2010 International Conference on Pattern Recognition

Nonlinear mappings for generative kernels on latent variable models

Anna Carli*, Manuele Bicego*†, Sisto Baldo* and Vittorio Murino*†
*Dipartimento di Informatica
Università di Verona
Verona, Italy
Email: anna.carli–manuele.bicego–sisto.baldo–vittorio.murino@univr.it
†Istituto Italiano di Tecnologia (IIT)
Genova, Italy

***Abstract*—Generative kernels have emerged in the last years as an effective method for mixing discriminative and generative approaches. In particular, in this paper, we focus on kernels defined on generative models with latent variables (e.g. the states in a Hidden Markov Model). The basic idea underlying these kernels is to compare objects, via a inner product, in a feature space where the dimensions are related to the latent variables of the model. Here we propose to enhance these kernels via a nonlinear normalization of the space, namely a nonlinear mapping of space dimensions able to exploit their discriminative characteristics. In this paper we investigate three possible nonlinear mappings, for two HMM-based generative kernels, testing them in different sequence classification problems, with really promising results.**

***Keywords*-nonlinear mappings; generative kernels;**

I. Introduction

In recent years, generative kernels [1]–[4] have emerged as an approach to mix generative methods (like Hidden Markov Models or Bayesian Networks) and discriminative techniques (like Support Vector Machines). Generally speaking, there is a proved complementarity of discriminative and generative estimations: asymptotically (in the number of labelled training examples), classification error of discriminative methods is lower than that of generative ones [5], comparing logistic regression and naive Bayes classifiers. On the other side, generative counterparts are effective with less, possibly unlabelled, data.

In this paper, we focus on a particular class of generative kernels, namely kernels defined on generative models with latent variables (for example the states in a Hidden Markov Model – HMM): the most famous example is the Marginalized Kernel [4]. Very recently, another kernel has been proposed to be used with HMMs, called State-Space Kernel [6]. The idea of this class of kernels is to map the objects of the problem in a space where each dimension (or a set of dimensions) describes the contribution of one of the latent variables of the model. For example, in the State-Space, each direction measures how often the system is in a particular state given the model and the observation. The inner product in such generative-derived spaces typically represents the kernel.

The main idea of the approach described here derives from the fact that the different directions of the generative space (which are related to latent variables) could have different characteristics in terms of discriminative and descriptive power, and some space transformations might be useful. For example, the well known Fisher Kernel has been improved by a space normalization in [7]; moreover, the Marginalized Kernel does not work without a re-scaling of the space[1], as shown in [4]. The common characteristic of all these space transformations is the linearity of the scaling function. Nevertheless there are situations where the linearity assumption is too restrictive, and a benefit may be obtained from a nonlinear scaling via a nonlinear mapping.

In this paper, we investigate this last solution, modifying kernels in order to include a nonlinear normalization, namely, a nonlinear mapping of space dimensions able to highlight or exploit their discriminative characteristics. The specific form of such kernel depends on the choice of the nonlinear mapping and on the latent variable model it relies upon: here, we focus on HMM-based generative embeddings.

In a preliminary work [8], we investigated a possible choice of a nonlinear mapping, based on a powering operation, obtaining promising results. Encouraged by these performances, in this paper we pursue a further study, comparing the powering operation with two other different nonlinear mappings: the logarithm and the logistic function. Even if all of these functions, in principle, are able to equilibrate the contributions of each latent variable of the model, they have different characteristics, which will be discussed and analyzed in the paper. A thorough experimental evaluation has been used to validate our intuitions, using the Marginalized and the State-Space kernels in a SVM-based classification framework involving three tasks (two 2-D shape recognition problems and one gesture classification task). In the following, the basic theory and the nonlinear transformations considered are described in Sections 2 and 3, respectively. In Section 4, the experimental trials on three

[1]It is straightforward to extract the generative embedding – namely the space – from the kernel defined in [4] – see [8].

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.523

IEEE computer society

data sets are reported, together with a discussion of the results achieved, also sketching the future development of the approach.

II. The general idea

We consider a particular class of generative kernels which lie on latent variables. An object x is mapped in a vectorial space through the model components, i.e. the latent variables. The features of the resulting space $\mathcal{H}$ summarize information about how latent variables describe or model the observation x. We will call this information in the resulting space as $\boldsymbol{g_h}$ – the pedix $\boldsymbol{h}$ highlights the dependence of such information from the latent variables $\boldsymbol{h}$, where $\boldsymbol{h} = \{h_1, \ldots, h_N\}$ denotes the set of hidden variables of the generative latent model.

The kernel, typically defined as the inner product in the resulting Hilbert space, may then be decomposed into a sum of inner products, each related to a specific latent variable:

$$K(x, x') = \langle \boldsymbol{g_h}(x), \boldsymbol{g_h}(x') \rangle = \sum_{i=1}^{N} \langle \boldsymbol{g}_{h_i}(x), \boldsymbol{g}_{h_i}(x') \rangle \quad (1)$$

where $\boldsymbol{g}_{h_i}(x)$ denotes a vector of features related to a particular hidden variable h_i.

This formulation can in principle be applied to any generative model with latent variables which are used to form features. In the Marginalized Kernel case [4] $\boldsymbol{g}_{h_i}$ is a vector of length equal to the number of symbols in the alphabet, while in the State-Space Kernel [6] $\boldsymbol{g}_{h_i}$ is a scalar value. The basic idea of the approach we propose here is that the nonlinear mapping of the different directions of the derived generative space may highlight their discriminative characteristics [8]. This is accomplished by performing a nonlinear mapping f of dimensions of the original Hilbert space. The new kernel formulation is then defined as

$$NK(x, x') = \sum_{i=1}^{N} \langle f\left(\boldsymbol{g}_{h_i}(x)\right), f\left(\boldsymbol{g}_{h_i}(x')\right) \rangle \quad (2)$$

where NK represents an inner product in a new space whose dimensions are obtained from dimensions of the original latent variable space through the nonlinear mapping f.

III. The nonlinear mappings

Here, we discuss different possible mappings. In particular, in this work we considered the power, the natural logarithm and the logistic functions.

In the first case, the function f of the nonlinear mapping is defined as[2]:

$$f\left(\boldsymbol{g}_{h_i}(x)\right) = \left(\boldsymbol{g}_{h_i}(x)\right)^{\rho} \qquad \forall i = 1, \ldots, N \quad (3)$$

where $\rho > 0$ is a parameter. We notice that for $\rho = 1$ the original kernel is re-obtained. In any case, we assume $\rho \leq 1$, since this solution has some appealing characteristics as we will see in the following. The powering operation is not new in the kernel scenario, even if our use is innovative. The most famous example is the polynomial kernel where the powering of the inner product is considered – while here we propose the powering of each single component of the vectors involved in the inner products. Further, another example can be found in [2], where there is the definition of the so-called Probability Product Kernel – which implies the powering of probability products. Also in this case, there is a remarkable difference, since the definition in [2] lies on powering the components of an integral over the observation space of two known probability distributions, whereas our approach considers the integration (summation) in the latent variable space, namely the integration is over the model components.

[2]In all these functions, if $\boldsymbol{g}_{h_i}$ is a vector, we consider the element-wise application of the function.

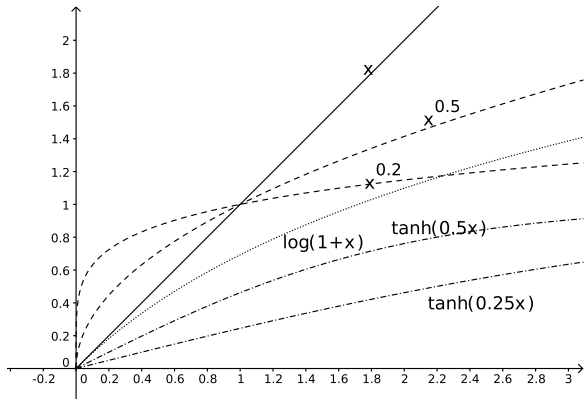

Figure 1. Behaviors of the investigated nonlinear mappings.

In the second function (natural logarithm), f is defined as:

$$f\left(\boldsymbol{g}_{h_i}(x)\right) = \log\left(1 + \boldsymbol{g}_{h_i}(x)\right) \quad (4)$$

This represents an interesting function, since it has no parameters to be set. With respect to the power operation, it has the peculiarity of not raising lower components, still reducing large ones.

Finally, in the logistic functions case, the function f is defined as:

$$f(\boldsymbol{g}_{h_i}(x)) = \tanh\left(\frac{\rho}{2}\boldsymbol{g}_{h_i}(x)\right) = \frac{1 - e^{-\rho \boldsymbol{g}_{h_i}(x)}}{1 + e^{-\rho \boldsymbol{g}_{h_i}(x)}} \quad (5)$$

with $0 < \rho < 2$.

Let us comment on the different behaviors of these functions, which are shown in Fig. 1, together with the identity map $f(x) = x$ (corresponding to the original, unmodified methods). In such figure, the power function (slashed graphs) is represented for $\rho = 0.2$ and $\rho = 0.5$, the logistic function (slash-dotted graphs) for $\rho = 0.5$ and $\rho = 1$, the logarithmic function has no parameter (dotted graph).

A common feature of these transformations is that they are concave, with vanishing derivatives at $+\infty$. Actually, it

seems that these two characteristics are crucial in the choice of nonlinear mappings: in our experimental study we found that convex functions do not produce any improvement of the kernel (for example, we tried the powering operation with $\rho > 1$).

Moreover, all the considered nonlinear mappings are asymptotically *nonexpansive*: they *reduce distances* between $\boldsymbol{g}_{h_i}(x)$ and $\boldsymbol{g}_{h_i}(y)$, provided that these quantities are large enough. More specifically, the logarithm and the logistic functions (for $0 < \rho < 2$) are *globally* nonexpansive, while the power ($0 \leq \rho < 1$) is nonexpansive for $\boldsymbol{g}_{h_i}(x), \boldsymbol{g}_{h_i}(y) > 1$, but, on the contrary, it expands distances for $\boldsymbol{g}_{h_i}(x), \boldsymbol{g}_{h_i}(y) < 1$. Furthermore, the power transformation $f(x) = x^\rho$ magnifies small values of x (for $x < 1$), while shrinking large values of the variable. This effect becomes stronger and stronger as ρ approaches 0.

Coming back to our kernel case, the effect of a powering operation with $\rho \leq 1$ is to raise the contribution of smaller components of $\boldsymbol{g}_{h_i}$ and to reduce the contribution of larger components of $\boldsymbol{g}_{h_i}$, thus re-equilibrating the contributions of each latent variable. This may be seen as a way of augmenting the entropy of the contributions of latent variables. On the contrary, assuming $\rho > 1$ has the opposite behavior, sparsifying the contributions of the latent variables. The natural logarithm $f(x) = \log(1 + x)$ performs a more energetic shrinking of large values of x, but it is essentially the identity map for values of x near 0. Finally, the logistic map $f(x) = \tanh(\frac{\rho}{2}x)$ shrinks the range of x to the interval $0 \leq x < 1$: in some sense, it behaves in a way which is even more radical than the logarithm. The parameter ρ controls the initial slope of the curve ($\rho = 2f'(0)$).

A final remark: clearly, the choice of the parameters in the logistics and in the power functions is crucial. Different values may lead to different behaviors of the kernel functions. In this study we do not propose any method for the choice of these values, considering them as free parameters of the kernels. A further study on the effect of these values is currently under investigation, but, from a preliminary experimental evaluation, there is evidence of a correlation between some invariants of the Gram matrix and the accuracies, when varying the parameter ρ.

IV. Experimental evaluation and discussion

The proposed methods have been evaluated in three sequence classification problems, using fully ergodic HMMs as generative models, Marginalized Kernel [4] and State-Space Kernel [6] as generative kernels, and the three nonlinear mappings described in the previous section. The exact formulation of the quantity $\boldsymbol{g}_{h_i}(x)$ for HMMs and these kernels may be found in [8][3].

[3]In such paper, two possible applications of the nonlinear mapping idea are proposed. In our experiments, we tested both versions reporting only the best results among the two.

To test the proposed approach we performed three experiments, in two application domains, comparing the original version of the kernel to its enhanced counterpart obtained by nonlinear mappings. The first application domain is 2-D shape classification, where we chose to study the Chicken Pieces Database, denoted also as *Chicken* data [9] (446 contours of chicken pieces – with five classes). We employed two different sequence representations to model contours, chain codes and curvature angles. In the first case, a standard 8-direction chain encoding procedure is applied to each image. Then, discrete HMMs are used to model these classes of symbol sequences. In the second case, we derive curvature sequences as in [10], [11]. Classes of curvature sequences are modeled by continuous Gaussian HMMs. More details on how to employ HMM to recognize 2-D shapes may be found in [10]. The original set is split into the training and test sets, in the ratio of $50\% - 50\%$. The classification runs are averaged over 20 hold-out experiments.

The second application concerns a gesture classification problem, where we used high-quality recordings of Australian sign language signs (Auslan) [12]. The problem we considered is composed by 10 signs (classes), with 27 samples per sign; each sample is a sequence of 22-D observations, with an averaged length of 57 frames. Continuous Gaussian HMMs are employed in this case, directly modeling the signals acquired from the sensors. In order to get comparable results to [12], the performance of our classification schemes is computed by using 20 repetitions of a 5-fold cross-validation.

HMMs were trained using Baum-Welch procedure [13]: in the continuous case the training was initialized with a standard GMM clustering, whereas in the discrete case 20 random initializations were tried, picking the best in a likelihood sense. A preliminary evaluation (not shown here due to lack of space) revealed that the best number of states of the HMM was 8 for the chain codes case, 5 for the curvature case, and 3 for the gesture recognition problem.

The parameters, in the power and logistic cases, have been varied in a logarithmic scale between 0 and 1. The best results, for each nonlinear mapping, are shown in Table I, for the three different experiments. From this table, it is evident that the proposed nonlinear mappings have a beneficial impact on the performances of both the State-Space Kernel and the Marginalized Kernel for all data sets. Moreover, in the *Chicken* case, the obtained results are really competitive with the state of the art, considering the difficulty of the data set (in [14] the authors reported a 83% of accuracy, which is, to the best of our knowledge, the best result on this dataset). Concerning the three mappings, it is evident that the powering function is more beneficial than logistic and logarithm: probably, its behavior for smaller components (raising them) is beneficial and crucial.

Table I
RESULTS. SSK IS THE STATE-SPACE KERNEL [6], AND MK IS THE MARGINALIZED KERNEL [4]. NLM STANDS FOR NONLINEAR MAPPING (BETWEEN BRACKETS THE APPLIED FUNCTION). THE STANDARD ERRORS OF THE MEAN, FOR THE BEST PARAMETER CHOICES, ARE ALL LOWER THAN 0.007.

Embedding	Chicken Chain Codes	Chicken Curvature	Auslan
Original SSK	0.751	0.736	0.798
NLM SSK (power)	0.813	0.807	0.904
NLM SSK (logar)	0.753	0.755	0.838
NLM SSK (logis)	0.770	0.780	0.826
Original MK	0.775	0.767	0.533
NLM MK (power)	0.855	0.780	0.932
NLM MK (logar)	0.829	0.776	0.901
NLM MK (logis)	0.817	0.776	0.856

V. CONCLUSIONS

In this paper we investigated the suitability of three nonlinear mappings for preprocessing the space underlying two HMM-based generative kernels. Obtained results on three different experiments confirm the applicability of the proposed approach.

ACKNOWLEDGMENT

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

REFERENCES

[1] T. S. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," in *Advances in Neural Information Processing Systems*, vol. 11, 1999, pp. 487–493.

[2] T. Jebara, I. Kondor, and A. Howard, "Probability product kernels," *Journal of Machine Learning Research*, vol. 5, pp. 819–844, 2004.

[3] M. Cuturi, K. Fukumizu, and J.-P. Vert, "Semigroup kernels on measures," *Journal of Machine Learning Research*, vol. 6, pp. 1169–1198, 2005.

[4] K. Tsuda, T. Kin, and K. Asai, "Marginalized kernels for biological sequences," *Bioinformatics*, vol. 18, pp. S268–S275, 2002.

[5] A. Ng and M. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes," in *Advances in Neural Information Processing Systems*, 2002.

[6] M. Bicego, E. Pekalska, D. Tax, and R. Duin, "Component–based discriminative classification for Hidden Markov Models," *Pattern Recognition*, vol. 42, no. 11, pp. 2637–2648, 2009.

[7] N. Smith and M. Gales, "Speech recognition using SVMs," in *Advances in Neural Information Processing Systems*, vol. 14, 2002, pp. 1197–1204.

[8] A. Carli, M. Bicego, S. Baldo, and V. Murino, "Non-linear generative embeddings for kernels on latent variable models," in *Proc. IEEE ICCV09 Workshop on Subspace Methods*, 2009, pp. 154–161.

[9] G. Andreu, A. Crespo, and J. Valiente, "Selecting the toroidal self-organizing feature maps (TSOFM) best organized to object recognition," in *Proc. IEEE International Conference on Neural Networks*, vol. 2, 1997, pp. 1341–1346.

[10] M. Bicego and V. Murino, "Investigating Hidden Markov Models' capabilities in 2D shape classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 281–286, 2004.

[11] M. Neuhaus and H. Bunke, "Edit distance-based kernel functions for structural pattern classification," *Pattern Recognition*, vol. 39, no. 10, pp. 1852–1863, 2006.

[12] M. Kadous, "Learning comprehensible descriptions of multivariate time series," in *Proc. International Conference on Machine Learning*, 1999, pp. 454–463.

[13] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[14] A. Perina, M. Cristani, U. Castellani, V. Murino, and N.Jojic, "A hybrid generative/discriminative classification framework based on free-energy terms," in *Proc. IEEE International Conference on Computer Vision*, 2009.