COMBINING FREE ENERGY SCORE SPACES WITH INFORMATION THEORETIC KERNELS: APPLICATION TO SCENE CLASSIFICATION

M. Bicego[1,2]*, A. Perina*[1]*, V. Murino*[1,2]*, A. Martins*[3]*, P. Aguiar*[4]*, M. Figueiredo*[3]

[1] *Dipartimento di Informatica, University of Verona - Verona, Italy*
[2] *Istituto Italiano di Tecnologia (IIT) - Genova, Italy*
[3] *Instituto de Telecomunicações, Instituto Superior Técnico - Lisboa, Portugal*
[4] *Instituto de Sistemas e Robótica, Instituto Superior Técnico - Lisboa, Portugal*

ABSTRACT

Most approaches to learn classifiers for structured objects (*e.g.*, images) use generative models in a classical Bayesian framework. However, state-of-the-art classifiers for vectorial data (*e.g.*, support vector machines) are learned discriminatively. A generative embedding is a mapping from the object space into a fixed dimensional score space, induced by a generative model, usually learned from data. The fixed dimensionality of these generative score spaces makes them adequate for discriminative learning of classifiers, thus bringing together the best of the discriminative and generative paradigms. In particular, it was recently shown that this hybrid approach outperforms a classifier obtained directly for the generative model upon which the score space was built.

Using a generative embedding involves two steps: (i) defining and learning the generative model and using it to build the embedding; (ii) discriminatively learning a (maybe kernel) classifier on the adopted score space. The literature on generative embeddings is essentially focused on step (i), usually using some standard off-the-shelf tool for step (ii). In this paper, we adopt a different approach, by focusing also on the discriminative learning step. In particular, we combine two very recent and top performing tools in each of the steps: (i) the free energy score space; (ii) non-extensive information theoretic kernels. In this paper, we apply this methodology in scene recognition. Experimental results on two benchmark datasets shows that our approach yields state-of-the-art performance.

Index Terms— Scene categorization, generative embeddings, score spaces, information theoretic kernels.

1. INTRODUCTION

Most approaches to the statistical learning of classifiers belong to one of two classical paradigms: generative and discriminative [12, 15]. 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 generative class models.

In the past decade, several hybrid generative-discriminative approaches have been proposed with the goal of taking advantage of the best of both paradigms [5, 8]. In this context, significant interest has been aroused by the so-called generative score space methods (or generative embeddings), where the basic idea is to exploit a generative model to map the objects to be classified into a feature space, where discriminative techniques, namely kernel-based, can be used. This is particularly suitable to deal with non-vectorial data (strings, trees, images), since it maps objects (maybe of different dimensions) into a fixed dimension space.

The seminal work on generative embeddings is [5], 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 [1, 2, 13].

A very recent approach, termed *free energy score space* (FESS) [13, 14], has shown to outperform other generative embeddings (including those in [5] and [17]) in several applications. The FESS expresses how well each data point fits different parts of the generative model, using the variational free energy as a lower bound on the negative log-likelihood. It has been found that the FESS is highly informative for discriminative learning, yielding state-of-the-art results in several computational biology and computer vision problems (namely, scene/object recognition) [13, 14].

Typically, the feature vectors resulting from the generative embedding (FESS or others) are used to feed some kernel-based classifier, namely, a *support vector machine* (SVM) with simple linear or radial basis function (RBF) kernels. In this paper, we propose to combine the FESS embedding with the recently introduced information theoretic (IT) kernels [11]. These new kernels, which are based on a non-extensive generalization of the classical Shannon information theory, are defined on (possibly unnormalized) probability measures. In [11], they were successfully used in text categorization tasks, based on multinomial (bag-of-words type) text representations. Here, the idea is to consider the points of the FESS as multinomial probability distributions, thus valid arguments for the information theoretic kernels.

We illustrate the excellent performance of combining the FESS embedding with the IT kernels on the scene classification problem [2, 9]. For this problem, the FESS embedding is built on a topic-based generative model known as *probabilistic latent semantic analysis* (pLSA) [4]. The good performance of pLSA in generative approaches to scene classification suggests that it is an adequate generative model for this problem. In [13, 14], it was shown that a classifier defined on the pLSA-based FESS outperforms a purely generative pLSA-based method. The experimental results reported in this paper show that the improvement is even more significant when the FESS is combined with the non-extensive IT kernels to build SVM and K-nearest neighbors (K-NN) classifiers. On two challenging scene classification problems – the Vogel-Schiele dataset (a.k.a. Corel, [18]) and the Fei Fei – Perona dataset [9] – our classifiers significantly outperform the current state-of-the-art methods.

The rest of the paper is organized as follows. Section 2, reviews the FESS approach. Section 3 describes the IT kernels. The proposed way of using the information theoretic kernels on the FESS is formalized in Section 4. Experiments on scene classification are reported in Section 5, and Section 6 concludes the paper.

2. FREE ENERGY SCORE SPACES

A generative model is essentially a joint distribution $P(y, x|\theta)$ for a pair of (usually vector) random variables, where θ contains the model parameters. Consider a set $\mathbf{x} = \{x^1, \ldots, x^T\}$ of i.i.d. observations from this distribution and the corresponding set of $\mathbf{y} = \{y^1, \ldots, y^T\}$ of hidden variables. Under the i.i.d. assumption, $P(\mathbf{y}, \mathbf{x}|\theta) = \prod_{t=1}^{T} P(y^t, x^t|\theta)$.

Given $\mathbf{x}$, the usual goals are to estimate θ (learning) and the hidden variables $\mathbf{y}$ (inference), but these tasks may be computationally intractable. Variational inference overcomes this difficulty by approximating the exact posterior $P(\mathbf{y}|\mathbf{x}, \theta)$ by a function $Q(\mathbf{y}) \in \mathcal{Q}$, where $\mathcal{Q}$ is a family of distributions allowing tractable inference. For example, under the i.i.d. assumption herein adopted, the common choice is to use factorized distributions $Q(\mathbf{y}) = \prod_{t=1}^{T} q_t(y^t)$. The so-called *free energy* $\mathcal{F}(Q, \theta)$ [7] is simply the *Kullback-Leibler divergence* (KLD) between the approximation Q and the exact posterior $P(\mathbf{y}|\mathbf{x}, \theta)$, plus a function independent of Q, that is,

$$\begin{aligned}\mathcal{F}(Q, \theta) &= D_{\text{KL}}\left[Q(\mathbf{y}) \| P(\mathbf{y}|\mathbf{x}, \theta)\right] - \ln P(\mathbf{x}|\theta) \\ &= \sum_{y} Q(\mathbf{y}) \ln \frac{Q(\mathbf{y})}{P(\mathbf{y}, \mathbf{x}|\theta)}. \qquad (1)\end{aligned}$$

The KLD is always non-negative, it is zero iff its arguments are equal, and is convex. Therefore, minimizing $\mathcal{F}(Q, \theta)$ with respect to Q leads to the best approximation in $\mathcal{Q}$ to the true posterior $P(\mathbf{y}|\mathbf{x}, \theta)$. The minimization of $\mathcal{F}(Q, \theta)$ is carried out by alternating minimization with respect to Q and θ, while holding the other fixed [7]. If $P(\mathbf{y}|\mathbf{x}, \theta) \in \mathcal{Q}$, this coincides with the standard *expectation-maximization* (EM) algorithm.

As pointed out in [13, 14], in the presence of i.i.d. data, we can assume $P(\mathbf{y}, \mathbf{x}|\theta) = \prod_t P(y^t, x^t|\theta)$, and using a factorized approximation $Q(\mathbf{y}) = \prod_t q_t(y^t)$, the free energy in Eq. (1) can be decomposed as $\mathcal{F}(Q, \theta) = \sum_t \mathcal{F}^t(q_t, \theta)$, where

$$\mathcal{F}^t(q_t, \theta) = \sum_{y^t} q_t(y^t)\, \ln q_t(y^t) - \sum_{y^t} q_t(y^t) \ln P(y^t, x^t|\theta).$$

The first term in $\mathcal{F}^t$ is the entropy of the variational approximation to the posterior. The second term, which has the form of a cross-entropy, measures how well observation x^t is explained by the model for that θ.

If the joint distribution $P(y, x|\theta)$ itself factorizes due to its internal conditional dependency structure, it is possible to further decompose each $\mathcal{F}^t$ into a sum of local terms [13]. For example, if the generative model is described by a Bayesian network, its joint distribution can be written as

$$P(y, x|\theta) = P(v|\theta) = \prod_{i=1}^{M} P(v_i|\mathbf{PA}_i, \theta),$$

where $v = \{y, x\}$ is the entire set of variables (hidden and visible), and $\mathbf{PA}_i$ is the set of parents of v_i. Using this factorization, each $\mathcal{F}^t$ can then be written as a summation of terms

$$\mathcal{F}^t = E + \sum_{i=1}^{M} f_i(x^t, \theta),$$

where E represent the entropy and

$$f_i(x^t, \theta) = -\sum_{y^t} q_t(y^t) \ln P(v_i^t|\mathbf{PA}_i, \theta).$$

Notice that each v_i^t is either a component of x^t (an observed variable) or of y^t (a hidden variable). If it is a component of x^t, it is fixed at the observed value; if it is a component of y^t, it is being marginalized out by the summation.

The same idea can be used to write the entropy E as a summation of terms $f_i(x^t, \theta)$; differenlty from the cross-entropy, E is broken following the factorization of the posterior distribution $q_t(y^t)$. For further detail see [13, 14].

The decomposition shown in the previous paragraph suggests the following feature extractor, as proposed in [13]: a function ζ that, for a choice of θ, maps a given point x into $\mathbb{R}^M$ according to

$$\zeta(x, \theta) = [f_1(x, \theta), \ldots, f_M(x, \theta)].$$

where each $f_i(x, \theta)$ comes either from entropy or cross-entropy. This feature extractor may be used in several different ways. For example, in a classification problem, we learn a parameter estimate $\hat{\theta}_n$ for each of the N classes, using the above mentioned EM-type algorithm. Given this collection of parameter estimates, $\Theta = \{\hat{\theta}_1, \ldots, \hat{\theta}_N\}$, the corresponding ζ functions are concatenated to form the score vector (belonging to the *free energy score space*),

$$\phi_{\Theta}^{FE}(x) = \Big[\zeta(x, \hat{\theta}_1), \ldots, \zeta(x, \hat{\theta}_N)\Big] \in \mathbb{R}^{NM}. \qquad (2)$$

In [13, 14], these feature vectors were used to build SVM classifiers with classical *radial basis function* (RBF) kernels.

3. INFORMATION THEORETIC KERNELS

Kernels on probability measures have been shown very effective in classification problems involving text, images, and other types of data [3, 6]. Given two probability measures p_1 and p_2, representing two objects, several information theoretic kernels (ITKs) can be defined [11]. The Jensen-Shannon kernel is defined as

$$k^{\text{JS}}(p_1, p_2) = \ln(2) - JS(p_1, p_2), \qquad (3)$$

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

$$JS(p_1, p_2) = H\Big(\frac{p_1 + p_2}{2}\Big) - \frac{H(p_1) + H(p_2)}{2}, \qquad (4)$$

where $H(p)$ is the usual Shannon entropy.

The Jensen-Tsallis (JT) kernel is given by

$$k_q^{\text{JT}}(p_1, p_2) = \ln_q(2) - T_q(p_1, p_2), \qquad (5)$$

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

$$T_q(p_1, p_2) = S_q\left(\frac{p_1 + p_2}{2}\right) - \frac{S_q(p_1) + S_q(p_2)}{2^q} \qquad (6)$$

is the Jensen-Tsallis q-difference, and $S_q(r)$ is the Jensen-Tsallis entropy, defined, for a multinomial $r = (r_1, ..., r_L)$, with $r_i \geq 0$ and $\sum_i r_i = 1$, as

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

In [11], 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 versions of the JT kernels are defined as follows:

- The weighted JT kernel (version A) is given by

$$k_q^A(\mu_1, \mu_2) = S_q(\pi) - T_q^\pi(p_1, p_2), \tag{7}$$

where $\pi = (\pi_1, \pi_2) = \left(\frac{\omega_1}{\omega_1+\omega_2}, \frac{\omega_2}{\omega_1+\omega_2}\right)$ and

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

- The weighted JT kernel (version B) is defined as

$$k_q^B(\mu_1, \mu_2) = \left(S_q(\pi) - T_q^\pi(p_1, p_2)\right)(\omega_1 + \omega_2)^q. \tag{8}$$

4. PROPOSED APPROACH

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

$$k(x, x') = k_q^i\left(\phi_\Theta^{FE}(x), \phi_\Theta^{FE}(x')\right), \tag{9}$$

where $i \in \{\text{JT, A, B}\}$ indexes one of the Jensen-Tsallis kernels (5), (7), or (8), and ϕ_Θ^{FE} is the free energy score function defined in (2). Notice that this kernel is well defined because all the components of ϕ_Θ^{FE} are non-negative.

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 [11] 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 [16, Proposition 3.22] guarantee that the kernel k defined in (9) inherits the positive definiteness of k_q^i, thus can be safely used in SVM learning algorithms.

Although this approach is applicable to any classification problem, in this paper we consider the image interpretation/analysis task studied in [13]: scene categorization.

For each image, SIFT feature vectors [10] are computed from 16×16 patches, centered on a grid with 8 pixels spacing. The SIFT features obtained from a training set of images is clustered (using K-means) to yield a codebook of W "visual words". With this codebook, a given image can be represented by a W-dimensional vector, resembling a bag-of-words representation of a text document: the i-th entry counts the number of SIFT vectors that are closer to the i-th element of the codebook than to any other element.

As a generative model, we adopt pLSA [4], which has been shown to yield good performance in scene classification with purely generative approaches [9, 2]. The pLSA model is essentially a mixture of Z multinomial components (usually called latent topics) representing the joint distribution of words and document classes. In each component of this mixture, the words and classes are conditionally independent, although of course they are globally dependent.

5. EXPERIMENTAL EVALUATION

The proposed approach has been firstly tested on the Vogel and Schiele dataset [18], also known as Corel – sample images are shown in Fig. 1. This dataset (with 6 classes) is originally divided into training and testing subsets. We used one half of the training subset to obtain a codebook of $W = 175$ "visual words" and to estimate N pLSA models (each with $Z = 40$ latent topics), one for each class. The FESS features (2) are straightforward to extract from the description of pLSA in [4]. Finally, the second half of training set is used to learn the classifier (or simply to define it, in the case of K-NN). The constant C of the SVM learning algorithm is optimized by 10-fold cross validation (CV); parameter K of K-NN is either set to 1 or tuned by 10-fold CV.

Fig. 1. Sample images from the Corel dataset.

Table 1. Average (over 10 runs) classification accuracies (standard deviation in parenthesis) for the COREL dataset.

	SVM	1-NN	K-NN
$k^{\text{JS}} = k_1^{\text{JT}}$	0.899 (0.007)	0.633 (0.012)	0.627 (0.015)
k_q^{JT} (auto q)	0.931 (0.006)	0.629 (0.011)	0.610 (0.014)
k_q^{JT} (best q)	0.928 (0.006)	0.635 (0.008)	0.628 (0.012)
k_q^A (auto q)	0.943 (0.005)	0.639 (0.013)	0.634 (0.013)
k_q^A (best q)	0.944 (0.005)	0.645 (0.012)	0.640 (0.010)

The classification accuracies on the testing subset, shown in Table 1, are averages over 10 repetitions of the experiments. In that table, "best q" means that the value of q used in the JT kernels is the one leading to the highest accuracy, while "auto q" means that this parameter was adjusted by 10-fold CV. We omit results with k_q^B, since k_q^A always yielded better accuracies. These results show that the SVM classifier clearly outperforms the K-NN. Another important conclusion is that the performance with the value of q chosen by CV ("auto-q") is very close to that obtained with the "best-q".

Figure 2 plots the SVM accuracies, for different kernels, as a function of parameter q. The plot also shows the accuracies obtained with q chosen by cross-validation. In line with the results from [11], the best performances are obtained for $q < 1$.

Table 2 compares our results with others found in the literature or computed by us. The accuracy obtained by the proposed method is better than that of [13], which, to the best of our knowledge, was the state-of-the-art on this dataset.

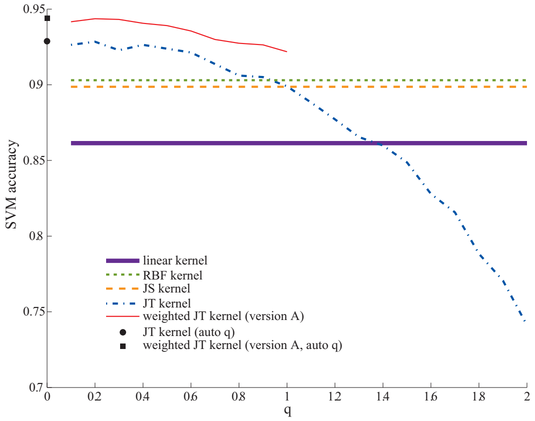

Fig. 2. SVM accuracies, on the COREL dataset, with several kernels, as a function of q.

A further set of experiments was performed using the Fei Fei – Perona dataset [9], with the same experimental protocol proposed in [13]. The results in Table 3 confirm the effectiveness of the proposed approach, in comparison with the state-of-the-art.

Table 2. Accuracies of several classifiers on the COREL dataset.

Method	Accuracy
[18]	0.751
[2] (pLSA)	0.857
[13] (pLSA + FESS + SVM-RBF)	0.903
pLSA + FESS + SVM-linear kernel	0.861
pLSA + FESS + SVM-polynomial kernel	0.830
pLSA + FESS + logistic regression	0.762
pLSA + FESS + sparse logistic regression	0.792
k_q^B (best q, not usable in practice)	0.944
k_q^B (auto q)	**0.943**

6. CONCLUSIONS

We have proposed using information theoretic kernels in combination with the free energy score space embedding for discriminative learning of classifiers. We applied the proposed methodology to scene categorization tasks, using SIFT features and a pLSA model to build the FESS embedding. On challenging benchmark datasets (Corel and Fei Fei – Perona), we obtained state-of-the-art accuracy.

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

7. REFERENCES

[1] M. Bicego, V. Murino, M. Figueiredo, "Similarity-based classification of sequences using hidden Markov models", *Pattern Recognition*, vol. 37, pp. 2281–2291, 2004.

[2] A. Bosch, A. Zisserman, X. Munoz, "Scene classification via pLSA," *Proc. ECCV*, Graz, Austria, 2006.

Table 3. Classification accuracies (together with some other state of the art results) for the Fei Fei Perona dataset.

IT Kernel	SVM
linear	0.808
$k^{\text{JS}} = k_1^{\text{JT}}$	0.874
k_q^{JT} (auto q)	0.891
k_q^{JT} (best q)	0.891
k_q^A (auto q)	0.880
k_q^A (best q)	0.883

Reference	Accuracy
[9]	0.652
[2]	0.734
[19]	0.811
[13]	0.843
our best (k_q^{JT})	0.891

[3] M. Cuturi, K. Fukumizu, J. Vert, "Semigroup kernels on measures," *JMLR*, vol. 6, pp. 1169–1198, 2005.

[4] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. 22nd Int. ACM Conf. on Res. and Dev. in Info. Retr. – SIGIR*, Berkeley, 1999.

[5] T. Jaakkola, D. Haussler, "Exploiting generative models in discriminative classifiers," *NIPS*, 1999.

[6] T. Jebara, R. Kondor, A. Howard, "Probability product kernels," *JMLR*, vol. 5, pp. 819–844, 2004.

[7] M. Jordan, Z. Ghahramani, T. Jaakkola, L. Saul, "An introduction to variational methods for graphical models," *Mach. Learn.*, vol. 37, pp. 183–233, 1999.

[8] J. Lasserre, C. Bishop, T. Minka, "Principled hybrids of generative and discriminative models," *IEEE CVPR*, New York, 2006.

[9] L. Fei Fei, P. Perona, "A Bayesian hierarchical model for learning natural scene categories,"*IEEE CVPR*, San Diego, 2005.

[10] D. Lowe, "Object recognition from local scale-invariant features", *IEEE ICCV*, Kerkyra, Greece, 1999.

[11] A. Martins, N. Smith, E. Xing, P. Aguiar, M. Figueiredo, "Nonextensive information theoretic kernels on measures," *JMLR*, vol. 10, pp. 935–975, 2009.

[12] A. Ng, M. Jordan, "On discriminative vs generative classifiers: A comparison of logistic regression and naive Bayes," *NIPS*, 2002.

[13] A. Perina, M. Cristani, U. Castellani, V. Murino, N. Jojic, "A hybrid generative/discriminative classification framework based on free-energy terms," *IEEE ICCV*, Kyoto, Japan, 2009.

[14] A. Perina, M. Cristani, U. Castellani, V. Murino, N. Jojic, "Free energy score space", *NIPS*, 2009.

[15] Y. Rubinstein, T. Hastie, "Discriminative vs informative learning", *Proc. 3rd Int. Conf. on Knowledge Disc. and Data Min. – KDD*, Newport Beach, 1997.

[16] J. Shawe-Taylor, N. Cristianini, *Kernel Methods for Pattern Recognition*, Cambridge University Press, 2004.

[17] K. Tsuda, M. Kawanabe, G. Rätsch, S. Sonnenburg, K.-R. Müller, "A new discriminative kernel from probabilistic models," *Neural Computation*, vol. 14, pp. 2397–2414, 2002.

[18] J. Vogel, B. Schiele, "Semantic modeling of natural scenes for content-based image retrieval," *IJCV*, vol. 72, pp. 133–157, 2007.

[19] S. Lazebnik, C. Schmid J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," *CVPR*, 2006.