Neurocomputing 101 (2013) 161–169

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Combining information theoretic kernels with generative embeddings for classification

Manuele Bicego [a,*], Aydın Ulaş [a], Umberto Castellani [a], Alessandro Perina [e], Vittorio Murino [a,b], André F.T. Martins [c], Pedro M.Q. Aguiar [d], Mário A.T. Figueiredo [c]

[a] Dipartimento di Informatica, University of Verona, Verona, Italy
[b] Istituto Italiano di Tecnologia – IIT, Genova, Italy
[c] Instituto de Telecomunicações, Instituto Superior Técnico, Lisboa, Portugal
[d] Instituto de Sistemas e Robótica, Instituto Superior Técnico, Lisboa, Portugal
[e] Microsoft Research, Redmond, WA, USA

ARTICLE INFO

Article history:
Received 20 January 2012
Received in revised form
23 June 2012
Accepted 25 August 2012
Communicated by H. Yu
Available online 18 September 2012

Keywords:
Hybrid generative–discriminative schemes
Generative embeddings
Probabilistic latent semantic analysis
Information theoretic kernels

ABSTRACT

Classical approaches to learn classifiers for structured objects (*e.g.*, images, sequences) use generative models in a standard Bayesian framework. To exploit the state-of-the-art performance of discriminative learning, while also taking advantage of generative models of the data, generative embeddings have been recently proposed as a way of building hybrid discriminative/generative approaches. A generative embedding is a mapping, induced by a generative model (usually learned from data), from the object space into a fixed dimensional space, adequate for discriminative classifier learning. Generative embeddings have been shown to often outperform the classifiers obtained directly from the generative models upon which they are built.

Using a generative embedding for classification involves two main steps: (i) defining and learning a generative model and using it to build the embedding; (ii) discriminatively learning a (maybe kernel) classifier with the embedded data. The literature on generative embeddings is essentially focused on step (i), usually taking some standard off-the-shelf tool for step (ii). Here, we adopt a different approach, by focusing also on the discriminative learning step. In particular, we exploit the probabilistic nature of generative embeddings, by using kernels defined on probability measures; in particular we investigate the use of a recent family of non-extensive information theoretic kernels on the top of different generative embeddings. We show, in different medical applications that the approach yields state-of-the-art performance.

1. Introduction

Most approaches to the statistical learning of classifiers belong to one of two classical paradigms: *generative* and *discriminative* [1,2], also known in the statistics literature as *sampling* and *diagnostic*, respectively [3]. Generative approaches are based on probabilistic class models and *a priori* class probabilities, learnt from training data and combined via Bayes law to yield posterior probabilities. Discriminative learning methods aim at learning class boundaries or posterior class probabilities directly from data, without relying on intermediate generative class models.

In the past decade, several hybrid generative–discriminative approaches have been proposed with the goal of combining the best of both paradigms [4,5]. These approaches can loosely be divided into three groups: blending methods, iterative methods, and staged methods. In a few words, blending methods (e.g. [5,6]) try to optimize a single objective function that contains different terms coming from the generative and discriminative model. Iterative methods (e.g. [7–9]) are algorithms involving a generative and a discriminative model that are trained in an iterative process, each influencing the other. Finally, in staged methods [4,10–12], the models are trained in separate procedures, but one of the models – usually the discriminative one – is trained on features provided by the first. This later family is currently the most frequently applied and studied, and it includes the class of methods known as generative embeddings (or score spaces), where the basic idea is to exploit a generative model to map the objects to be classified into a feature space. This is particularly suited for non-vectorial data (strings/sequences, trees, images), as it maps objects of possibly different dimensions (*e.g.*, strings of different lengths) into a fixed dimension space.

The seminal work on generative embeddings is [4], where the so-called *Fisher score* was introduced. In that work, the features of a given object are the derivatives of the log-likelihood function under the assumed generative model, with respect to the model parameters, computed at that object. Other examples of generative embeddings can be found in [10–12].

* Corresponding author.
E-mail address: manuele.bicego@univr.it (M. Bicego).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.08.014

Typically, the feature vectors resulting from the generative embedding are used to feed some kernel-based classifier, such as a *support vector machine* (SVM) with standard linear or radial basis function (RBF) kernels. In this paper, we follow an alternative route: instead of relying on standard kernels, we investigate the use of a recently introduced family of information theoretic (IT) kernels [13]. The main idea is that the IT kernels we can exploit the probabilistic nature of the generative embeddings, improving even more the classification results of the hybrid approaches. In particular we investigate a particular class of IT kernels, based on a non-extensive generalization of the classical Shannon information theory, and defined on unnormalized or normalized (*i.e.*, probability) measures. In [13], they were successfully used in text categorization tasks, based on multinomial text representations (*e.g.*, bags-of-words, character *n*-grams). Here, the idea is to consider the points of the generative embedding as multinomial probability distributions, thus valid arguments for the information theoretic kernels.

We illustrate the performance of combining different generative embeddings with the IT kernels on different medical applications: colon cancer detection on gene expression data, schizophrenia detection on brain MRI images, and renal cell cancer classification on tissue microarray data. Following recent work, we adopt the so-called pLSA (*probabilistic latent semantic analysis*) as a generative model, the usefulness of which has been recently shown in different applications [11,14–16]. The experimental results reported in this paper testify for the adequacy and state-of-the-art performance of the combination of IT kernels with generative embeddings.

Summarizing, the main contributions of the paper are:

- The investigation of the use of a novel class of information theoretic (IT) kernels [13] as a similarity measure between objects in a generative embedding space.
- A thorough investigation of different generative embedding (GE) schemes, some of them being very recent. Such a large and extensive comparison, involving eight different generative embeddings, was missing from the literature.
- The application of this hybrid scheme (GE+IT kernels) to the medical domain. Actually it is worth to notice that we exploit the same scheme for three very different medical applications, which start from very different representations: 3D surfaces (brain classification), images (renal cancer), and microarray expression matrices (colon cancer).

The remaining sections of the paper are organized as follows. In Section 2, the fundamental ideas of generative embeddings are reviewed, together with the basics of the schemes here investigated, while Section 3 describes the IT kernels. The proposed way of using the IT kernels with the generative embeddings is formalized in Section 4. Details on applications and experimental results are reported in Section 5, and Section 6 concludes the paper.

2. Generative embeddings

Pursuing principled hybrid discriminative–generative classifier learning methods is, arguably, one of the currently most interesting challenges in machine learning research. The underlying motivation is the clear complementarity of discriminative and generative strategies: asymptotically (in the number of labeled training examples), classification error of discriminative methods is lower than for generative ones [1]. On the other side, generative schemes are effective with less data; furthermore, they allow for easier/simpler handling of missing data and inclusion of prior knowledge about the data. Among these hybrid generative–discriminative methods, the interest in "generative embeddings" (also called generative score spaces) has been increasing in recent years, as is testified by an increasing literature on the class of methods (see, among other, [4,11,14,17–21]).

Generative embeddings involve three key building blocks: (i) a generative model (or a family thereof) is adopted and learned from the data; (ii) this learned model is used to obtain a mapping between the original object space and a fixed-dimension vector space (usually called a *score space*); (iii) the objects in the training set are mapped into the score space and used by some discriminative learning technique. The key idea is the mapping of objects of possibly different dimension into fixed-dimensional feature vectors, using a model of how this objects are generated. This opens the door to the use of standard discriminative techniques (such as support vector machines or logistic regression) and has been shown to achieve higher classification accuracy than purely generative or discriminative approaches.

Once a generative embedding is obtained, in order to use a kernel-based discriminative learning approach, it is necessary to adopt a kernel that expresses similarity between pairs of points in the score space, maybe also derived from the adopted generative model. The most famous example of one such kernel is the *Fisher kernel* [4], which is simply a Riemennian inner product, using the inverse Fisher matrix of the generative model as the underlying metric. In this paper, we will use kernels defined on the score space that are independent of the generative model.

In the following subsections, we will describe the generative embeddings used in this paper, after reviewing the pLSA generative model based on which they are built.

2.1. Probabilistic latent semantic analysis (pLSA)

Consider a set of documents $\mathcal{D}=\{d_1,\ldots,d_{|\mathcal{D}|}\}$, each containing an arbitrary number of words, all taken from a vocabulary of $\mathcal{W}=\{w_1,\ldots,w_{|\mathcal{W}|}\}$; of course, without loss of generality, we may simply refer to the documents and words by their indices, thus we simplify the notation by writing $\mathcal{D}=\{1,\ldots,|\mathcal{D}|\}$ and $\mathcal{W}=\{1,\ldots,|\mathcal{W}|\}$. This collection of documents is summarized in a bag-of-words fashion (*i.e.*, ignoring the word order) into a $|\mathcal{W}|\times|\mathcal{D}|$ occurrence matrix $\mathbf{C}=[C_{ij},\ i=1,\ldots,|\mathcal{W}|,\ j=1,\ldots,|\mathcal{D}|]$, where element C_{ij} indicates the number of occurrences of the i-th word in the j-th document.

pLSA [22] is a generative mixture model for matrix $\mathbf{C}$ where the presence of each word in each document is mediated by a latent random variable, $Z\in\mathcal{Z}=\{1,\ldots,|\mathcal{Z}|\}$ (known as the *topic* or *aspect* variable). More specifically, pLSA is a mixture model for the joint distribution of the pair of random variables $D\in\mathcal{D}$ and $W\in\mathcal{W}$, where the event $(W=i,D=j)$ means that there is an occurrence of the i-th word in the j-th document. pLSA expresses the joint probability distribution $P(W=i,D=j)$ as a mixture of distributions such that, in each component of the mixture (*i.e.*, for each *topic*), the random variables W and D are independent (*i.e.*, $\mathbb{P}(W=i,D=j|Z=z)=\mathbb{P}(W=i|Z=z)\mathbb{P}(D=j|Z=z)$); formally,

$$\mathbb{P}(W=i,D=j)=\sum_{z=1}^{|\mathcal{Z}|}\mathbb{P}(Z=z)\mathbb{P}(W=i|Z=z)\mathbb{P}(D=j|Z=z). \quad (1)$$

This model is parameterized by a set of $1+2|\mathcal{Z}|$ multinomial distributions: the distribution of the latent topic variable $\{\mathbb{P}(Z=1),\ldots,\mathbb{P}(Z=|\mathcal{Z}|)\}$; the distributions of words $\{\mathbb{P}(W=1|Z=z),\ldots,\mathbb{P}(W=|\mathcal{W}||Z=z)\}$ for each $z\in\{1,\ldots,|\mathcal{Z}|\}$; the distributions of documents $\{\mathbb{P}(D=1|Z=z),\ldots,\mathbb{P}(D=|\mathcal{D}||\ Z=z)\}$ for each $z\in\{1,\ldots,|\mathcal{Z}|\}$. Let us write these parameters compactly in a vector $\mathbf{p}=[p_1,\ldots,p_{|\mathcal{Z}|}]$, where $p_z\equiv\mathbb{P}(Z=z)$ and a pair of matrices $\mathbf{Q}$ and $\mathbf{R}$, where $Q_{zw}\equiv\mathbb{P}(W=w|Z=z)$ and $R_{zd}\equiv\mathbb{P}(D=d|Z=z)$. Of course, both $\mathbf{Q}$ and $\mathbf{R}$ are stochastic matrices: $Q_{zw}\geq 0$, $R_{zw}\geq 0$, $\sum_{w=1}^{|\mathcal{W}|}Q_{zw}=1$, and $\sum_{d=1}^{|\mathcal{D}|}R_{zd}=1$.

Given a set of N independent samples $\{(w_n,d_n) \in \mathcal{W} \times \mathcal{D},\ n = 1,\ldots,N\}$ from this generative model, the log-likelihood function (from which the parameters $\mathbf{p}$, $\mathbf{Q}$, and $\mathbf{R}$ are to be estimated) can be easily shown to be

$$\mathcal{L}(\mathbf{p},\mathbf{Q},\mathbf{R}) = \sum_{w=1}^{|\mathcal{W}|}\sum_{d=1}^{|\mathcal{D}|} C_{wd}\log(\mathbb{P}(W=w,D=d)), \tag{2}$$

$$\mathcal{L}(\mathbf{p},\mathbf{Q},\mathbf{R}) = \sum_{w=1}^{|\mathcal{W}|}\sum_{d=1}^{|\mathcal{D}|} C_{wd}\log\left(\sum_{z=1}^{|\mathcal{Z}|} p_z Q_{zw} R_{zd}\right), \tag{3}$$

where C_{wd} is the number of times the pair (w, d) occurs in the set of observations, that is, the number of times that the w-th word occurs in the d-th document (as defined above). This shows that matrix $\mathbf{C}$ contains the sufficient statistics to estimate the parameters of the pLSA model. Of course, maximizing (3) w.r.t. $\mathbf{p}$, $\mathbf{Q}$, and $\mathbf{R}$ cannot be done in closed form, but can naturally be addressed via the EM algorithm [22].

It is important to note that the (multinomial) random variable D takes values in the list of documents in the training set. For this reason, pLSA is not a full generative model of documents in the sense that it has no way to assign a probability to a previously unseen document.

In possession of estimates of the model parameters, $\widehat{\mathbf{p}}$, $\widehat{\mathbf{Q}}$, and $\widehat{\mathbf{R}}$, it is possible to estimate quantities such as the probability that a given topic is present in a given document

$$\mathbb{P}(Z=z|D=d) = \frac{\widehat{R}_{zd}\widehat{p}_z}{\sum_{s=1}^{|\mathcal{Z}|}\widehat{R}_{sd}\widehat{p}_s}. \tag{4}$$

2.2. pLSA-based generative embeddings

Generative embeddings can be divided into two families: those based on the generative model parameters and those based on hidden variables of those models. The former class derives the features by using differential operations with respect to the model parameters, while the latter derive feature maps using the log-likelihood function of the model, focusing on the random variables rather than on the parameters.

2.2.1. Parameter-based generative embeddings

In this subsection, we review three of the best-known generative embeddings based on the generative model parameters

The Fisher score (*FS*) was the first example of generative embedding [4], and it consists of using as feature vector the tangent vector of the data log likelihood with respect to the model parameters. In the case of the pLSA model [23], each document $d \in \{1,\ldots,|\mathcal{D}|\}$ is mapped into the gradient of its log-probability w.r.t. the model parameters, which we collect into a vector $\boldsymbol{\theta} \equiv (\mathbf{p},\mathbf{Q},\mathbf{R})$. The log-probability of a document $d \in \{1,\ldots,|\mathcal{D}|\}$, denoted as $l(d)$, is obtained by marginalization,

$$\begin{aligned} l(d_i) &= \log \mathbb{P}(D=d) = \log \sum_{w=1}^{|\mathcal{W}|} \mathbb{P}(W=w,D=d) \\ &= \log \sum_{w=1}^{|\mathcal{W}|}\sum_{z=1}^{|\mathcal{Z}|} p_z Q_{zw} R_{zd}. \end{aligned} \tag{5}$$

The pLSA-based Fisher score maps each document d into a vector of dimension containing the derivatives of $l(d)$ w.r.t. to the elements of $\boldsymbol{\theta}$. In this score space, we define the kernel simply as the Euclidean inner space. Alternatively (although we do not consider that choice here), the kernel could be defined as the Riemennian inner product, using the inverse Fisher matrix as the metric [4].

The TOP kernel (where TOP is an acronym for *T*angent *O*f *P*osterior log-odds [17]) was designed for two-class problems and is based on the gradient of the posterior log-odds ratio. Formally, given parameter estimates of two pLSA generative models for the two classes, $\boldsymbol{\theta}^{(-1)}$ and $\boldsymbol{\theta}^{(+1)}$, a given document d is mapped into the gradient of the posterior log-odds ratio $\log \mathbb{P}(C=+1|d,\boldsymbol{\theta}) - \log \mathbb{P}(C=-1|d,\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(-1)},\boldsymbol{\theta}^{(+1)})$. Finally, the TOP kernel is defined simply as the Euclidean inner product in the resulting vector space.

The log-likelihood ratio (LLR) embedding [20] is similar to the Fisher score, except that it uses one generative model per class, rather than a single model. Formally, for a C-class problem, the LLR embedding maps a given document d into the concatenation of the gradients of $\log \mathbb{P}(d|\boldsymbol{\theta}^{(1)}),\ldots,\log \mathbb{P}(d|\boldsymbol{\theta}^{(C)})$, w.r.t. the respective parameters. Consequently, the dimensionality of the LLR embedding is C times larger than that of the Fisher embedding.

2.2.2. Latent-variable-based embeddings

These methods, arising from considerations in [14], derive generative feature mappings from the log-likelihood, using the hidden variables of the model, rather than on its parameters.

The free energy score space (*FESS*) is based on the observation that the free energy bound on the complete log-likelihood decomposes into a sum of terms [14]; the mapping of a given document is the vector containing the terms in this decomposition. The details of the free energy bound and the resulting embedding (the FESS) are too long to include here, so the reader is referred to [14].

The posterior divergence (*PD*) embedding is a modification of the FESS embedding [19] which also takes into account how much each sample affects the model. Details on the pLSA-based PD embedding and on its relationship with FESS case can be found in [19].

The mixture of topics (*MT*) embedding simply maps a given document d into the $|\mathcal{Z}|$-dimensional vector containing the conditional probabilities $\mathbb{P}(Z=1|D=d),\ldots,\mathbb{P}(Z=|\mathcal{Z}||D=d)$. Recall that these probabilities (given the parameter estimates) are computed according to (4).

2.2.3. Some remarks

Recently, pLSA has been used, not only for text problems, but in several other application areas, including computer vision, bioinformatics (gene expression data), and medical image analysis [11,24,25]. In imaging problems, the idea is use pLSA to model the co-occurrence of image features (*visual words*) [11,25].

One obvious question that arises when using pLSA models is the selection of the number of topics $|\mathcal{Z}|$. In all our application, we have estimated this number by using the well-known *Bayesian information criterion* (BIC) [26], which penalizes the likelihood with a term that depends on number of model parameters. In the pLSA model, the number of free parameters is $|\mathcal{Z}|-1+|\mathcal{Z}|(|\mathcal{D}|+|\mathcal{W}|-2)$. Thus, the number of topics is chosen as the minimizer w.r.t. $|\mathcal{Z}|$ of the penalized log-likelihood

$$-\sum_{w=1}^{|\mathcal{W}|}\sum_{d=1}^{|\mathcal{D}|} C_{wd}\log\left(\sum_{z=1}^{|\mathcal{Z}|} p_z Q_{zw} R_{zd}\right) + [|\mathcal{Z}|-1+|\mathcal{Z}|(|\mathcal{D}|+|\mathcal{W}|-2)]\log(\sqrt{N}).$$

In our experiments, we consider two versions of the FESS and MT embeddings. In the first version, we train one pLSA model per class and concatenate the resulting feature vectorss (we will refer these as FESS-1 and MT-1); in the second version, we train a pLSA model for the whole data, ignoring the class label (we will refer these as FESS-2 and MT-2). In summary, we will consider eight different generative embeddings: MT-1, MT-2, FESS-1, FESS-2, LLR, FS, TOP, and PD.

3. Information theoretic kernels

Kernels on probability measures have been shown very effective in classification problems involving text, images, and other types of data [13,27,28]. Given two probability measures p_1 and p_2, representing two objects, several information theoretic kernels (ITKs) can be defined [13]. The Jensen–Shannon kernel is defined as

$$k^{\mathrm{JS}}(p_1,p_2) = \ln(2) - JS(p_1,p_2), \tag{6}$$

with $JS(p_1,p_2)$ being the Jensen–Shannon divergence

$$JS(p_1,p_2) = H\left(\frac{p_1+p_2}{2}\right) - \frac{H(p_1)+H(p_2)}{2}, \tag{7}$$

where $H(p)$ is the usual Shannon entropy. The Jensen–Tsallis (JT) kernel is given by

$$k_q^{\mathrm{JT}}(p_1,p_2) = \ln_q(2) - T_q(p_1,p_2), \tag{8}$$

where $\ln_q(x) = (x^{1-q}-1)/(1-q)$ is the q-logarithm,

$$T_q(p_1,p_2) = S_q\left(\frac{p_1+p_2}{2}\right) - \frac{S_q(p_1)+S_q(p_2)}{2^q} \tag{9}$$

is the Jensen–Tsallis q-difference, and $S_q(r)$ is the Tsallis non-extensive entropy, defined, for a multinomial distribution $r = (r_1,\ldots,r_L)$ as

$$S_q(r_1,\ldots,r_L) = \frac{1}{q-1}\left(1 - \sum_{i=1}^{L} r_i^q\right).$$

In [13], versions of these kernels applicable to unnormalized measures were also defined. Let $\mu_1 = \omega_1 p_1$ and $\mu_2 = \omega_2 p_2$ be two unnormalized measures, where p_1 and p_2 are the normalized counterparts (probability measures), and ω_1 and ω_2 arbitrary positive real numbers (weights). The weighted JT kernel (version A) is given by

$$k_q^A(\mu_1,\mu_2) = S_q(\pi) - T_q^{\pi}(p_1,p_2), \tag{10}$$

where $\pi = (\pi_1,\pi_2) = (\omega_1/(\omega_1+\omega_2), \omega_2/(\omega_1+\omega_2))$ and

$$T_q^{\pi}(p_1,p_2) = S_q(\pi_1 p_1 + \pi_2 p_2) - (\pi_1^q S_q(p_1) + \pi_2^q S_q(p_2)).$$

The weighted JT kernel (version B) is defined as

$$k_q^B(\mu_1,\mu_2) = (S_q(\pi) - T_q^{\pi}(p_1,p_2))(\omega_1+\omega_2)^q. \tag{11}$$

4. Proposed approach

The approach herein proposed consists in defining a kernel between two observed objects x and x' as the composition of the score function with one of the JT kernels presented above. Formally,

$$k(x,x') = k_q^i(\phi_\Theta(x), \phi_\Theta(x')), \tag{12}$$

where $i \in \{\mathrm{JT,A,B}\}$ indexes one of the Jensen–Tsallis kernels (8), (10), or (11), and ϕ_Θ is one of the generative embeddings defined in Section 2.

We consider two types of kernel-based classifiers: K-NN and SVM. Recall that positive definiteness is a key condition for the applicability of a kernel in SVM learning. It was shown in [13] that k_q^A is a positive definite kernel for $q \in [0,1]$, while k_q^B is a positive definite kernel for $q \in [0,2]$. Standard results from kernel theory [29, Proposition 3.22] guarantee that the kernel k defined in (12) inherits the positive definiteness of k_q^i, thus can be safely used in SVM learning algorithms.

5. Experimental evaluation

We have applied the proposed approach to three (binary) classification problems in the medical domain, which will be described in detail below: binary classification of brain MRI images into schizophrenia/non-schizophrenia; cancer detection in tissue microarray (TMA) images; colon cancer detection in gene expression microarray data.

All the accuracies are computed using the averaged hold out cross validation (30 repetitions). The standard errors of means, in all runs, were all less than 0.0252, 0.0032 and 0.0179, for the Brain Classification Task, the Renal Cancer Classification task and the Colon Cancer Classification task, respectively. The value of parameter q of the IT kernels is estimated using 5-fold cross-validation on the training set. Since results were similar, we omit the weighted JT kernel (version B): we will refer to weighted JT kernel (version A) as W-Jen–Tsal. As classifiers, we use *support vector machines* (SVM), with the well-known parameter C adjusted by 5-fold cross-validation on the training set, as well as the *K-nearest neighbors* classifier, with $K=1$, *i.e.*, the nearest neighbor (NN) rule.

Many comparisons have been carried out, in order to highlight the specific contribution of every part of the hybrid approach. In particular, we compared the proposed pipeline with the results obtained in the original space with standard kernels (linear and radial basis function – RBF – kernels), in the generative embedding space with the standard kernels, and in the original space with IT kernels. All the details are presented in the results subsection.

5.1. Application details

We will now describe the three applications in more detail. In particular, we will describe how the pLSA model is used in each problem, that is, what is the meaning of terms "words" and "documents" in each particular type of data. The datasets used in the experiments are summarized in Table 1.

5.1.1. Classification of brain MRI

In the first application, the goal is to analyze MRI brain scans in order to distinguish between normal subjects and subjects affected by schizophrenia (see Fig. 1). We adopt an approach based on a *region of interest* (ROI) [30], where the idea is to focus on a region of the brain considered informative for the task at hand. In particular, our analysis focuses on the left thalamus, whose abnormal activity has been already correlated with schizophrenia [31]. The pLSA model is used as proposed in [25]: local 3D brain shape features are computed from the MRI data; the "visual words" are obtained by quantization of the features; histograms of visual words in each subject (*i.e.*, each "document", in the pLSA language) finally lead to the counting matrix **C** from which the pLSA model is learned.

In our experiments the data set consists of MRI data of 30 healthy and 30 schizophrenic subjects.[1] The visual features were extracted using the so-called *heat kernel signature* (HKS) [32],

[1] This data was collected in the context of the SIMBAD project, which is an European FET project dealing with similarity-based approaches to pattern recognition; see http://simbad-fp7.eu/.

which is able to encode simultaneously the contribution of local features for a fixed set of scales into a single shape descriptor [33]. The dictionary of words was obtained by quantizing all the features into 100 bins.

5.1.2. Renal cancer classification via tissue microarray

In the second application, the aim is to analyze tissue microarray (TMA) images in order to identify whether a given renal cell nucleus is malignant or benign. For this purpose, TMA are obtained and the images are normalized and segmented for nuclei; finally the true labels are assigned by a pool of pathologists (see Fig. 2). To build the "visual words", features are extracted from the segmented nuclei (as in [34]) and then quantized into 168 bins. In particular, we used the *pyramid histograms of oriented gradients* (PHOG, see [35] for details) computed over a 2-level pyramid of patches.

In our experiments, we use a set of three patients (more details can be found in [34]) from which 474 nuclei (*i.e.*, "documents", in pLSA terms) were segmented; 321 (67%) benign and 153 (33%) malignant.

5.1.3. Colon cancer classification from gene expression microarray data

In the third application, the goal is to analyze gene expression microarray data in order to distinguish between healthy people and people affected by colon cancer. The starting point is a microarray gene expression matrix, where the element at position (i, j) represents the expression level of the ith gene in the jth subject/sample. Topic models (of which pLSA is an instance) have been recently and successfully applied in this context (see, *e.g.*, [24,36]). Actually, it is possible to establish an analogy between a word-document pair and a gene-sample pair; it seems reasonable to interpret samples as documents and genes as words. In this way, the gene expression levels in a sample may interpreted as the word counts in a document. Consequently, we can simply take a gene expression matrix and (of course, after a preprocessing step, for example, to remove possibly negative numbers [24]) interpret it as a count matrix $\mathbf{C}$ from which a pLSA model can be estimated.

Table 1
Summary of the applications and the corresponding numbers of "words" and "documents".

Problem	# Classes	# Documents	# Words
Brain MRI classification	2	60	100
Renal cancer classification	2	474	168
Colon cancer classification	2	62	500

The experiments were carried out on the dataset from [37], which is composed of 40 colon tumor cases and 22 normal colon tissue samples, each characterized by the expression level of 2000 genes. As is common in gene expression microarray data analysis, a beneficial effect may be obtained by selecting a sub-group of genes, using prior knowledge that genes varying little across samples are less likely to be informative. Hence, we decided to perform the experiments by retaining the top 500 genes ranked by decreasing variance, as in [36].

5.2. Results and discussion

In this section, the obtained results are displayed. In order to highlight the different specific aspects of proposed experimental evaluation, the obtained results are organized in different tables:

1. In Table 2 the performances of the different IT kernels on the different generative embeddings are displayed. This represents the proposed approach, the table showing the performances of the different variants (different IT kernels, different generative embeddings).
2. In Table 3 we compare the performances of the IT kernels with respect to the standard kernels on the different generative embeddings. In particular we compared the best IT kernel (as obtained from the previous tables) with RBF and linear kernels. For every entry, we show the best classifier result (among NN and SVM). The σ parameter of RBF kernels has been adjusted by 5-fold cross-validation on the training set. This table aims at showing the contribution of the IT kernels in the proposed hybrid approach.
3. In Table 4 we compare the performances of the IT kernels on the generative embeddings with the same IT kernels in the original spaces. In particular we show best GE result for every IT kernel. This table aims at showing the contribution of the generative embedding step in the proposed hybrid approach.
4. Finally, in Table 5 we extract a summary of all methods: we displayed results obtained on the original space with standard kernels (linear and RBF kernels), on original space with IT kernels, on generative embedding space with standard kernels, and on generative embedding spaces with IT kernels. For every configuration we displayed the best result.

In all tables, "lin", "RBF", "JS", "JT", "WJT" represent the different kernels: linear, RBF, Jensen–Shannon, Jensen–Tsallis, and Weighted Jensen Tsallis, respectively, as described in Section 3. "NN" and "SVM" are the nearest neighbor and SVM classifiers, respectively. Finally, the acronyms of the generative embeddings follow the notation described in Section 2.2: "MT-2" is the mixture topics embedding for a single pLSA, "MT-1" is the posterior topic mixture

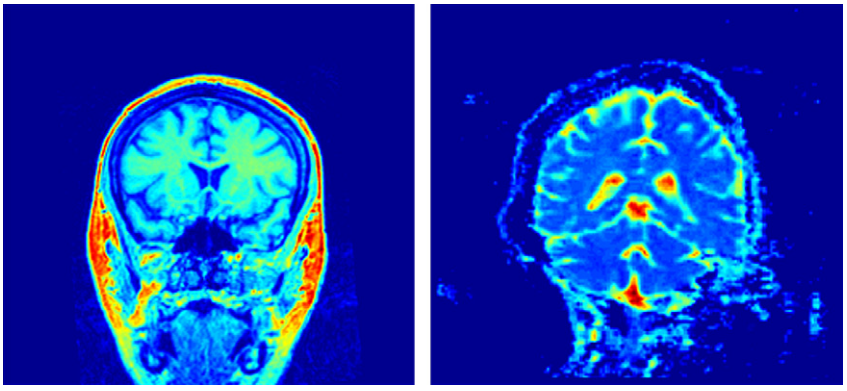


Fig. 1. Two MRI slices. Left: 3D morphological imaging. Right: diffusion weighting imaging.

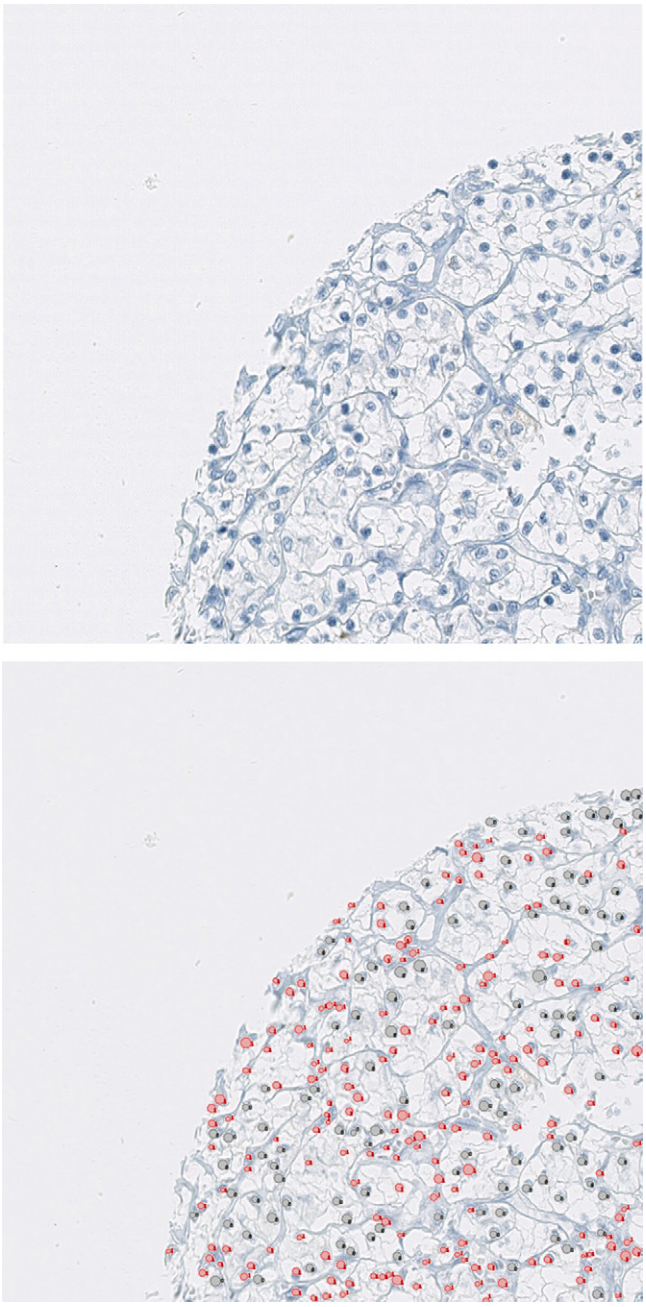


Fig. 2. Top: one quadrant (1500×1500 pixels) of a TMA spot image. Bottom: a pathologist exhaustively labeled all cell nuclei and classified them into malignant (black) and benign (red). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

with one pLSA per class, "FESS-2" is the *free energy score space* for a single pLSA, while "FESS-1" is the FESS using one pLSA per class, "LLR" is the log-likelihood ratio embedding, "FS" is the Fisher space, "TOP" refers to the TOP kernel, and "PD" is the posterior divergence embedding. Finally, "Orig" refers to the original space (namely the space without applying the generative embedding).

The results in the tables suggest the following observations:

- From Table 3 we can observe that, in many cases, the use of IT kernels with generative embeddings is moderately better than standard kernels over the same embeddings; the difference is quite clear in some cases. In particular, the main improvement is obtained in the application where the dictionary is large (e.g. colon cancer, which has 500 words). Actually this is reasonable, these kernels have been introduced in the linguistic scenarios [13], where the dictionary dimension is typically rather large.
- It is clear from the same table that the best generative embedding is the very recent *posterior divergence* (PD), which is outperformed only in few cases by other embeddings. This is confirmed over all applications. From a theoretical point of view, we observe this

Table 2
Accuracy rates of the different IT kernels over the different generative embeddings for the three applications (see the main text for details): (a) the brain MRI, (b) renal cancer, (c) colon cancer.

Embedding	JS–NN	JS–SVM	JT–NN	JT–SVM	WJT–NN	WJT–SVM
(a)						
TPM-1	0.542	0.596	0.503	0.627	0.584	0.643
TPM-2	0.589	0.689	0.543	0.658	0.631	0.702
FESS-1	0.569	0.500	0.500	0.369	0.584	0.500
FESS-2	0.627	0.600	0.500	0.674	0.601	0.720
LLR	0.588	0.616	0.500	0.638	0.614	0.636
FSH	0.584	0.702	0.553	0.673	0.619	0.699
TOP	0.519	0.500	0.500	0.500	0.500	0.500
PD	0.748	0.500	0.627	0.806	0.726	0.808
(b)						
TPM-1	0.648	0.742	0.612	0.741	0.632	0.742
TPM-2	0.660	0.742	0.595	0.733	0.625	0.743
FESS-1	0.643	0.706	0.619	0.688	0.630	0.702
FESS-2	0.653	0.736	0.609	0.743	0.625	0.744
LLR	0.640	0.765	0.577	0.765	0.607	0.763
FSH	0.660	0.760	0.581	0.745	0.611	0.754
TOP	0.632	0.684	0.616	0.686	0.620	0.687
PD	0.987	0.986	0.425	0.984	0.652	0.986
(c)						
TPM-1	0.775	0.816	0.739	0.861	0.768	0.857
TPM-2	0.774	0.862	0.772	0.868	0.800	0.878
FESS-1	0.711	0.675	0.683	0.635	0.700	0.670
FESS-2	0.744	0.822	0.717	0.826	0.726	0.830
LLR	0.713	0.778	0.676	0.755	0.688	0.774
FSH	0.777	0.862	0.773	0.856	0.800	0.875
TOP	0.705	0.669	0.672	0.676	0.692	0.674
PD	0.814	0.863	0.743	0.862	0.859	0.863

Table 3
Accuracy rates of different kernels over the different generative embeddings for the three applications.

GE	Brain MRI			Renal cancer			Colon cancer		
	Lin	RBF	IT	Lin	RBF	IT	Lin	RBF	IT
TPM-1	0.516	0.677	0.643	0.690	0.718	0.742	0.732	0.645	0.861
TPM-2	0.686	0.673	0.702	0.735	0.750	0.743	0.842	0.832	0.878
FESS-1	0.561	0.690	0.584	0.709	0.742	0.706	0.720	0.762	0.711
FESS-2	0.629	0.693	0.720	0.737	0.744	0.744	0.829	0.835	0.830
LLR	0.573	0.692	0.638	0.713	0.755	0.765	0.722	0.692	0.778
FSH	0.618	0.696	0.702	0.740	0.762	0.760	0.852	0.827	0.875
TOP	0.519	0.639	0.519	0.694	0.691	0.687	0.704	0.632	0.705
PD	0.752	0.702	0.808	0.976	0.825	0.987	0.814	0.842	0.863

Table 4
Accuracy rates of the different IT kernels over generative embedding and the original space for the three applications.

Method	Brain		Renal cancer		Colon cancer	
	Orig	GE	Orig	GE	Orig	GE
JS–NN	0.602	0.748	0.640	0.987	0.758	0.814
JS–SVM	0.743	0.702	0.742	0.986	0.769	0.863
JT–NN	0.503	0.627	0.627	0.619	0.660	0.773
JT–SVM	0.706	0.806	0.736	0.984	0.842	0.868
WJT–NN	0.500	0.726	0.607	0.652	0.659	0.859
WJT–SVM	0.738	0.808	0.734	0.986	0.816	0.878

Table 5
Summary of all possible variants for the three applications.

Task	Orig+StdK	GE+StdK	Orig+ITK	GE+ITK
Brain	0.770	0.752	0.743	0.808
Renal cancer	0.776	0.976	0.742	0.987
Colon cancer	0.829	0.852	0.842	0.878

Table 6
Summary of the best results and comparison with state-of-the-art methods.

Method/reference	Protocol	Accuracy
Brain MRI classification		
GE+ITK	Hold out	0.808
[33]	Leave one out	0.833
[38]	Leave one out	0.883
Renal cancer classification on TMA images		
GE+ITK	Hold out	0.987
[39]	10-fold CV	0.797
Colon cancer classification with gene expression microarray data		
GE+ITK	Hold out	0.878
[40]	10-fold CV	0.888
[41]	Leave one out	0.887
[42]	Leave one out	0.935
[43]	0.7/0.3 CV	0.873

generative embedding descriptor is very rich: like other score spaces (FS, FESS) it takes into account how well a sample fits the model and, like FESS, how uncertain the fitting is. Moreover it also assesses the change in model parameters brought on by the input sample, i.e. how much a sample affects the model. These three measures are not simply stacked together, but they are derived from the incremental EM algorithm which, in the E-step only looks at one or few selected samples to update the model in each iteration. Moreover, from the specific application scenarios, we tried to analyze this very appealing behavior by looking at the averaged dimensionality of the generative embedding spaces. There are 5 embeddings for which the average dimensionality is less than 100, for other three is more than 5000. PD has an average dimensionality of 1000, so possibly representing a good compromise between expressiveness and curse of dimensionality. Another point is that PD may slightly prefer pLSA. Actually, also in [19], the most remarkable improvement obtained with PD over other 3 embeddings (given a proper choice of the number of topics) was obtained with pLSA as generative model.

- From Table 4 we can observe that the use of a generative embedding is almost always beneficial with respect to the original space, when using IT kernels. In the Renal Cancer classification task the improvement is really impressive, going from a maximum performance in the original space of 0.742 to 0.987. This suggests that a generative embedding approach is really suited when the number of documents is high, so that the generative model can be adequately trained. (Renal cancer has 474 documents compared with 60 and 62 for the other two applications.)
- From Table 2 it seems that there is no significant difference among the various IT kernels, even if it may be argued that the weighted Jensen–Tsallis seems to have a slight advantage over the others.
- From the same table, it may be seen that there is not a huge difference between the performances of the SVM and the NN, which confirms the goodness of the devised similarity measures.
- In Table 5 it can be seen that the generative embedding plus IT kernels seems to be the best scheme in all the three applications, confirming the intuition that selecting a proper similarity measure in the generative embedding space may improve even more the performances.
- A summary of the best combination over the different schemes, together with some state-of-the-art results, is reported in Table 6. Even if the other results are obtained through a different protocol, it is evident that the proposed approach is in line with the results reported in the literature.
- A final observation is related to the behavior of the parameter q of the Jensen–Tsallis and the Weighted Jensen Tsallis kernels, which has been computed via cross-validation on the training set in all experiments. What we noticed is that the chosen q was almost always less than 1, which is in line with the results obtained in [13]. Although we do not have, at this moment, a formal justification for this fact, it may be due to the following behavior of the JT (and WJT) kernels. For $q<1$, the maximizer of $k_q^{JT}(p,v)$ (or of $k_q^{B}(p,v)$) with respect to p is not v, but another distribution closer to uniform. This is not the case for the Jensen–Shannon kernel k^{JS}, which coincides with k_1^{JT}, for which the minimizer of $k^{JS}(p,v)$ with respect to p is precisely v. This behavior of k_q^{JT} plays the role of a regularizer (favoring uniform distributions) on the multinomial.

6. Conclusions

In this paper, we have proposed to combine several generative embeddings (some of which very recently proposed) with information theoretical kernels, to obtain a new class of hybrid generative/discriminative methods to learning classifiers from data. The generative embeddings here considered are based on pLSA (probabilistic latent semantic analysis) modelling of the data, whereas the information theoretic kernels are based on a non-extensive version of information theory. We have tested the proposed approach on three medical classification problems; the reported experimental results are competitive with other state-of-the-art methods, showing that the proposed approach is promising and deserves further development.

Acknowledgments

We acknowledge support from the FET programme (EU FP7), under the SIMBAD project (contract 213250).

References

[1] A.Y. Ng, M.I. Jordan, On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes, in: Proceedings of the Conference on Advances in Neural Information Processing Systems, NIPS, 2001, vol. 14, pp. 841–848.
[2] Y.D. Rubinstein, T. Hastie, Discriminative vs informative learning, in: Proceedings of the International Conference on Knowledge Discovery and Data Mining, KDD '97, AAAI Press, 1997, pp. 49–53.
[3] B. Ripley, Pattern Recognition and Neural Networks, Cambridge University Press, 1996.
[4] T.S. Jaakkola, D. Haussler, Exploiting generative models in discriminative classifiers, in: Proceedings of the Conference on Advances in Neural Information Processing Systems, NIPS, 1998, vol. 11, Cambridge, MA, USA, pp. 487–493.
[5] J.A. Lasserre, C.M. Bishop, T.P. Minka, Principled hybrids of generative and discriminative models, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR '06, vol. 1, Washington, DC, USA, 2006, pp. 87–94.
[6] G. Bouchard, B. Triggs, The trade-off between generative and discriminative classifiers, in: Proceedings of the 16th IASC Symposium on Computational Statistics, 2004, pp. 721–728.
[7] A. Ferreira, M. Figueiredo, Hybrid generative/discriminative training of radial basis function networks, in: European Symposium on Artificial Neural Networks, ESANN, 2008, pp. 599–604.
[8] A. Fujino, N. Ueda, K. Saito, Semi-supervised learning for a hybrid generative/discriminative classifier based on the maximum entropy principle, IEEE Trans. Pattern Anal. Mach. Intell. 30 (3) (2008) 424–437.
[9] C. Sminchisescu, A. Kanaujia, D. Metaxas, Learning joint top-down and bottom-up processes for 3D visual inference, in: Proceedings of the IEEE CS Conference on Computer Vision and Pattern Recognition, 2006, pp. 1743–1752.

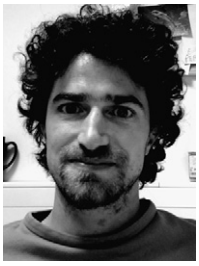
[10] M. Bicego, V. Murino, M.A. Figueiredo, Similarity-based classification of sequences using hidden Markov models, Pattern Recognition 37 (2004) 2281–2291.
[11] A. Bosch, A. Zisserman, X. Munoz, Scene classification via pLSA, in: Proceedings of the European Conference on Computer Vision, ECCV '06, 2006, pp. 517–530.
[12] A. Perina, M. Cristani, U. Castellani, V. Murino, N. Jojic, A hybrid generative/discriminative classification framework based on free-energy terms, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV '09, 2009, pp. 2058–2065.
[13] A.F.T. Martins, N.A. Smith, E.P. Xing, P.M.Q. Aguiar, M.A.T. Figueiredo, Nonextensive information theoretic kernels on measures, J. Mach. Learn. Res. 10 (2009) 935–975.
[14] A. Perina, M. Cristani, U. Castellani, V. Murino, N. Jojic, Free energy score space, in: Proceedings of the Conference on Advances in Neural Information Processing Systems, NIPS, 2009, vol. 22, pp. 1428–1436.
[15] G. Chandalia, M.J. Beal, Using fisher kernels from topic models for dimensionality reduction, in: NIPS Workshop on Novel Applications of Dimensionality Reduction, 2006.
[16] J.-C. Chappelier, E. Eckard, PLSI: the true fisher kernel and beyond, in: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases, ECML-PKDD '09, ECML PKDD '09, 2009, pp. 195–210.
[17] K. Tsuda, M. Kawanabe, G. Rätsch, S. Sonnenburg, K.-R. Müller, A new discriminative kernel from probabilistic models, Neural Comput. 14 (2002) 2397–2414.
[18] N. Smith, M. Gales, Speech recognition using SVMs, in: Proceedings of the Conference on Advances in Neural Information Processing Systems, NIPS, 2001, vol. 14, pp. 1197–1204.
[19] X. Li, T.S. Lee, Y. Liu, Hybrid generative–discriminative classification using posterior divergence, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR '11, 2011, pp. 2713–2720.
[20] N.D. Smith, M.J.F. Gales, Using SVMs to Classify Variable Length Speech Patterns, Technical Report CUED/F-INFENG/TR-412, Cambridge University Engineering Department, 2002.
[21] M. Bicego, E. Pekalska, D.M.J. Tax, R.P.W. Duin, Component-based discriminative classification for hidden Markov models, Pattern Recognition 42 (2009) 2637–2648.
[22] T. Hofmann, Unsupervised learning by probabilistic latent semantic analysis, Mach. Learn. 42 (2001) 177–196.
[23] T. Hofmann, Learning the similarity of documents: an information-geometric approach to document retrieval and categorization, in: Proceedings of the Conference on Advances in Neural Information Processing Systems, NIPS, 1999, pp. 914–920.
[24] M. Bicego, P. Lovato, B. Oliboni, A. Perina, Expression microarray classification using topic models, in: Proceedings of the 2010 ACM Symposium on Applied Computing, SAC '10, New York, NY, USA, 2010, pp. 1516–1520.
[25] U. Castellani, A. Perina, V. Murino, M. Bellani, G. Rambaldelli, M. Tansella, P. Brambilla, Brain morphometry by probabilistic latent semantic analysis, in: Proceedings of the International Conference on Medical Image Computing and Computer-assisted Intervention MICCAI '10, MICCAI, 2010, pp. 177–184.
[26] G. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (1979) 461–464.
[27] M. Cuturi, K. Fukumizu, J.-P. Vert, Semigroup kernels on measures, J. Mach. Learn. Res. 6 (2005) 1169–1198.
[28] T. Jebara, R. Kondor, A. Howard, Probability product kernels, J. Mach. Learn. Res. 5 (2004) 819–844.
[29] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, 2004.
[30] N.R. Giuliani, V.D. Calhouna, G.D. Pearlson, A. Francis, R.W. Buchanan, Voxel-based morphometry versus region of interest: a comparison of two methods for analyzing gray matter differences in schizophrenia, Schizophr. Res. 74 (2005) 135–147.
[31] C. Corradi-DellAcqua, L. Tomelleri, M. Bellani, G. Rambaldelli, R. Cerini, R. Pozzi-Mucelli, M. Balestrieri, M. Tansella, P. Brambilla, Thalamic-insular dysconnectivity in schizophrenia: evidence from structural equation modeling, Human Brain Mapp. 33 (2012) 740–752.
[32] J. Sun, M. Ovsjanikov, L. Guibas, A concise and provably informative multi-scale signature based on heat diffusion, in: Proceedings of the Symposium on Geometry Processing, SGP '09, 2009, pp. 1383–1392.
[33] U. Castellani, P. Mirtuono, V. Murino, M. Bellani, G. Rambaldelli, M. Tansella, P. Brambilla, A new shape diffusion descriptor for brain classification, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI '11, 2011.
[34] P. Schüffler, T. Fuchs, C.S. Ong, V. Roth, J. Buhmann, Computational TMA analysis and cell nucleus classification of renal cell carcinoma, in: Proceedings of the 32nd DAGM Conference on Pattern recognition, Springer, 2010, pp. 202–211.
[35] A. Bosch, A. Zisserman, X. Munoz, Representing shape with a spatial pyramid kernel, in: Proceedings of the 6th ACM International Conference on Image and Video Retrieval, CIVR '07, 2007, pp. 401–408.
[36] S. Rogers, M. Girolami, C. Campbell, R. Breitling, The latent process decomposition of cDNA microarray data sets, IEEE/ACM Trans. Comput. Biol. Bioinf. 2 (2005) 143–156.
[37] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, A. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proc. Natl. Acad. Sci. 96 (1999) 6745–6750.
[38] U. Castellani, A. Ulaş, V. Murino, M. Bellani, M. Tansella, P. Brambilla, Selecting scales by multiple kernel learning for shape diffusion analysis, in: MICCAI Workshop on "Mathematical Foundations of Computational Anatomy", MFCA '11, 2011, pp. 148–158.
[39] A. Ulaş, P.J. Schüffler, M. Bicego, U. Castellani, V. Murino, Hybrid generative–discriminative nucleus classification of renal cell carcinoma, in: M. Pelillo, E.R. Hancock (Eds.), Proceedings of the International Workshop on Similarity-Based Pattern Analysis—SIMBAD '11, Lecture Notes in Computer Science, vol. 7005, Springer, 2011, pp. 77–88.
[40] S. Deegalla, H. Bostrom, Fusion of dimensionality reduction methods: a case study in microarray classification, in: Proceedings of the International Conference on Information Fusion, 2009, pp. 460–465.
[41] D. German, B. Afsari, T. Choon, D. Naiman, Microarray classification from several two-gene expression comparisons, in: Proceedings of the International Conference on Machine Learning and Applications, 2008, pp. 583–585.
[42] H. Liu, L. Liu, H. Zhang, Ensemble gene selection by grouping for microarray data classification, J. Biomed. Inf. 43 (2010) 81–87.
[43] L. Wang, J. Zhu, H. Zou, Hybrid huberized support vector machines for microarray classification and gene selection, Bioinformatics 24 (2008) 412–419.

Manuele Bicego received his Laurea degree and Ph.D. degree in Computer Science from University of Verona in 1999 and 2003, respectively. From 2004 to 2008 he was at the University of Sassari. Currently he is assistant professor (ricercatore) at the University of Verona. From June 2009 to February 2011 he was also team leader at the Istituto Italiano di Tecnologia (IIT, Genova, Italy).

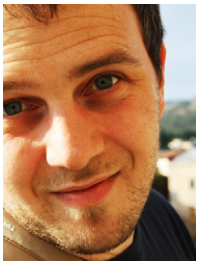
His research interests include statistical pattern recognition, mainly probabilistic models (GMM, HMM) and kernel machines (e.g. SVM), with application to video analysis, biometrics and, recently, bioinformatics. He is author of several papers in the above subjects, published in international journals and conferences. He is an associate editor of ELCVIA and JoI journals. He has served as member of the scientific committee of different international conferences, and he is a reviewer for several international conferences and journals.

He is member of the IEEE Systems, Man, and Cybernetics society and of the IAPR Society, Italian Chapter (GIRPR).

Aydin Ulaş received his B.S., M.S. and Ph.D. degrees in computer science from Boğaziçi University, Istanbul, in 1999, 2001 and 2008 respectively. He is currently doing his postdoc in the University of Verona, Italy and working on the FP7 project SIMBAD (Similarity based Pattern Analysis and Recognition). His research interests include model selection, classifier combination, statistical comparison of classification algorithms, medical imaging, bioinformatics, and machine learning. He is a reviewer for several international conferences and journals and he is a member of IEEE Computational Intelligence Society and IAPR Turkish chapter (TÖTIAD).

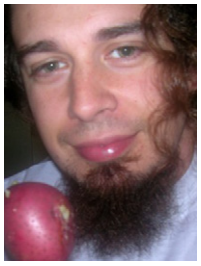


Umberto Castellani is Ricercatore (i.e., Research Assistant) of the Department of Computer Science at the University of Verona. He received his Dottorato di Ricerca (Ph.D.) in Computer Science from the University of Verona in 2003 working on 3D data modelling and reconstruction. During his Ph.D., he had been a Visiting Research Fellow at the Machine Vision Unit of the Edinburgh University, in 2001. In 2007, he was an Invited Professor at the LASMEA Laboratory in Clermont-Ferrand, France. In 2008, he was a Visiting Researcher at the PRIP Laboratory at Michigan State University (USA). His research is focused on 3D data processing, statistical learning, and medical image analysis. He has coauthored several papers which were published in leading conference proceedings and journals. He is a member of Eurographics and IEEE.



Alessandro Perina received the Ph.D. degree in Computer Science from the University of Verona with a thesis on classification with generative models. From 2006 to 2010 he has been member of the Vision, Image Processing and Sound group (VIPS) at the University of Verona. He is now a Postdoctoral researcher at Microsoft Research, Redmond working with the eScience group. His research interests are in computer vision and machine learning.

Vittorio Murino is a Full Professor with the Department of Computer Science, University of Verona, Italy, and Head of Computer Imaging facility at the Italian Institute of Technology (IIT), Genova, Italy. He is author or co-author of more than 180 papers in the fields of computer vision and pattern recognition (in particular, probabilistic techniques for image and video processing), with applications on video surveillance, biomedical image analysis, and recently, bioinformatics.



André T. Martins receive the E.E. degree from Instituto Superior Técnico (IST), the engineering school of the Technical University of Lisbon. In 2004 he joined PRIBERAM, where he has been doing R&D in natural language processing. In 2006, he became a Ph.D. student enrolled in the dual (Carnegie Mellon University and IST) Ph.D. program in Language and Information Technologies. His research interests are in machine learning, with a special focus on structured prediction and natural language processing. He has published in the most prestigious journals and conferences of the area (JMLR, IEEETPAMI, ICML, AISTATS, EMNLP, ACL) and recently co-organized the first edition of the Lisbon Machine Learning School (LxMLS' 2011).

Pedro M.Q. Aguiar received Ph.D. degree in electrical and computer engineering from the Instituto Superior Técnico, Technical University of Lisbon, Lisbon, Portugal, in 2000. He is currently an Assistant Professor with the Instituto Superior Técnico, Technical University of Lisbon, Lisbon, Portugal. He is also affiliated with the Institute for Systems and Robotics, Lisbon, Portugal, and has been Visiting Scholar with Carnegie-Mellon University, Pittsburgh, PA, and a Consultant with Xerox Palo Alto Research Center, Palo Alto, CA. His main research interests are in image analysis and computer vision.

Mário A.T. Figueiredo received E.E., M.Sc., and Ph.D. degrees in electrical and computer engineering, all from Instituto Superior Técnico (IST), Technical University of Lisbon, Portugal, in 1985, 1990, and 1994, respectively. Since 1994, he has been with the Department of Electrical and Computer Engineering of IST, where he is now a full Professor. He is also a researcher and area coordinator at Instituto de Telecomunicações. His interests include statistical pattern recognition, machine learning, image processing, and computer vision. He is a Fellow of the IEEE and of the IAPR. He received the Portuguese IBM Scientific Prize in 1995 and the IEEE Signal Processing Society Best Paper Award of 2011. He is/was associate editor of several journals, including the IEEE Transactions on Image Processing, the IEEE Transactions on Pattern Analysis and Machine Intelligence, and Pattern Recognition Letters. He was guest co-editor of special issues of several IEEE transactions and journals.