Short Papers

On the Selection and Classification of Independent Features

Marco Bressan and Jordi Vitrià

Abstract—This paper is focused on the problems of feature selection and classification when classes are modeled by statistically independent features. We show that, under the assumption of class-conditional independence, the class separability measure of divergence is greatly simplified, becoming a sum of unidimensional divergences, providing a feature selection criterion where no exhaustive search is required. Since the hypothesis of independence is infrequently met in practice, we also provide a framework making use of class-conditional Independent Component Analyzers where this assumption can be held on stronger grounds. Divergence and the Bayes decision scheme are adapted to this class-conditional representation. An algorithm that integrates the proposed representation, feature selection technique, and classifier is presented. Experiments on artificial, benchmark, and real-world data illustrate our technique and evaluate its performance.

Index Terms—Feature selection, divergence, independent component analysis, naive Bayes.

1 Introduction

In the context of statistical pattern classification, the assumption of class-conditional independence greatly simplifies estimation through the marginalization of class-conditional densities. In this paper, we explore the consequences of class-conditional independence in the field of feature selection. It can be seen that the statistical criterion of divergence [15] is greatly simplified under this premise. Since class-conditional independence is not a frequently met assumption in practice, we also make use of Independent Component Analysis in order to obtain a (class-conditional) representation where this hypothesis can be held on stronger grounds.

In Section 2, we expose the concepts of independence and conditional independence, stressing the fact that the first does not imply the latter. We then detail the effect of class-conditional independence on the Bayesian classification scheme and on the measure of divergence. The consequence of this marginalization on the Bayesian decision scheme is well-known, resulting in the naive Bayes classifier. Less known is the fact that marginalized densities transform divergence into a sum of unidimensional divergence. The selection of optimal features for this discriminability criterion and, under these conditions, is a nonexhaustive procedure.

In practice, the most frequent situation is to encounter features with high levels of dependence between one another, so direct use of our proposed criterion or of the naive Bayes rule is unjustified and might lead to large error rates. Section 3 introduces Independent Component Analysis (ICA) and explains the way it can be employed, through class-conditional representations, to force independence on the random vector representing a certain class. Under this representation and certain assumptions, independence is met so both naive Bayes and the proposed feature selection criterion can be used. Nevertheless, the nature of our representation requires to adapt both approaches. Naive Bayes is adapted by making use of the change of variables theorem and we observe that, by interpreting divergence in terms of expected log-likelihood ratios, we can formulate a class-conditional divergence-based feature selection criterion. Finally, Section 3.4 presents the training and test algorithms that integrate the proposed representation, feature selection technique, and classifier. Several experiments were performed showing the performance of our method. A first experiment, over an artificial data set [21], shows that, when class-conditional independence is present, divergence is a robust approach that needs no a priori knowledge on class-distributions. A second experiment is performed on the Letter Image Recognition Data from the UCI Repository [2]. This experiment illustrates the importance of the independence assumption when we use the simplified version of divergence and apply a naive Bayes Classifier. For this problem and classifier, our local approach achieves maximum classification results for all possible feature subsets. Finally, a third experiment was performed using a total of $40,000$ 512-dimensional color histograms corresponding to 10 different classes extracted from 948 images corresponding to the Corel database [7]. In this experiment, we show how our method outperforms other usual methods even for very low dimensions.

2 Conditional Independence

Let X and Y be random variables taking values in Ω. And, let $p(x,y)$, $p(x)$, $p(y)$, and $p(x|y)$ be, respectively, the joint density of (X,Y), the marginal densities of X and Y, and the conditional density of X given $Y=y$. We say that X and Y are independent $p(x,y)=p(x)p(y)$ or, equivalently, if $p(x|y)=p(x)$ [8]. It proves useful to understand independence from the following statement derived from the last equality: Two variables are independent when the value one variable takes gives us no knowledge on the value of the other variable. The definition of independence can be extended to the multivariate case $(X_1,\ldots,X_N)$ as $p(\boldsymbol{x})=p(x_1)\ldots p(x_N)$. Conditional independence is defined as a natural extension of these definitions through the incorporation of the conditional operator: $p(x,y|z)=p(x|z)p(y|z)$ and, equivalently, $p(x|y,z)=p(x|z)$.

A frequent mistake is to think that global independence implies conditional independence, Simpson's paradox [20] probably being the most well-known counterexample. The falseness of this implication can also be visualized considering random variables (X,Y) with uniform distribution in the square $\Omega=[0,1]\times[0,1]$ and Z, the random variable defined as 1 for the set $\{(x,y)\in\Omega, x>y\}$ and 0 otherwise. It is clear in this case that, given Z, knowledge on the value of, for instance, X provides information on Y: It should be greater or less than X, depending on the value of Z.

The case in which class-conditional independence is encountered has interesting consequences in the field of statistical pattern classification. Given K classes in $\Omega=\{C_1,\ldots C_K\}$ and a set of features represented by an N-dimensional random vector $\boldsymbol{x}=(x_1,\ldots,x_N)$, the Maximum A Posteriori (MAP) and the Maximum Likelihood (ML) solutions both make use of the class-conditional densities $p(\boldsymbol{x}|C_k)$. If equiprobable priors are considered and these densities assumed independent, the Bayes rule provides the ML solution commonly known as naive Bayes rule [10],

$$C_{Naive}=\arg\max_{k=1\ldots K}\prod_{n=1}^{N}P(x_n|C_k). \tag{1}$$

Thanks to the marginalization that takes place in the density estimation, the naive Bayes classifier is simple, effective, and fast. It has been applied with success in several pattern recognition tasks [22], [17]. Its statistical nature implies interesting theoretic and predictive properties and, if the conditional independence assumption holds and the univariate densities are properly estimated, no other classifier can outperform naive Bayes in terms of misclassification probability.

- *The authors are with the Centre de Visió per Computador (CVC), Departament Informàtica, Universitat Autònoma de Barcelona, Bellaterra, Spain, 08193. E-mail: {marco, jordi}@cvc.uab.es.*

Manuscript received 9 Jan. 2002; revised 9 Oct. 2002; accepted 13 Jan. 2003.
Recommended for acceptance by P. Meer.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 115664.

0162-8828/03/$17.00 © 2003 IEEE Published by the IEEE Computer Society

2.1 Divergence and Conditional Independence

The problem of feature selection for classification can be stated as, given a set of features representing our data, select a subset such that working with the reduced set proves advantageous for classification. Goodness of a feature subset is measured through a criterion usually based on class separability. In most of the cases, evaluation of such criteria requires costly and completely new computations for each possible subset, turning feature subset selection into a combinatorial problem. Statistical class separability measures take into account the distance among the conditional distributions. These measures require an estimate of the conditional densities. Problems with this estimation are overcome using parametric or semiparametric techniques. Standard feature selection criteria such as the Bhattacharyya and Mahalanobis distances, Gaussian divergence, or Fisher ratio take this approach [11]. Divergence, instead, makes no prior assumption on the class-conditional densities. The drawback derived from this lack of assumptions is the eventual inaccuracy of the estimations, particularly in the presence of high dimensional data. This problem is overcome if conditional independence can be safely assumed. We will also see that this hypothesis allows the selection of a feature subset of any size without the need for an exhaustive search. Additionally, divergence has a straightforward interpretation from both an information theoretic and a probabilistic framework where it is directly related with the Bayes error [14].

A commonly used distance measure for (class-conditional) densities, for its connection with information theory, is the Kullback-Leibler distance [15],

$$KL(C_i, C_j) = \int_\Omega p(\boldsymbol{x}|C_i) \log \frac{p(\boldsymbol{x}|C_i)}{p(\boldsymbol{x}|C_j)} d\mathbf{x}, \quad (2)$$

where $1 \leq i, j \leq K$. The asymmetry of Kullback-Leibler motivates the symmetric measure of divergence, long ago used for feature selection [18], defined as

$$\mathcal{D}_{ij} = \mathcal{D}(C_i, C_j) = KL(C_i, C_j) + KL(C_j, C_i). \quad (3)$$

Besides being symmetric, divergence is zero between a distribution and itself, always positive and monotonic on the number of features. A paradox arises directly from the property of monotonicity: If including more features only increases class separability, why should we decide to remove features in the first place? The fact that the Bayesian classifier is also monotonic on the number of features should only add confusion to this issue. As will be confirmed in the experiments, unmet assumptions (independence) and errors in the density estimation break this monotonicity, so subset selection is justified, especially when considering the close relationship between estimation error and dimensionality (curse of dimensionality). Equally important is the dramatic increase feature selection can have in the speed of a pattern recognition system, at a very low cost in classification accuracy.

When the condition of class-conditional independence is met, it can be seen that divergence is additive on the features by introducing the definition of conditional independence in (3),

$$\mathcal{D}_{ij} = \sum_{n=1}^{N} \mathcal{D}_{ij}^n, \quad (4)$$

where $\mathcal{D}_{ij}^n$ indicates the marginal divergence for the nth feature. For this particular case, the property of monotonicity is evident. Also, unidimensional density estimation can be performed and the computation of divergence for a feature subset $S \subseteq \{1, \ldots, N\}$ (noted by $\mathcal{D}_{ij}^S$) is straightforward. We can also observe that

$$(n_1 \notin S, n_2 \notin S) \wedge (\mathcal{D}_{ij}^{n_1} \leq \mathcal{D}_{ij}^{n_2}) \Rightarrow (\mathcal{D}_{ij}^{S \cup n_1} \leq \mathcal{D}_{ij}^{S \cup n_1}). \quad (5)$$

This property of order suggests that, at least for the two class case, the best feature subset is the one that contains the features with maximum marginal divergence and, thus, provides a very simple rule for feature selection without involving any search procedure: Given a feature subset size, preserve only those features with maximal marginal divergence.

Although divergence only provides a measure for the distance between two classes there are several ways of extending it to the multiclass case, providing an effective feature selection criterion. The most common approach is taking the average over all class pairs

$$\mathcal{D}_A^n = \frac{2}{K(K-1)} \sum_{i=1}^{K} \sum_{j>i} \mathcal{D}_{ij}^n. \quad (6)$$

$\mathcal{D}_A^n$ represents the average divergence present in feature n. This approach is simple and preserves the exposed property of order for feature subsets. Average divergence can also be related to the Bayes error [9], justifying this choice. In the experiments and unless stated otherwise, the $\mathcal{D}_A$ criterion was used.

3 A Class-Conditional Representation

In the previous section, we have shown how conditional independence, through the marginalization of the densities, can simplify both Bayesian decision and feature subset selection when the divergence criterion is used. In practice, this assumption is rarely met. We have also mentioned that global independence does not necessarily imply class-conditional independence. For the case in which class-conditional independence is not true, we now introduce a supervised representation where this assumption can be held on stronger grounds.

3.1 Independent Component Analysis

The Independent Component Analysis (ICA) of an N-dimensional random vector is the linear transform which minimizes the statistical dependence between its components. This representation in terms of independence proves useful in an important number of applications such as data analysis and compression, blind source separation, blind deconvolution, denoising, etc. [1], [16], [12]. The basic ICA model [12] can be expressed as

$$\boldsymbol{W}(\boldsymbol{x} - \overline{\boldsymbol{x}}) = \boldsymbol{s}, \quad (7)$$

where $\boldsymbol{x}$ corresponds to the random vector representing our data, $\overline{\boldsymbol{x}}$ its mean, $\boldsymbol{s}$ is the random vector of *independent components* with dimension $M \leq N$, and $\boldsymbol{W}$ is called the *filter* or *projection matrix*. This model is frequently presented in terms of $\boldsymbol{A}$, the pseudoinverse of $\boldsymbol{W}$, called the *mixture matrix*. Names are derived from the original blind source separation application of ICA. If the components of vector $\boldsymbol{s}$ are independent, at most one is Gaussian and its densities are not reduced to a point-like mass, it can be seen that $\boldsymbol{W}$ is completely determined [6]. Main drawbacks of ICA to be taken into account in practice are its linear nature and the fact, a large number of samples are required for robust estimation, particularly for high-dimensional data.

In practice, the estimation of the filter matrix $\boldsymbol{W}$ and, thus, the independent components can be performed through the optimization of several objective functions such as likelihood, kurtosis, information flow, or mutual information. Though several algorithms have been tested, the method employed in this article is the one known as FastICA. This method attempts to maximize non-Gaussianity through the search of maximum negentropy directions. Negentropy, a normalized version of entropy, is a robust measure of non-Gaussianity which can be accurately approximated through the expectation of general nonquadratic functions. Optimization is achieved through a gradient descent algorithm speeded up by an approximative Newton iteration scheme. This method is explained in [12] where it is also related with other ICA estimation approaches, such as maximum likelihood estimation, mutual information minimization or tensorial-based methods.

3.2 Class-Conditional Independent Component Analysis (CC-ICA)

As mentioned, global feature independence is not sufficient for conditional independence. In [3], a class-conditional ICA (CC-ICA) model is introduced that, through class-conditional representations, ensures conditional independence. This scheme was successfully applied in the framework of classification for object recognition. The basic CC-ICA model is estimated from the training set for each class. If $\boldsymbol{W}_k$ and $\boldsymbol{s}_k$ are the projection matrix and the independent components for class C_k with dimensions $M_k \times N$ and M_k, respectively, then, from (7)

$$\boldsymbol{s}_k = \boldsymbol{W}_k(\boldsymbol{x} - \overline{\boldsymbol{x}}_k), \tag{8}$$

where $\boldsymbol{x} \in C_k$ and $\overline{\boldsymbol{x}}_k$ is the class mean, estimated from the training set. Most ICA methods require, or at least advise, data whitening as preprocessing. Since some simple denoising is also recommended, dimensionality reduction and whitening through PCA is very common practice as a preprocessing stage for ICA [4], [12], one of its advantages being that, for whitened data, the unmixing matrix should be orthogonal. In this case, $\boldsymbol{W}_k = \boldsymbol{B}_k\boldsymbol{D}_{\boldsymbol{k}}{}^{-1/2}\boldsymbol{E}_k$, where $\boldsymbol{E}_k$ is the $M \times N$ PCA eigenvector matrix, $\boldsymbol{D}_k$ the $M \times M$ diagonal matrix with the corresponding eigenvalues, and $\boldsymbol{B}_k$ the $M \times M$ ICA projection (unmixing) matrix for the whitened data. In this case $\boldsymbol{v} = \boldsymbol{E}_k(\boldsymbol{x} - \boldsymbol{x}_k)$ are the principal components such that $\boldsymbol{s} = \boldsymbol{B}_k\boldsymbol{D}_k^{-1/2}\boldsymbol{v}$. Assuming the class-conditional representation actually provides independent components, we have that the class-conditional probability in transformed space, noted as $p_k(\boldsymbol{s}) \overset{def}{=} p(\boldsymbol{s}_k)$, can now be expressed in terms of unidimensional densities,

$$p(\boldsymbol{v}|C_k) = \nu_k p_k(\boldsymbol{s}) = \nu_k \prod_{m=1}^{M_k} p_k(s_m), \tag{9}$$

with ν_k a normalizing constant. Actually, from the change of variables rule,

$$\nu_k = |\det(\boldsymbol{B}_k\boldsymbol{D}_k^{-1/2})| = |\det(\boldsymbol{D}_k^{-1/2})| = \prod_m \frac{1}{\sqrt{\lambda^m}},$$

so this constant can be estimated together with the CC-ICA models.

From now on we will say Ω is an *ICA Space* if all its class-conditional distributions correspond to independent variables and thus can be expressed as a product of unidimensional distributions.

If independence is not known in advance and a CC-ICA representation is used then, by using whitened data, replacing (9) in (1), and using log-likelihoods [3],

$$C_{Naive} = \arg\max_{k=1\dots K} \sum_{m=1}^{M_k} \log p_k(s^m) + \log(\nu_k). \tag{10}$$

If a subset of features has been selected for class C_k, then the sum in (10) is only performed on the corresponding features. In addition, several ICA estimation algorithms, such as those based on a maximum likelihood approach, already provide the class-conditional marginal densities in the estimation [5], [19]. If this is not the case for our choice of algorithm, the class-conditional marginal densities $p_k(s^m)$ can be estimated using classical density estimation techniques such as Gaussian mixture models, Laplace mixture models, or nonparametric kernel methods. It can be seen that minimization of mutual information results in strongly non-Gaussian distributions. This a priori knowledge can be used to restrict possible densities to particular density families such as the generalized Gaussian. Being unidimensional, estimation is fast and straightforward.

3.3 Divergence for CC-ICA

If our features do not lie in an ICA space, through (8) we can have K linear representations, each one providing class-conditional independence. With this approach, the selection of a single feature for the whole ICA space involves the selection of possibly distinct single features belonging to different representations. Divergence needs adapting to result useful in these class-conditional representations. After some algebraic operations detailed in the Appendix, we obtain the following expression for divergence,

$$\mathcal{D}_{ij} = \sum_{m=1}^{M_i} \mathcal{B}_{ij}^m + \sum_{m=1}^{M_j} \mathcal{B}_{ji}^m, \tag{11}$$

where $\mathcal{B}_{ij}^m$ can be approximated by

$$\mathcal{B}_{ij}^m \approx \frac{1}{\#C_i} \sum_{\boldsymbol{x}\in C^i} \log p^i(\boldsymbol{w}_i^{mT}(\boldsymbol{x} - \overline{\boldsymbol{x}}_i)) - \frac{1}{\#C_j} \sum_{\boldsymbol{x}\in C^j} \log p^i(\boldsymbol{w}_i^{mT}(\boldsymbol{x} - \overline{\boldsymbol{x}}_i)), \tag{12}$$

with $\boldsymbol{w}_i^m$ the vector indicating the mth row of $\boldsymbol{W}_i$ and $\overline{\boldsymbol{x}}_i$ the class mean. Notice that $\mathcal{B}_{ij}^m$ measures the separation of classes C_i and C_j in the mth component of the representation learned for class C_i. Given a feature subset size, divergence is maximized by preserving for C_i, those features maximizing $\mathcal{B}_{ij}$ and, for C_j, those maximizing $\mathcal{B}_{ji}$. This will cause different feature subsets on each class-conditional representation, meaning that, while certain features might be appropriate for separating class C_i from class C_j in the i^{th} representation, possibly distinct features will separate class C_j from class C_i in the j^{th} representation.

Extension to the multiclass case can be performed in the same fashion as with divergence, with one of the indexes fixed for the representation. The average class-conditional divergence for feature m in C^i is

$$\{\mathcal{B}_i^m\}_A = \frac{1}{K-1} \sum_{j=1, j\neq i}^{K} \mathcal{B}_{ij}^m. \tag{13}$$

3.4 The Algorithm

Fig. 1 details the learning algorithm for the case in which a class-conditional representation is chosen. If class-conditional independence is assumed, then the algorithm is greatly simplified. Steps 2 and 3 can be skipped and divergence estimated from Step 6. In this case, the divergence is once again symmetrical resulting in a single set of discriminant features for all classes.

Fig. 2 illustrates the way classification is performed. Once again, a previous assumption of class-conditional independence, greatly simplifies the algorithm. In this case, projection is unnecessary and only the marginal class-conditional densities corresponding to the selected features are evaluated and employed for computing the class-conditional probability.

4 Experiments

A first experiment is performed on the artificial two-class example Trunk used to illustrate the curse of dimensionality [21]. The two classes have multivariate normal 20-dimensional distributions with covariance given by the identity matrix and means $\mu_1 = [1/\sqrt{1}, 1/\sqrt{2}, \dots, 1/\sqrt{20}]$, $\mu_2 = -\mu_1$. In a recent survey on feature selection [13], Jain and Zongker propose this example to investigate the quality of certain feature subsets considering that the optimal $d\text{-feature}$ subset is known in advance: the first d features. They propose a measure of average quality for the feature selection criterion varying the number of training patterns per class and averaging the results of five artificially generated data sets on every possible $d\text{-feature}$ subset. The maximum possible value for this average quality is one, meaning that the 20 possible feature subsets were the optimal subset for the five data sets.

```
TRAINING STAGE
for each class k=1...K
    Estimate ICA parameters $\overline{\boldsymbol{x}}_k$, $\boldsymbol{W}_k$ and $\nu_k$.
    if marginal densities are unavailable, project class samples $\boldsymbol{s} = p_k(\boldsymbol{w}_k^m(\boldsymbol{x} - \overline{\boldsymbol{x}}_k))$.
    Estimate $p_k(s^m)$
end loop
From estimated marginal densities compute $\mathcal{B}_{ij}^m$, using (12)
for each class k=1...K
    Select the F most discriminant features from $\{\mathcal{B}_k^m\}_A$, indexed by $I_k = \{m_1, \ldots, m_F\}$
end loop
```

Fig. 1. Algorithm for learning class-conditional representations and discriminant.

Notice that this data set is actually an ICA space: The class-conditional densities are uncorrelated Gaussians, thus independent. So, there is no need to transform the data. We use (3) to compute the unidimensional divergence values. In Fig. 3, we reproduce the results in [13] using the Mahalanobis distance between means as a criterion and the optimal branch and bound feature subset selection algorithm. We also plot the results of our method, estimating the marginal densities with a 2-Gaussian Mixture Model (no prior knowledge of the data assumed) and with a Gaussian with unknown mean and deviation (Gaussian data assumed). For the latter, divergence has a closed form [11]. From Fig. 3, we observe divergence is a fairly robust criterion with performance above Jain's criterion. Gaussian Mixture Models do not perform well when the number of samples is similar to the dimensionality but soon recover, meaning that we can do without the prior knowledge on the data distribution without seriously affecting the results.

A second experiment is performed on the Letter Image Recognition Data [2]. Each instance each of the $20,000$ images within this database represents a capital typewritten letter in one of 20 fonts. Each letter is represented using 16 integer valued features corresponding to statistical moments and edge counts. Training is done on the first $16,000$ instances and test on the final $4,000$. There are approximately 615 samples per class in the training set. In this case, feature independence cannot be assumed. Fig. 4 illustrates the results of the naive Bayes Classifier for different representations and feature subsets. The divergence feature selection criterion was used for ICA (a global ICA representation), CC-ICA, and ORIG (the original representation), while, for PCA, features were selected as ordered by the representation. The results of quadratic classification (Gaussian maximum likelihood) on PCA were also included as a reference.

```
TEST STAGE
for each class k=1...K
    Project test sample in CC-ICA space, $\boldsymbol{s} = \boldsymbol{W}_k(\boldsymbol{x}_{Test} - \boldsymbol{m}_k)$
    for each feature f=1...F
        Evaluate $p^k(s_{m_f}), m_f \in I_k$
    end loop
Compute class-conditional probability using Eq.(10).
end loop
```

Fig. 2. Algorithm for classifying sample x_{Test} using the scheme learned in Fig. 1.

We can observe in Fig. 4 the importance of the independence assumption when using both naive Bayes and the divergence criterium. The CC-ICA representation, by seeking this independence, achieves much better results than all the other implementations. On this database, we also tried naive Bayes on $100,000$ random 8-feature combinations for each class, with the result that no combination achieved our classification results (83.17 percent).

A third experiment was performed in order to illustrate the performance of independent feature selection on high-dimensional data. In this case local color histograms (dimension = 512) were extracted from different representative regions of 948 images belonging to the Corel Database [7]. The regions belong to 10 different classes corresponding to clouds, grass, ice, leaves, rocky mountains, sand, sky, snow mountains, trees, and water. A total of $40,000$ samples (histograms) were extracted, of which $30,000$ were used for training and the remaining for test. The number of class samples was equal among both training and test

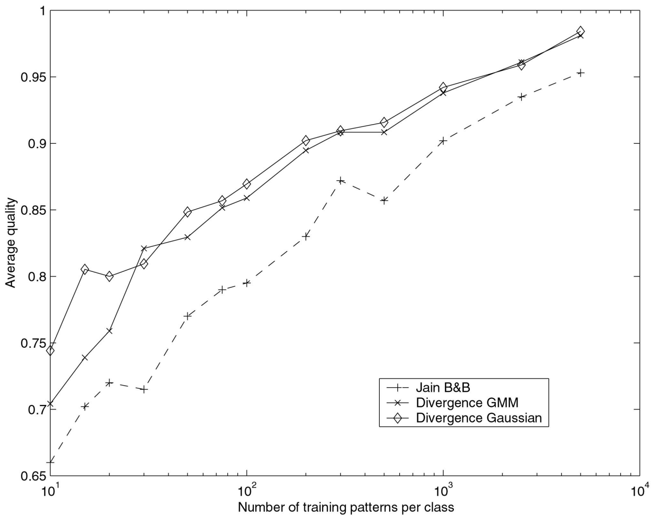

Fig. 3. Quality of selected feature subsets as a function of the size of the training data.

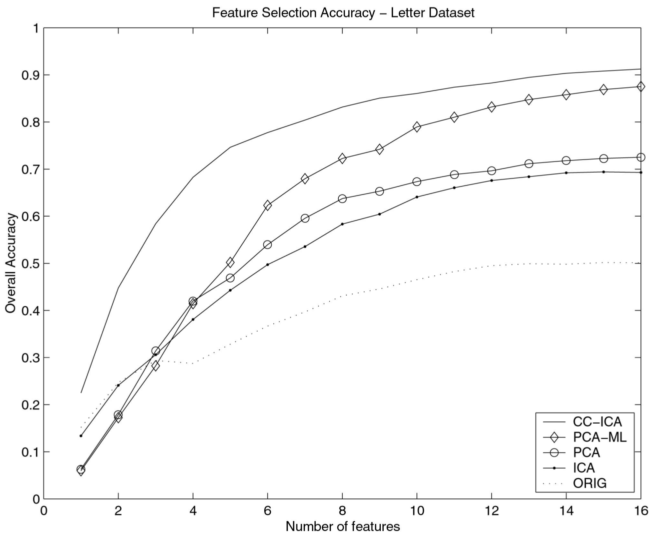

Fig. 4. Naive Bayes classifier performance on different representations and feature selection criteria. The importance of the independence assumption on naive Bayes performance is observed. Maximum likelihood on a PCA representation is added as a reference.

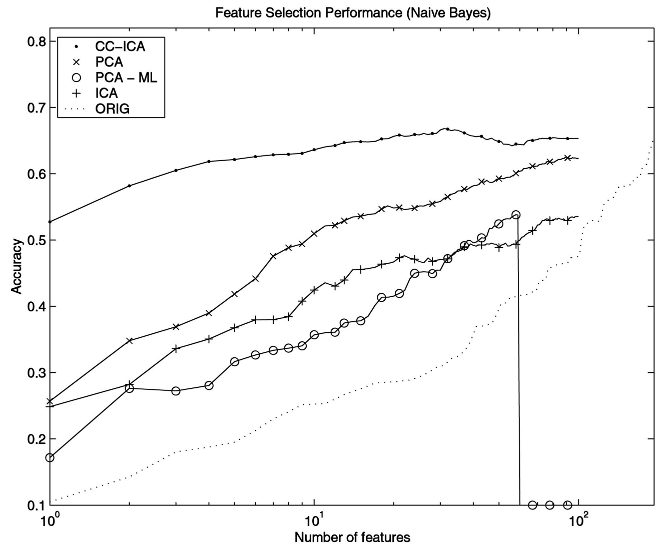

Fig. 5. Accuracies for the Corel database histogram-based naive Bayes classification.

sets. In all cases, the true class dimensionality was considerably below 512: classes have very restricted color variation. So, the CC-ICA was performed after PCA dimensionality reduction (preserving 98.5 percent of the variance) and whitening. The final dimensions for the class-conditional representations varied between 42 (ice) and 85 (leaves). Fig. 5 accounts for the results of the same experiment performed on the Letter database: naive Bayes classification under ICA, CC-ICA, PCA, and the original representation and maximum likelihood on PCA (PCA-ML) as a reference. As with the letter database, the CC-ICA scheme (10) outperforms the global representations notoriously for a low number of features. The performance of this method drops after 40 features, precisely the dimension of the least dimensional class. The answer to why PCA outperforms ICA for naive Bayes has to be found in the fact that independence does not imply conditional independence and the consequence this has over our feature selection criterion and the naive Bayes classifier. ORIG represents a K-NN classification on the original data using a mean Bhattacharyya distance with forward search for feature selection. Notice that PCA-ML drops to zero once the covariance matrix for a certain class becomes rank-deficient.

Accuracy results in (5) are poor: below 70 percent in the best case. This is due to the high confusion the class signatures (color). This is the reason why the average match percentile (AMP) is frequently used to evaluate the results of this type of experiment, where a rank in the classification proves sufficient. The AMP for our best case (31 features) is 92.64 percent.

5 Conclusions

Conditionally independent features greatly simplify pattern classification and feature selection problems. For this last case, the separability measure of divergence is reduced to a sum of unidimensional divergences and the optimal feature subset of any cardinality can be found without involving an exhaustive search. Since conditional independence is not usually encountered on real-world data sets, we provide a context, class-conditional independent component analysis, where it can be assumed on stronger grounds. Divergence is adapted to this context. Even though no assumption is made on the classifier, the natural choice for our situation is naive Bayes. This classifier is also adapted to our class-conditional representation.

Three different experiments were performed. These experiments illustrate the robustness and performance of the introduced techniques on artificial, real-world benchmarked, and high-dimensional data, as well as comparing the results with alternative approaches.

The CC-ICA representation is still a linear approximation of a (possibly) nonlinear problem, so actual independence is seldom achieved and the assumptions made for our method are weakened. Current research on nonlinear or overcomplete independent component analysis could eventually provide a framework where the exposed theory can be even more effectively put into practice. Another major inconvenience of our approach is the fact that independent component analysis learning requires a large number of samples, particularly on high-dimensional data, and, in our case, we need a large number of samples per class. When this condition is not met, we cannot find trustable class conditional representations nor make any assumption on the divergence.

A promising line of research arises when the maximum likelihood rule is replaced by pairwise classifiers, which are easily adapted to the CC-ICA situation and avoid using average divergence.

Appendix

The expression for divergence under class-conditional representations can be derived from the following operations. Since the Kullback-Leibler distance can be understood as the class-conditional expectation of the log-likelihood ratio (expected density overlap), (3) can be rewritten as in terms of conditional expectations

$$\mathcal{D}_{ij} = E_i\{\log p(\boldsymbol{x}|C_i) - \log p(\boldsymbol{x}|C_j)\} + E_j\{\log p(\boldsymbol{x}|C_j) - \log p(\boldsymbol{x}|C_i)\}. \tag{14}$$

Rearranging the terms in this sum and replacing by (9), we have

$$\begin{aligned}\mathcal{D}_{ij} = {} & E_i\left\{\sum_{m=1}^{M_i} \log p_i(\boldsymbol{w}_i^{mT}(\boldsymbol{x} - \overline{\boldsymbol{x}}_i)) + \log(\nu_i)\right\} \\ & - E_j\left\{\sum_{m=1}^{M_i} \log p_i(\boldsymbol{w}_i^{mT}(\boldsymbol{x} - \overline{\boldsymbol{x}}_i)) + \log(\nu_i)\right\} \\ & + E_j\left\{\sum_{m=1}^{M_j} \log p_j(\boldsymbol{w}_j^{mT}(\boldsymbol{x} - \overline{\boldsymbol{x}}_j)) + \log(\nu_j)\right\} \\ & - E_i\left\{\sum_{m=1}^{M_j} \log p_j(\boldsymbol{w}_j^{mT}(\boldsymbol{x} - \overline{\boldsymbol{x}}_j)) + \log(\nu_j)\right\},\end{aligned} \tag{15}$$

where $\boldsymbol{w}_k^m$ is the $m\mathrm{th}$ row of the filter matrix learnt for class C_k and $\overline{\boldsymbol{x}}_k$ the estimated class mean. Normalization constants are canceled (which, by the way, shows that divergence is invariant for invertible linear transformations) and the sum can be taken out of the expectation operators such that, if we define

$$\mathcal{B}_{ij}^m = \left(E_i\{\log p^i(\boldsymbol{w}_i^{mT}(\boldsymbol{x}-\overline{\boldsymbol{x}}_i))\} - E_j\{\log p^i(\boldsymbol{w}_i^{mT}(\boldsymbol{x}-\overline{\boldsymbol{x}}_i))\}\right), \quad (16)$$

we have (11). Equation (12) is obtained by approximating with the empirical expectation.

ACKNOWLEDGMENTS

This work was supported by MCYT grant TIC2000-0399-C02-01 and the Secretaria de Estado de Educacion, Universidades, Investigacion y Desarrollo from the Ministerio de Educacion y Cultura of Spain. The authors are also grateful to the reviewers for their careful reading and numerous suggestions.

REFERENCES

[1] A. Bell and T. Sejnowski, "The 'Independent Components' of Natural Scenes Are Edge Filters," *Neural Computation,* vol. 11, pp. 1739-1768, 1999.

[2] C. Blake and C. Merz, "UCI Repository of Machine Learning Databases," 1998.

[3] M. Bressan, D. Guillamet, and J. Vitria, "Using an ICA Representation of High Dimensional Data for Object Recognition and Classification," *Proc. IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 1004-1009, 2001.

[4] J. Cardoso and P. Comon, "Independent Component Analysis, a Survey of Some Algebraic Methods," *Proc. Int'l Symp. Circuits and Systems (ISCAS '96),* vol. 2, pp. 93-96, 1996.

[5] S. Choi, A. Cichocki, and S. Amari, "Flexible Independent Component Analysis," *J. VLSI Signal Processing,* vol. 26, nos. 1/2, pp. 25-38, Aug. 2000.

[6] P. Comon, "Independent Component Analysis—a New Concept?" *Signal Processing,* vol. 36, pp. 287-314, 1994.

[7] *Corel Stock Photo Library,* Corel Corp., Ontario, Canada, 1990.

[8] A.P. Dawid, "Conditional Independence in Statistical Theory (with Discussion)," *J. Royal Statistical Soc., Ser. B,* vol. 41, pp. 1-31, 1979.

[9] H. Decell and J. Quirein, "An Iterative Approach to the Feature Selection Problem," *Proc. Purdue Univ. Conf. Machine Processing of Remotely Sensed Data,* vol. 1, pp. 3B1-3B12, 1972.

[10] R. Duda, P. Hart, and D. Stork, *Pattern Classication,* second ed. John Wiley & Sons, 2001.

[11] K. Fukunaga, *Introduction to Statistical Pattern Recognition,* second ed. Academic Press, 1990.

[12] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis.* John Wiley & Sons, 2001.

[13] A.K. Jain and D.E. Zongker, "Feature Selection: Evaluation, Application, and Small Sample Performance," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 2, pp. 153-158, Feb. 1997.

[14] T. Kailath, "The Divergence and Bhattacharyya Distance Measures in Signal Selection," *IEEE Trans. Comm. Technology,* vol. 15, no. 1, pp. 52-60, Feb. 1967

[15] S. Kullback, *Information Theory and Statistics.* John Wiley & Sons, 1968.

[16] T. Lee, M. Lewicki, and T. Seynowski, "A Mixture Models for Unsupervised Classification of Non-Gaussian Sources and Automatic Context Switching in Blind Signal Separation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 10, pp. 1-12, Oct. 2000.

[17] D. Lewis, "Naive Bayes at Forty: The Independence Assumption in Information Retrieval," *Proc. ECML-98, 10th European Conf. Machine Learning,* C. N'edellec and C. Rouveirol, eds., pp. 4-15, 1998.

[18] T. Marill and D. Green, "On the Effectiveness of Receptors in Recognition Systems," *IEEE Trans. Information Theory,* vol. 9, pp. 1-17, 1963.

[19] J. Miskin, "Ensemble Learning for Independent Component Analysis," PhD thesis, Selwyn College, Cambridge, Dec. 2000.

[20] E. Simpson, "The Interpretation of Interaction in Contingency Tables," *J. Royal Statistical Soc., Ser. B,* vol. 13, pp. 238-241, 1951.

[21] G. Trunk, "A Problem of Dimensionality: A Simple Example," *IEEE Trans. Pattern Analysis and Machine Intelligence* vol. 1, no. 3, pp. 306-307, July 1979.

[22] Y. Yang, S. Slattery, and R. Ghani, "A Study of Approaches to Hypertext Categorization," *J. Intelligent Information Systems,* 2002.

Moment Computation for Objects with Spline Curve Boundary

Stanislav Sheynin and
Alexander Tuzikov, *Member*, *IEEE*

Abstract—A new approach is proposed for computation of area and geometric moments for a plane object with a spline curve boundary. The explicit formulae are obtained for area and low order moment calculation. The complexity of calculation depends on the moment order, spline degree, and the number of control points used in spline representation. The formulae proposed use the advantage that the sequence of spline control points is cyclic. It allowed us to reduce substantially the number of summands in them. The formulae might be useful in different applications where it is necessary to perform measurements for shapes with a smooth boundary.

Index Terms—Area, moment, parametric curve, spline, explicit formulae.

1 INTRODUCTION

It is of interest for different applications to compute geometric moments of plane or volumetric objects. Moments of zero order define the object area and volume, respectively. The object centroid is computed using first order moments and the orientation (we mean axes of inertia)—from second order moments. It is well-known that geometric moments and moment invariants are very useful for recognition of objects and images [1], [2].

A geometric moment m_{pq} of order $p+q$ for a plane object P is defined as follows:

$$m_{pq}(P) = \iint_P x^p y^q dxdy. \quad (1)$$

The explicit formulae for moment computation of 2D *polygonal* objects were derived in [3]. These formulae were extended in [4] for 3D *polyhedral* objects and higher dimensional polytopes. Other related results can be found in [5], [6], [7], [8], [9].

Recently, new results were obtained for 2D and 3D objects with a boundary defined by *parametric curves* and *surfaces* [10], [11], [12], [13], [14]. These results are due to the fact that some parametric representations of a curve or a surface are allowed to compute moments directly. The complexity of computation, in this case, depends on the order of the curve/surface applied and the moment order to be computed. The formulae for area computation of objects bounded by Bézier and *B*-spline curves were proposed in [14]. They are based on computation of the signed area of a sector between the curve and the coordinate origin. A similar approach was presented in [10] for computation of areas and volumes. Computation of volume and moments for cubic patches was discussed and evaluated in [11]. The results presented in [10], [11] were further extended in [13]. For the case when the object is bounded by a set of parametric *B*-spline surfaces the authors represent the moment formulae as multilinear forms of control points coordinates with some coefficients. These coefficients are integrals of *B*-spline basis functions. It is shown in [12] that moment computation is equivalent to applying a multidimensional filter on the curve coefficients followed by computing a scalar product. The authors discuss the properties and computation of filter kernel and present several examples for spline curves.

- *The authors are with the National Center of Informational Resources and Technologies, National Academy of Sciences of Belarus, Akademicheskaja 25, 220072 Minsk, Belarus. E-mail: {sheynin, tuzikov}@mpen.bas-net.by.*

Manuscript received 23 Aug. 2002; revised 3 Dec. 2002; accepted 13 Feb. 2003.
Recommended for acceptance by R. Sharma.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 117173.