Improving Naive Bayes using Class-Conditional ICA

Marco Bressan and Jordi Vitrià*

Centre de Visiò per Computador, Dept. Informàtica,
Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain.
Tel. +34 93 581 30 73 Fax. +34 93 581 16 70
[marco, jordi]@cvc.uab.es

Abstract. In the past years, Naive Bayes has experienced a renaissance in machine learning, particularly in the area of information retrieval. This classifier is based on the not always realistic assumption that class-conditional distributions can be factorized in the product of their marginal densities. On the other side, one of the most common ways of estimating the Independent Component Analysis (ICA) representation for a given random vector consists in minimizing the Kullback-Leibler distance between the joint density and the product of the marginal densities (mutual information). From this that ICA provides a representation where the independence assumption can be held on stronger grounds. In this paper we propose class-conditional ICA as a method that provides an adequate representation where Naive Bayes is the classifier of choice. Experiments on two public databases are performed in order to confirm this hypothesis.

1 Introduction

For years, the most common use of the Naive Bayes Classifier has been to appear in classification benchmarks outperformed by other, generally more recent, methods. Despite this fate, in the past few years this simple technique has emerged once again, basically due to its results both in performance and speed in the area of information retrieval and document categorization [25, 15]. Recent experiments on benchmark databases have also shown that Naive Bayes outperforms several standard classifiers even when the independence assumption is not met [6]. Additionally, the statistical nature of Naive Bayes implies interesting theoretic and predictive properties and, if the independence assumption is held and the univariate densities properly estimated, it is well known that no other classifier can outperform Naive Bayes in the sense of misclassification probability. Attemps to overcome the restriction imposed by the independence assumption have motivated attempts to relax this assumption via a modification of the classifier [23], feature extraction in order to hold the assumption on stronger grounds, and approaches to underestimate the independence assumption by showing it doesn´t make a big difference [6, 18]. This paper is clearly on the second line of research: we propose a class-conditional Independent Component Analysis Representation (CC-ICA) together with an appropriate feature selection procedure in order to obtain a representation where statistical independence is maximized. This representation has already proved successful in the area of object recognition and classification of high dimensional data [17].

* 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.

For multivariate random data, Independent Component Analysis (ICA) provides a linear representation where the projected components (usually called independent components) have maximized statistical independence. Additionally, in many problems the unidimensional densities of the independent components belong to restricted density families, such as supergaussian or subgaussian, exponential densities, etc. This prior knowledge allows a simple parametric approach to the estimations. The success of performing Naive Bayes over an ICA representation has an additional explanation. It has been shown that Naive Bayes performance improves under the presence of low-entropy distributions [18]. In many problems, this is precisely the type of distribution achieved by an ICA representation [1, 8, 10, 24].

In Section 2 we introduce the concept of independence and conditional independence, making some observations that justify the need for class-conditional representations. Here, we also introduce the Bayes Decision scheme and the particular case corresponding to the Naive Bayes classifier. Section 3 introduces Independent Component Analysis (ICA) and explains how it can be employed, through class-conditional representations, to force independence on the random vector representing a certain class. Naive Bayes is adapted to our representation. The problem of estimating the resulting marginal densities is also covered in this section. In Section 4, using the concept of divergence, briefly provides a scheme to select those features that preserve class separability from each representation in order to classify using a restricted set of features. Finally, experiments are performed on the Letter Image Recognition Data, from the UCI Repository [3] and the MNIST handwritten digits database [13]. These experiments illustrate the importance of the independence assumptions by applying the Naive Bayes classifiers to different representations and comparing the results. The representations used are the original representation, a class-conditional PCA representation (since PCA uncorrelates the data, under our line of reasoning, it can be understood as a second-order step towards independence) and finally our CC-ICA representation.

2 Independence and the Bayes Rule

Let X and Y be random variables and $p(x, y)$, $p(x)$, $p(y)$ 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 if any of the following two equivalent definitions hold [5]:

$$p(x, y) = p(x)p(y) \tag{1}$$

$$p(x|y) = p(x) \tag{2}$$

It proves useful to understand independence from the following statement derived from (2): Two variables are independent when the value one variable takes gives us no knowledge on the value of the other variable. For the multivariate case $(X_1, ..., X_N)$, independence can be defined by extending (1) as $p(x) = p(x_1)...p(x_N)$.

In the context of statistical classification, given K classes in a D-dimensional space $\Omega = \{C_1, ...C_K\}$ and a set of new features $\mathbf{x_T} = (x_1, ..., x_D)$ we wish to assign $\mathbf{x_T}$ to a particular class minimizing the probability of misclassification. It can be seen that the solution to this problem is to assign $\mathbf{x_T}$ to the class that maximizes the *posterior probability* $P(C_k|\mathbf{x_T}$. The Bayes rule formulates this probability in terms of the likelihood and the prior probabilities, which are simpler to estimate. This transformation, together with the assumption of

independence and equiprobable priors results on the Naive Bayes rule,

$$C_{Naive} = \arg \max_{k=1\ldots K} \prod_{d=1}^{D} P(x_d|C_k) \tag{3}$$

The simplification introduced in (3), transforming one D-dimensional problem into D 1-dimensional problems, is particularly useful in the presence of high dimensional data, where straightforward density estimation proves uneffective [19, 7]. Notice that class-conditional independence is required so a representation that achieves global independence of the data (sometimes referred to as "linked independence") is useless in this sense. A frequent mistake is to think that the independence of the features implies class-conditional independence, being Simpson's paradox [20] probably the most well known counterexample. The falseness of this implication can also be visualized considering a set of bivariate features (x, y) with uniform distribution in the square $\Omega = [0, 1] \times [0, 1]$ and classes $C_1 = \{(x, y) \in \Omega, x > y\}$, $C_2 = \overline{C_1}$. We conclude that in order to assume class-conditional independence, it is not enough to work in an independent feature space. For this particular case, in which class-conditional independence is not true, we now introduce a local representation where this assumption can be held on stronger grounds.

3 Independent Component Analysis

The 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. [2, 14, 24, 11]. Assuming the random vector we wish to represent through ICA has no noise, the ICA Model can be expressed as

$$\mathbf{W}(\mathbf{x} - \overline{\mathbf{x}}) = \mathbf{s} \tag{4}$$

where $\mathbf{x}$ corresponds to the random vector representing our data, $\overline{\mathbf{x}}$ its mean, $\mathbf{s}$ is the random vector of *independent components* with dimension $M \leq N$, and $\mathbf{W}$ is called the *filter* or *projection matrix*. This model is frequently presented in terms of $\mathbf{A}$, the pseudoinverse of $\mathbf{W}$, called the *mixture matrix*. Names are derived from the original blind source separation application of ICA. If the components of vector $\mathbf{s}$ are independent, at most one is Gaussian and its densities are not reduced to a point-like mass, it can be seen that $\mathbf{W}$ is completely determined [4].

In practice, the estimation of the filter matrix $\mathbf{W}$ and thus the independent components can be performed through the optimization of several objective functions such as likelihood, network entropy 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 minimize the mutual information by finding maximum negentropy directions, proving to be fast and efficient [11]. Since mutual information is the Kullback-Leibler difference between a distribution and its marginal densities, we would be obtaining a representation where the Naive Bayes rule best approximates the Bayes Rule in the sense of Kullback-Leibler.

As mentioned, global feature independence is not sufficient for conditional independence. In [17] we introduced a class-conditional ICA (CC-ICA) model that, through class-conditional representations, ensures conditional independence. This scheme was successfully applied in the framework of classification for object recognition. The CC-ICA model is estimated from

the training set for each class. If $\mathbf{W}_k$ and $\mathbf{s}_k$ are the projection matrix and the independent components for class C_k with dimensions $M_k \times N$ and M_k respectively, then from (4)

$$\mathbf{s}^k = \mathbf{W}^k(\mathbf{x} - \overline{\mathbf{x}^k}) \tag{5}$$

where $\mathbf{x} \in C_k$ and $\overline{\mathbf{x}^k}$ is the class mean, estimated from the training set. Assuming the class-conditional representation actually provides independent components, we have that the class-conditional probability noted as $p^k(\mathbf{s}) \stackrel{def}{=} p(\mathbf{s}^k)$ can now be expressed in terms of unidimensional densities,

$$p(\mathbf{x}|C_k) = \nu_k p^k(\mathbf{s}) = \nu_k \prod_{m=1}^{M_k} p^k(s_m) \tag{6}$$

with $\nu_k = (\int p^k(\mathbf{s})ds)^{-1}$, a normalizing constant. Plugging in (6) in (3) and applying logarithms, we obtain the Naive Bayes rule under a CC-ICA representation,

$$C_{Naive} = \arg \max_{k=1\ldots K} \sum_{l=1}^{L} \left(\sum_{m=1}^{M_k} \log P^k(s_{lm}) \right) + L\nu_k \tag{7}$$

In practice, classification is performed as follows. Representative features are extracted from the objects belonging to class C_k, conforming training set T_k. T_k is then used to estimate the ICA model and projected into this model. From the projected features, the M_k one dimensional densities are estimated, together with the normalization constants. If we have no prior information on these marginal distributions, nonparametric or semiparametric methods can be used in the one dimensional estimation. Given a test object, its representative features are projected on each class, and the class-conditional likelihoods calculated. The test object is assigned to the class with the highest probability.
As a matter of fact, the nonparametric density estimation is not even necessary. In the next subsection we will see how the ICA Model gives us a priori information that can be used in the estimation of the marginal densities of the independent components.

3.1 Marginal Density Estimation

It can be seen that progressive maximization of mutual information is achieved in the directions of maximum nongaussianity [9]. This results in independent components with strongly nongaussian distributions. A natural, but sensible measure for the nongaussianity of unimodal distributions is kurtosis. Kurtosis measures how "peaky" a distribution can be. In the range of unimodal distributions the uniform distribution can be considered the least "peaky", Dirac's delta its opposite. Kurtosis or the fourth-order cumulant is defined as $\kappa(s) = E(s^4) - 3$ for a zero mean, unit variance variable (true for the independent components). It refers to the concentration of the variable around zero. The higher the concentration the higher the kurtosis. It can be seen that, expressed in this way, kurtosis is zero for a standard gaussian variable. Negative kurtotic variables are referred to as *subgaussian* and positive kurtotic *supergaussian* or *sparse* variables. In our problems we can use kurtosis as an additional statistic for prior information on the distribution of the independent components.
A close relationship between sparsity and ICA has been pointed out [8, 10, 24, 1]. In our particular problem, as in many others, a very high sparsity is observed in the independent components. Classification can be interpreted in terms of sparsity in the following way. If a test feature belongs to a certain class, then a sparse representation for this feature will

be provided. This means that the independent components of the projected feature will be nearly zero for most values and consequently should have a large probability. Instead, if the feature does not belong to the class, it should activate several independent components at the same time and consequently have a low probability. This property is illustrated in Figure (3.1) for two class-conditional representations obtained in the experiments.

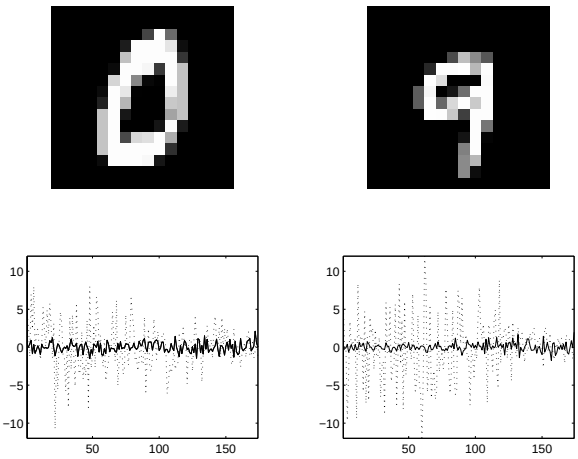

Fig. 1. For classes "0" and "9" of the MNIST database a representative was taken and shown in the top row. The bottom row plots the features of each representative on its own and on the other's class-conditional representations. Sparsity of the first representation (straight black line) is observed as well as random feature activation when class belonging is not met (dotted red line).

Though several parametric, semi-parametric and nonparametric approaches are possible, the experiments were performed using Laplacian or Gaussian mixtures. Figure (2) shows the performance of these estimations on real data.

4 Feature Selection

The fact the features we are dealing with are statistically independent can also be an advantage in the context of feature selection. Divergence, a frequently used measure for feature selection is additive for statistically independent variables.

Class separability is a standard criterion in feature selection for classification. Measures for class separability are generally obtained from the distance among the previously estimated class-conditional distributions. A commonly used distance measure for (class-conditional) densities, for its connection with information theory, is the Kullback-Leibler distance,

$$KL(C_i, C_j) = \int_{\Omega} p(\mathbf{x}|C_i) \log \frac{p(\mathbf{x}|C_i)}{p(\mathbf{x}|C_j)} d\mathbf{x} \tag{8}$$

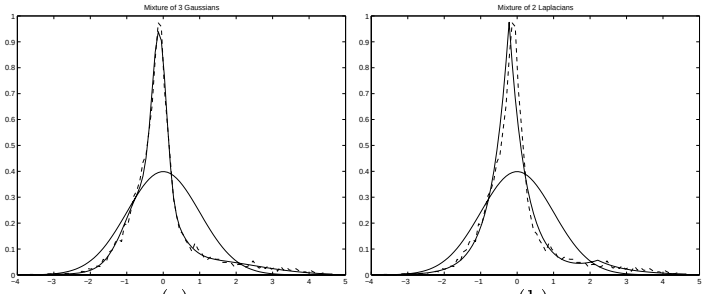

Fig. 2. Two of the proposed densities are estimated from a typical independent component obtained in the experiments (MNIST database). The true histogram of the component (dashed) and the standard gaussian density are also plotted as a reference.

where $1 \leq i, j \leq K$. The asymmetry of Kullback-Leibler motivates the symmetric measure of divergence, since long ago used for feature selection [16], defined as

$$\hat{D}_{ij} = \hat{D}(C_i, C_j) = KL(C_i, C_j) + KL(C_j, C_i) \tag{9}$$

Besides being symmetric, divergence is zero between a distribution and itself, always positive, monotonic on the number of features and provides an upper bound for the classification error [12]. The two main drawbacks of divergence are that it requires density estimation and has a nonlinear relationship with classification accuracy. While the second drawback is usually overcomed by using a transformed version of divergence, introduced by Swain and Davis [22, 21], the first inconvenient is not present when class-conditional features are independent. For this case, it can be seen that divergence is additive on the features. So, for this particular case, unidimensional density estimation can be performed and the calculation of divergence for a feature subset $S \subseteq \{1, ..., N\}$ (noted by $\hat{D}_{ij}^{S}$) is straightforward. A very important property besides montonicity shared by transformed divergence and divergence, is that

$$(n_1 \notin S, n_2 \notin S) \wedge (D_{ij}^{n_1} \leq D_{ij}^{n_2}) \Rightarrow (D_{ij}^{S \bigcup n_1} \leq D_{ij}^{S \bigcup n_1}) \tag{10}$$

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 (transformed) divergence, and thus provides a very simple rule for feature selection without involving any search procedure.

Although, (transformed) 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 method is to use the average divergence, defined as the the average divergence over all class pairs. This approach is simple and preserves the exposed property of order for feature subsets, but it is not reliable as the variance of the pairwise divergences increases. A more robust approach is to sort features by their maximum minimum (two-class) divergence. This works fine for small subsets but decays as the size of the subset increases: sorting features by maximum minimum divergence is a very conservative election.

In the CC-ICA context we have K local linear representations, each one making $\mathbf{x}|C_k$ independent. This involves the selection of possibly distinct single features belonging to

different representations. We now provide an alternative definition of divergence, adapted to local representations.
The log-likelihood ratio (L) is defined as,

$$L_{ij}(\mathbf{x}) = \log p(\mathbf{x}|C_i) - \log p(\mathbf{x}|C_j) \tag{11}$$

$L_{ij}(\mathbf{x})$ measures the overlap of the class-conditional densities in $\mathbf{x}$. It can be seen from (9) that $D_{ij} = E_{C_i}(L_{ij}) + E_{C_j}(L_{ji})$ where E_{C_i} is the class-conditional expectation operator. Approximating $E_{C_i}(g(x)) \approx (1/\#C_i) \sum_{x \in C_i} g(x) \stackrel{def}{=} \overline{g(x)}_{C_i}$, and reordering the terms, we have

$$D_{ij} \approx \left(\overline{\log p(x|C_i)}_{C_i} - \overline{\log p(x|C_i)}_{C_j} \right) + \left(\overline{\log p(x|C_j)}_{C_j} - \overline{\log p(x|C_j)}_{C_i} \right) \stackrel{def}{=} D'_{ij} + D'_{ji} \tag{12}$$

D'_{ij} measures the difference in the expected likelihood of classes i and j, assuming all samples are taken from class i. It is no longer symmetric but still additive for conditionally independent variables. Introducing (6) D'_{ij} can be expressed as,

$$D'_{ij} = \nu_i \sum_{m=1}^{M_i} \left(\overline{\log p^i(s_m)}_{C_i} - \overline{\log p^i(s_m)}_{C_j} \right) \stackrel{def}{=} \nu_i \sum_{m=1}^{M_i} D'^m_{ij} \tag{13}$$

Divergence is maximized by maximizing both D'_{ij} and D'_{ji}. The assymetry and locality of the latter will cause different feature subsets on each class 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 as with divergence. For instance, having fixed the representation, the average has to be taken over only one index,

$$D'^m_{A_i} = \frac{1}{K-1} \sum_{j=1, j \neq i}^{K} D'^m_{ij} \tag{14}$$

5 Experiments

A first experiment is performed on the Letter Image Recognition Data [3]. Each instance of the 20000 images within this database represents a capital typewritten letter in one of twenty fonts. Each letter is represented using 16 integer valued features corresponding to statistical moments and edge counts. Training is done on the first 16000 instances and test on the final 4000. There are approximately 615 samples per class in the training set. Fig. (5) 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. For all the Naive Bayes Classifiers, the mixture of two gaussians was used to estimate the resulting unidimensional densities. The results of Maximum Likelihood classification on PCA were also included as a reference.
We can observe in Fig. (5) 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. To test the feature selection criterion, on this database we also tried Naive Bayes on 10000

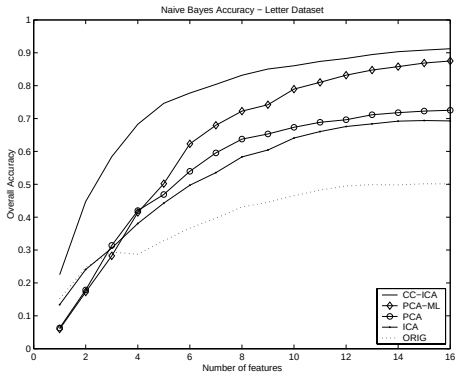

Fig. 3. Naive Bayes performance on the original, PCA (class-conditional), PCA-ML (global), ICA (global) and CC-ICA representations of the Letter Database. The importance of the independence assumption on Naive Bayes performance is observed. Maximum Likelihood on a global PCA representation is added as a reference.

random 8-feature combinations for each class, resulting that no combination achieved our classification results (83.17%).

The second experiment was performed on the MNIST handwritten digit database [13], which contains 60000 training and 10000 test samples. The images were resized from 28×28 to 16×16 resulting in 256 dimensional samples. 5000 and 750 samples per digit were randomly chosen for training and test sets, respectively. Overall accuracy using 1 through 128 features is plotted in Fig. (5). In all cases, a Naive Bayes classifier was used and the unidimensional densities estimated using the same approach (mixture of three gaussians) for adequate comparison. Also in all cases using simpler estimation methods such as gaussian or nonparametric (frequential) estimation performs worst than the exposed results. In the graph, PCA stands for a class-conditional PCA representation using the features as given by PCA. This approach performs poorly for a low number of features (< 50) but, after 60 features outperforms all the other methods, starting to decrease in performance after 100 features. Using the divergence feature selection criterion on PCA (PCA-FS) improves the performance of Naive Bayes on a PCA representation for a low number of features. CC-ICA obtains the best accuracy when the number of features is less than 60, obtaining an accuracy of .9 with as few of 50 features and .8 with only 9 features. The accuracy of CC-ICA is monotonic on the number of features. Several hypothesis can be thought of when analyzing lower accuracy of CC-ICA with respect to PCA for a large number of features. From the ICA perspective, it is well known that in large dimensions degenerate independent sources can arise. This seems to be our case since, in order to allow a dimensionality of 128, we have included sources with estimated kurtosis as high as 100. This affects both the classifier and the feature selection criterion.

Mencionar: In all cases unidimensional feature densities are estimated using the same approach (gaussian mixtures) for adequate comparison. Also in all cases using simpler estimation methods such as gaussian or nonparametric (frequential) estimation performs considerably worst than the exposed results.

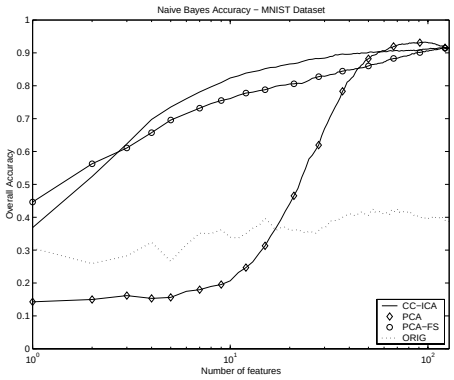

Fig. 4. Naive Bayes performance on the original representation and class-conditional PCA and ICA of the MNIST Database. Two feature selection criterions are employed on PCA. Logarithmic scale on the x-axis is employed.

6 Conclusions

The Naive Bayes classifier, though its generally unmet assumptions and notorious simplicity, still performs well over a large variety of problems. In this article, by making use of Independent Component Analysis, we present a class-conditional representation that allows to hold the Naive Bayes independence assumption on stronger grounds and thus improve the performance. Reinforcing the hypothesis is not the only reason for this improvement. It has been shown that Naive Bayes performance has a direct relationship with feature low entropy, and it is also well known that in several cases the independent components have low entropy (supergaussian/sparse distributions). For this representation we also introduce a scheme for selecting those (class-conditional) features most adequate for the task of classification. This scheme takes advantage of the property that states that feature divergence is additive on statistically independent features. Precisely the assumption we will make when using Naive Bayes.
A first experiment is performed in order to show that our proposed representation and feature selection criterion performs well even in low dimensional problems. The second experiment, on the MNIST database, evaluates Naive Bayes improvement in a high dimensional database. In both experiments results are compared against applying Naive Bayes on the original representation and on a PCA representation.

References

1. A. Bell and T. Sejnowski. An information-maximization approach for blind signal separation. *Neural Computation*, 7:1129–1159, 1995.
2. A. Bell and T. Sejnowski. The 'independent components' of natural scenes are edge filters. *Neural Computation*, 11:1739–1768, 1999.
3. C. Blake and C. Merz. Uci repository of machine learning databases, 1998.
4. P. Comon. Independent component analysis - a new concept? *Signal Processing*, 36:287–314, 1994.

5. A. P. Dawid. Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society, Ser. B*, 41:1–31, 1979.
6. P. Domingos and M. J. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29(2-3):103–130, 1997.
7. R. Duda, P. Hart, and D. Stork. *Pattern Classication*. John Wiley and Sons, Inc., New York, 2nd edition, 2001.
8. D. Field. What is the goal of sensory coding? *Neural Computation*, 6:559–601, 1994.
9. A. Hyvärinen. New approximatins of differential entropy for independent component analysis and projection pursuit. *Advances in Neural Processing Systems*, 10:273–279, 1998.
10. A. Hyvärinen. Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation. *Neural Computation*, 11:1739–1768, 1999.
11. A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley and Sons, 2001.
12. T. Kailath. The divergence and bhattacharyya distance measures in signal selection. *IEEE Trans. on Communication Technology COM-15*, 1:52–60, February 1967.
13. Y. LeCun and A. Labs-Research. *The MNIST DataBase of Handwritten digits*. http://www.research.att.com/ yann/ocr/mnist/index.html, 1998.
14. 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 Transactions on PAMI*, 22(10):1–12, 2000.
15. D. Lewis. Naive bayes at forty: The independence assumption in information retrieval. In C. N'edellec and C. Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, volume 1398.25, pages 4–15. Springer Verlag, Heidelberg, DE, 1998.
16. T. Marill and D. Green. On the effectiveness of receptors in recognition systems. *IEEE Trans. on Information Theory*, 9:1–17, 1963.
17. M.Bressan, D.Guillamet, and J.Vitria. Using an ica representation of high dimensional data for object recognition and classification. In *IEEE CSC in Computer Vision and Pattern Recognition (CVPR 2001)*, volume 1, pages 1004–1009, 2001.
18. I. Rish, J. Hellerstein, and J. Thathachar. An analysis of data characteristics that affect naive bayes performance. In C. N'edellec and C. Rouveirol, editors, *Proceedings of the Eighteenth Conference on Machine Learning -ICML2001*, pages –. Morgan Kaufmann, 2001.
19. D. W. Scott. *Multivariate Density Estimation*. John Wiley and sons, New York, NY, 1992.
20. E. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Ser. B*, 13:238–241, 1951.
21. P. Swain and S. Davis. *Remote sensing: the quantitative approach*. McGraw-Hill, 1978.
22. P. Swain and R. King. Two effective feature selection criteria for multispectral remote sensing. In *Proceedings of the 1st International Joint Conference on Pattern Recognition, IEEE 73 CHO821-9*, pages 536–540, 1973.
23. H. Turtle and W. Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187–222, July 1991.
24. R. Vigario, V. Jousmäki, M. Hämäläinen, R. Hari, and E. Oja. Independent component analysis for identification of artifacts in magnetoencephalographic recordings. *Advances in Neural Information Processing Systems*, 10:229–235, 1998.
25. Y. Yang, S. Slattery, and R. Ghani. A study of approaches to hypertext categorization. *Journal of Intelligent Information Systems. Kluwer Academic Press*, 2002.