirical density of the negative training examples. Fig. 3 compares the cumulative effect of this maximum power criterion with our descriptiveness measure in the case of horizontal ranklets. The graph is obtained by fixing $\beta = 10\%$ and plotting the sum of the fractions of negative examples rejected using the two possible sortings. The two methods end up reaching the same cumulative effect when all features are used. By figure inspection it appears that the net gain obtained by using the negative examples does not seem to be macroscopic. This point calls for further study and analysis.

The support of the 40 top discriminative features (for horizontal ranklets) is shown in the top row Fig. 4 (10 in each image from left to right). The middle row of the same

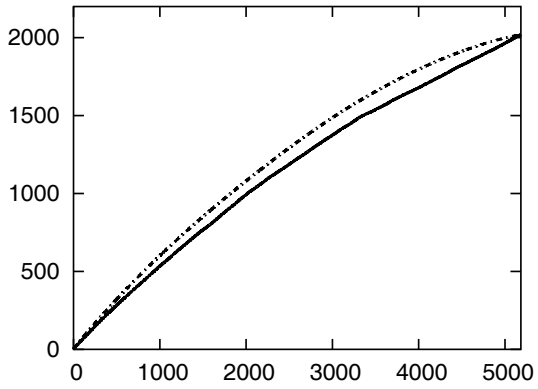

Fig. 3. The effect of using (solid line) or not using (dotted line) the negative examples in the feature selection process (see text).

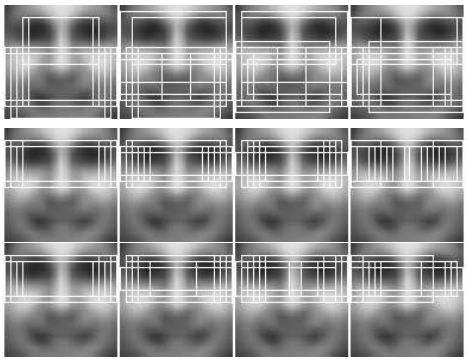

Fig. 4. Support of the top 40 discriminative features for horizontal ranklets (see text) obtained from: Top: positive examples in the symmetrized training; Middle: positive and negative examples in the symmetrized training; Bottom: positive and negative examples in the original training.

figure displays the support of the top 40 features sorted using also the negative examples. By inspection it is clear that the *eye area*, very reach of relevant features, is captured without using negative examples. When only positive examples are used, several ranklets of the *mouth area* appear to be also relevant. Interestingly, the most discriminative features represent edges or consistently untextured, uniform regions. This corresponds to the choice that is manually performed in [7]. The bottom row of Fig. 4 shows the 40 top discrinimative features obtained repeating the whole procedure *without* enforcing vertical symmetry through mirroring. The similarity between the middle and bottom row indicates that the original training is already characterized by a high degree of vertical symmetry.

Independence We now discuss the third and final stage in which we test for feature independence. After a coarse thresholding of all the computed features we built two graphs

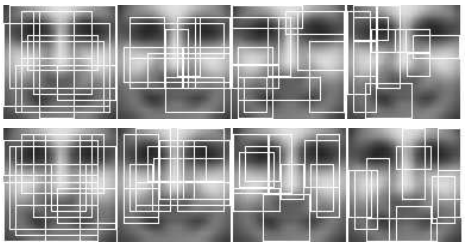

Fig. 5. Support of the top 40 discriminative features for horizontal ranklets *after* Spearman's test.

with 4000 nodes each, with nodes corresponding to the horizontal and vertical ranklets respectively. Pairwise Spearman's tests were run to test for independence with $\alpha = 0.5$. We ended up finding maximal cliques with 54 and 84 nodes of horizontal and vertical ranklets respectively. The support of the 40 top features (for horizontal ranklets) sorted by saliency is shown in the top row Fig. 5 (10 in each image from left to right). The bottom row of the same figure displays the support of the top 40 features repeating the whole procedure using also the negative examples for sorting the features. By inspection it can easily be appreciated that the proposed procedure, due to the pairwise independency requirement, appear to select features the support of which is more evenly distributed across the image. The influence of the independency constraint is clearer by looking at the eigenvalues of the covariance matrices of the descriptive features and of the independent subset. In figure 6 is shown that the eigenvalues corresponding to post-independency features descend more slowly than those relative to salient features.

5. CONCLUSIONS

We presented a data–driven approach to feature selection based on nonparametric statistics. The main points are a novel saliency measure and testing for independence using Spearman's statistics. These are supported both by theoretical considerations and by an empirical assessment over the MIT–CBCL database. Experiments confirm that negative examples introduce only marginal improvements, if any. Therefore, the proposed approach appears to be effective and well-suited for concept learning.

Acknowledgements

We thank L. Rosasco for useful discussions. This work is supported by the INFM Advanced Research Project MAIA, the FIRB Project ASTA^2 RBAU01877P, and the EC IST Programme under the PASCAL Network of Excellence IST-2002-506778.

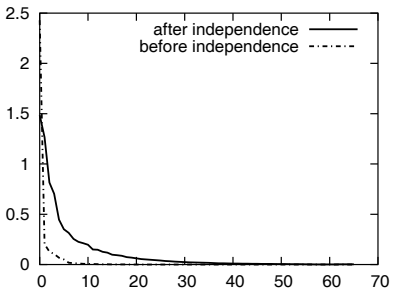

Fig. 6. Eigenvalues of the covariance matrix of all the descriptive features (dashed line) and of those in the pairwise independent clique (solid line).

6. REFERENCES

[1] A. Webb, *Statistical Pattern Recognition*, Oxford University Press, 1999.

[2] "Special issue on variable and feature selection," Journal of Machine Learning Reseach, march 2003.

[3] P. Viola and M. Jones, "Robust real-time object detection," in *II Int. Workshop on Stat. and Comput. Theories of Vision - modeling, learning, computing and sampling*, 2001.

[4] M. Alvira and R. Rifkin, "An empirical comparison of SNoW and SVMs for face detection," Tech. Rep. AI Memo 2001-004 – CBCL Memo 193, MIT, January 2001, http://www.ai.mit.edu/projects/cbcl.

[5] V. N. Vapnik, *The nature of statistical learning theory*, Springer–Verlag, 1995.

[6] T. Kadir and M. Brady, "Saliency, scale and image description," *Int. J. Comput. Vision*, vol. 45, no. 2, pp. 83–105, 2001.

[7] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *Internatonal Journal of Computer Vision*, vol. 38, no. 1, pp. 15–33, 2000.

[8] E. L. Lehmann, *Nonparametrics: Statistical methods based on ranks*, Holden-Day, 1975.

[9] MIT Center for Biological and Computational Learning, "CBCL face database no. 1," http://www.ai.mit.edu/projects/cbcl, 2000.

[10] F. Smeraldi, "Ranklets: orientation selective non-parametric features applied to face detection," in *Proc. of the 16th ICPR, Quebec QC*, August 2002, vol. 3, pp. 379–382.

[11] E. Franceschi, F. Odone, F. Smeraldi, and A. Verri, "Finding objects with hypothesis testing," in *Proceedings of International Workshop on Learning and Adaptable Visual Systems (LAVS)*, 2004.

[12] H. Schneiderman and T. Kanade, "A statistical method for 3D object recognition applied to faces and cars," in *Proceedings of CVPR*. 2000, pp. 746–751, IEEE.

[13] I. Guyon and E. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, pp. 1157–1182, March 2003.