Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence

A Tensor-Variate Gaussian Process for Classification of Multidimensional Structured Data

Qibin Zhao
RIKEN Brain Science Institute
Wakoshi, Saitama, Japan
qbzhao@brain.riken.jp

Liqing Zhang
Shanghai Jiao Tong University
Shanghai, China
zhang-lq@cs.sjtu.edu.cn

Andrzej Cichocki
RIKEN Brain Science Institute
Wakoshi, Saitama, Japan
cia@brain.riken.jp

Abstract

As tensors provide a natural and efficient representation of multidimensional structured data, in this paper, we consider probabilistic multinomial probit classification for tensor-variate inputs with Gaussian processes (GP) priors placed over the latent function. In order to take into account the underlying multimodes structure information within the model, we propose a framework of probabilistic product kernels for tensorial data based on a generative model assumption. More specifically, it can be interpreted as mapping tensors to probability density function space and measuring similarity by an information divergence. Since tensor kernels enable us to model input tensor observations, the proposed tensor-variate GP is considered as both a generative and discriminative model. Furthermore, a fully variational Bayesian treatment for multiclass GP classification with multinomial probit likelihood is employed to estimate the hyperparameters and infer the predictive distributions. Simulation results on both synthetic data and a real world application of human action recognition in videos demonstrate the effectiveness and advantages of the proposed approach for classification of multiway tensor data, especially in the case that the underlying structure information among multimodes is discriminative for the classification task.

Introduction

Tensors (also called multiway arrays) are generalization of vectors and matrices to higher dimensions and are equipped with corresponding multilinear algebra. Development of theory and algorithms for tensor decompositions (factorizations) has been an active area of study within the past decade, see e.g. (Cichocki et al. 2009; Kolda and Bader 2009), and the methods have been successfully applied to problems in unsupervised learning and exploratory data analysis. Multiway analysis enables us to effectively capture the multilinear structure of the data, which is usually available as *a priori* information on the data nature. There is a growing need for the development and application of machine learning methods to analyze multidimensional data, such as functional magnetic resonance (fMRI), electrocorticography (ECoG), electroencephalography (EEG) data, and 3D video sequences, thus emphasizing the need to take the information on the structure of the original data into account. Tensors provide a natural and efficient way to describe such multidimensional structured data, and the corresponding learning methods can explicitly exploit the *a priori* information of data structure and capture the underlying multimode relations to achieve useful decompositions of the data with good generalization ability. Recent research has addressed extensions of the kernel concept into tensor decompositions (Signoretto, De Lathauwer, and Suykens 2011; Xu, Yan, and Qi 2012), aiming to bring together the desirable properties of kernel methods and tensor decompositions for significant performance gain when the data are structured and nonlinear dependencies among latent variables do exist. In (Xu, Yan, and Qi 2012), the nonlinear tensor decomposition problem is addressed by a *Kronecker* product of kernels that are obtained from different groups of vector inputs. In (Signoretto, De Lathauwer, and Suykens 2011), the Chordal distance-based kernel for tensorial data is introduced with rotation and reflection invariance on the Grassmann manifold.

Gaussian process (GP) (Rasmussen and Williams 2006; Kersting and Xu 2009) is attractive for non-parametric probabilistic inference because knowledge can be specified directly in the prior distribution of latent function through the mean and covariance function. Inference can be achieved in a closed form for regression under a Gaussian likelihood, but approximation is necessary under non-Gaussian likelihoods. Gaussian process can be extended to binary classification problems by employing logistic or probit likelihoods (Nickisch and Rasmussen 2008), while multinomial logistic or multinomial probit likelihoods are employed in multiclass Gaussian process classification (Williams and Barber 1998; Chai 2012; Girolami and Rogers 2006). Since exact inference is analytically intractable for logistic and probit likelihoods, approximation inference is widely applied, such as Laplace approximation (Williams and Barber 1998), expectation propagation (Kim and Ghahramani 2006; Riihimäki, Jylänki, and Vehtari 2012) and variational approximation (Girolami and Rogers 2006).

In this paper, we extend multiclass GP classification to a tensor variate input space in order to consider the multiway structure of inputs into the model learning and predictions, which is important and promising for multidimen-

sional structured data classification. To this end, a new GP prior for the latent function is necessary, which can explicitly encode tensor structure information into the covariance function. Therefore, we propose a new family of multimode product kernels for tensorial data based on probabilistic generative models and information divergences. Unlike the *Kronecker* product kernel in (Xu, Yan, and Qi 2012; Saatci 2011), our tensor kernel is defined on tensor inputs and the objective is to make classification for tensor observations. The multinomial probit likelihood with variational Bayesian approximation are then employed for a multiclass tensor variate GP classification framework. In contrast with the standard GP, our proposed tensor-based GP (Tensor-GP) enables us to model both the input data by multiple generative models and the corresponding outputs by a probit likelihood model, which is promising to bring together the advantages of generative and discriminative models. In addition, Tensor-GP has several advantages over classical tensor-based methods, such as handling tensors with missing values, inference of hyperparameters, and providing uncertainty of predictions. Both simulations on synthetic data and a real-world application of video classification demonstrate the effectiveness and advantages of Tensor-GP, especially in the case that multiway structure is informative and discriminative for a specific classification problem.

Multilinear Algebra

For the development to follow, we first introduce the notation adopted in this paper. Tensors are denoted by calligraphic letters, e.g., $\mathcal{X}$; matrices by boldface capital letters, e.g., $\mathbf{X}$; and vectors by boldface lowercase letters, e.g., $\mathbf{x}$. The *order* of a tensor is the number of dimensions, also knows as ways or modes. The element $(i_1, i_2, \ldots, i_N)$ of an Nth-order tensor $\mathcal{X}$ is denoted by $x_{i_1 i_2 \ldots i_N}$ or $(\mathcal{X})_{i_1 i_2 \ldots i_N}$, in which indices typically range from 1 to their capital version, e.g., $i_n = 1, \ldots, I_n$. *Matricization*, also known as *unfolding*, is the process of reordering the elements of a tensor into a matrix. More specifically, the mode-n matricization of a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is denoted by $\mathbf{X}_{(n)} \in \mathbb{R}^{I_n \times I_1 \cdots I_{n-1} I_{n+1} \cdots I_N}$, while the vectorization of a tensor is denoted as $\text{vec}(\mathcal{X})$. The *inner product* of two same-sized tensors $\mathcal{X}, \mathcal{X}' \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ is defined by $\langle \mathcal{X}, \mathcal{X}' \rangle = \sum_{i_1 i_2 \ldots i_N} x_{i_1 i_2 \ldots i_N} x'_{i_1 i_2 \ldots i_N}$, and the squared Frobenius norm by $\|\mathcal{X}\|_F^2 = \langle \mathcal{X}, \mathcal{X} \rangle$.

The two most commonly used decompositions are the *Tucker* model and *CANDECOMP/PARAFAC* (CP) model, both of which can be regarded as higher-order generalizations of the matrix singular value decomposition (SVD). Let $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ denote an Nth-order tensor, then *Tucker* model is defined as follows:

$$\mathcal{X} = \mathcal{G} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \cdots \times_N \mathbf{U}^{(N)} \tag{1}$$

where $\mathcal{G} \in \mathbb{R}^{R_1 \times \cdots \times R_N}$ denotes the *core tensor* and $\mathbf{U}^{(n)} \in \mathbb{R}^{I_n \times R_n}$ denotes the mode-n *factor matrix*. If all factor matrices $\{\mathbf{U}^{(n)}\}_{n=1}^N$ are columnwise orthonormal and the core tensor $\mathcal{G}$ is all-orthogonal (i.e., any subtensors are orthogonal) and ordered, this decomposition is called higher-order singular value decomposition (HOSVD). If all the factor matrices have the same number of components, and the core tensor is super-diagonal, Tucker model simplifies to *CP* decomposition (Kolda and Bader 2009; De Sterck 2012) which can be also defined as a sum of rank-one tensors:

$$\mathcal{X} = \sum_{r=1}^{R} \lambda_r \, \mathbf{u}_r^{(1)} \circ \mathbf{u}_r^{(2)} \circ \cdots \circ \mathbf{u}_r^{(N)}, \tag{2}$$

where the symbol '$\circ$' denotes the outer product of vectors and R is defined as tensor *rank* (Kolda and Bader 2009). In general, *CP* model is considered to be a multilinear low-rank approximation while *Tucker* model is regarded to be a multilinear subspace approximation.

Multiclass Gaussian Processes Classification

We consider a classification problem consisting of Mth-order tensors $\mathcal{X}_n \in \mathbb{R}^{I_1 \times \cdots \times I_M}$ associated with target classes $y_n \in \{1, \ldots, C\}$, where $C > 2$, for $n = 1, \ldots, N$. All class labels are collected in the $N \times 1$ target vector $\mathbf{y}$, and all tensors are concatenated in a $(M+1)$th-order tensor $\boldsymbol{\mathcal{X}}$ of size $N \times I_1 \times \cdots \times I_M$. Given the latent function $\mathbf{f}_n = [f_n^1, f_n^2, \ldots, f_n^C]^T = \mathbf{f}(\mathcal{X}_n)$ at the observed input location $\mathcal{X}_n$, the class labels y_n are assumed independently and identically distributed as defined by a multinomial probit likelihood model $p(y_n|\mathbf{f})$. The latent vectors from all observations are denoted by $\mathbf{f} = [f_1^1, \ldots, f_N^1, f_1^2, \ldots, f_N^2, \ldots, f_1^C, \ldots, f_N^C]^T$. Our goal is to predict the class membership for a new input tensor $\mathcal{X}_*$ given the observed data $\mathcal{D} = \{\boldsymbol{\mathcal{X}}, \mathbf{y}\}$. We place Gaussian process priors on the latent function related to each class, which is the common assumption in multiclass GP classification (see (Rasmussen and Williams 2006; Riihimäki, Jylänki, and Vehtari 2012)). This specification results in the following zero-mean Gaussian prior for $\mathbf{f}$:

$$p(\mathbf{f}|\boldsymbol{\mathcal{X}}) = \mathcal{N}(\mathbf{0}, \mathbf{K}), \tag{3}$$

where $\mathbf{K}$ is a $CN \times CN$ blocked diagonal covariance matrix with matrices $\mathbf{K}^1, \ldots, \mathbf{K}^C$ (of size $N \times N$) on its diagonal corresponding to each class respectively. Element $K_{i,j}^c$ in cth class covariance matrix defines the prior covariance between f_i^c and f_j^c, which is governed by a kernel function $k(\mathcal{X}_i, \mathcal{X}_j)$, i.e., $K_{i,j}^c = k(\mathcal{X}_i, \mathcal{X}_j) = \text{Cov}(f_i^c, f_j^c)$ within the class c. Note that the kernel function should be defined in tensor-variate input space, hence commonly used kernel functions, such as Gaussian RBF, are infeasible. Therefore, a new framework of probabilistic product kernel for tensors are introduced and discussed in the next Section. In kernel function, hyperparameters are defined to control the smoothness properties and overall variance of latent functions, which usually are collected into one vector $\boldsymbol{\theta}$. For simplicity, we use the same $\boldsymbol{\theta}$ for all classes. For likelihood model, we consider the multinomial probit, which is a generalization of the probit model, given as

$$p(y_n|\mathbf{f}_n) = \text{E}_{p(u_n)} \left\{ \prod_{c=1, c \neq y_n}^{C} \Phi(u_n + f_n^{y_n} - f_n^c) \right\}, \tag{4}$$

where Φ denotes the cumulative density function of the standard normal distribution, and the auxiliary variable u_n is distributed as $p(u_n) = \mathcal{N}(0, 1)$.

By applying Bayes' theorem, the posterior distribution of the latent function is given by

$$p(\mathbf{f}|\mathcal{D},\boldsymbol{\theta}) = \frac{1}{Z}p(\mathbf{f}|\boldsymbol{\mathcal{X}},\boldsymbol{\theta})\prod_{n=1}^{N} p(y_n|\mathbf{f}_n), \tag{5}$$

where $Z = \int p(\mathbf{f}|\boldsymbol{\mathcal{X}},\boldsymbol{\theta})\prod_{n=1}^{N} p(y_n|\mathbf{f}_n)d\mathbf{f}$ is known as the marginal likelihood. Inference for a test input $\mathcal{X}_*$ is performed in two steps. First the posterior distribution of latent function $\mathbf{f}_*$ is given as $p(\mathbf{f}_*|\mathcal{D},\mathcal{X}_*,\boldsymbol{\theta}) = \int p(\mathbf{f}_*|\mathbf{f},\mathcal{X}_*,\boldsymbol{\theta})p(\mathbf{f}|\mathcal{D},\boldsymbol{\theta})d\mathbf{f}$. Then we compute the posterior predictive probability of $\mathcal{X}_*$, which is given by $p(y_*|\mathcal{D},\mathcal{X}_*,\boldsymbol{\theta}) = \int p(y_*|\mathbf{f}_*)p(\mathbf{f}_*|\mathcal{D},\mathcal{X}_*,\boldsymbol{\theta})d\mathbf{f}_*$. Since non-Gaussian likelihood model results in an analytically intractable posterior distribution, thus variational approximative methods can be used for approximative inference.

Probabilistic Product Kernels for Tensors

The kernels are considered by defining a topology implying the *a priori* knowledge about invariance in the input space. Although many kernels have been designed for a number of structured objects, few approaches exploit the structure of tensorial representations. In this section, we discuss the kernels for tensor-variate inputs, which can take multiway structure into account for similarity measures.

There are some valid reproducing kernels admit a straightforward generalization to Mth-order tensors, such as the kernel functions $k : \mathcal{X}\times\mathcal{X}\rightarrow\mathbb{R}$ given as

$$\text{Linear kernel:}\quad k(\mathcal{X},\mathcal{X}') = \langle \text{vec}(\mathcal{X}), \text{vec}(\mathcal{X}')\rangle,$$

$$\text{Gaussian-RBF:}\quad k(\mathcal{X},\mathcal{X}') = \exp\left(-\frac{1}{2\beta^2}\|\mathcal{X}-\mathcal{X}'\|_F^2\right).$$

In order to define the similarity measure that directly exploits multilinear algebraic structure of input tensors, a product kernel based on Chordal distance (projection Frobenius norm) on Grassmannian manifolds was proposed (Signoretto, De Lathauwer, and Suykens 2011).

Probabilistic kernels in vector input space have been investigated based on generative models and information divergences, such as *Fisher* kernel (Tsuda et al. 2004) and *Kullback-Leibler* kernel (Moreno, Ho, and Vasconcelos 2003). The Fisher kernel assumes a generative model that well explains all data samples and maps each sample into a gradient log-likelihood parameter space. Here, we propose a new probabilistic kernel framework for multiway tensors based on the assumption that each Mth-order tensor observation (e.g., $\mathcal{X}_n \in \mathbb{R}^{I_1\times\cdots\times I_M}$) is considered individually as M different generative models. More specifically, mode-m matricization $\mathbf{X}_{n_{(m)}}$ is regarded as an ensemble of multivariate instances with dimensionality of I_m and number of instances of $I_1I_2\cdots I_{m-1}I_{m+1}\cdots I_M$, generated from a parametric model $p(\mathbf{x}|\boldsymbol{\lambda}_m^n)$. In this manner, $\mathcal{X}$ has been successfully mapped into M-dimensional model-based probability distribution function space, i.e., $\{p(\mathbf{x}|\boldsymbol{\lambda}_m^n)|m = 1,\ldots,M\}$. Subsequently, similarity measure between two tensors $\mathcal{X}$ and $\mathcal{X}'$ in mode-m is defined as

$$S_m(\mathcal{X}||\mathcal{X}') = D\left(p(\mathbf{x}|\boldsymbol{\lambda}_m^{\mathcal{X}})\|q(\mathbf{x}|\boldsymbol{\lambda}_m^{\mathcal{X}'})\right), \tag{6}$$

where p, q represent mode-m probability density function for $\mathcal{X}$ and $\mathcal{X}'$ respectively and $D(p||q)$ is an information divergence between two distributions. One popular information divergence is the *symmetric Kullback-Leibler* (sKL) divergence (Moreno, Ho, and Vasconcelos 2003) expressed as

$$D_{sKL}\left(p(\mathbf{x}|\boldsymbol{\lambda})||q(\mathbf{x}|\boldsymbol{\lambda}')\right) = \frac{1}{2}\int_{-\infty}^{+\infty} p(\mathbf{x}|\boldsymbol{\lambda})\log\frac{p(\mathbf{x}|\boldsymbol{\lambda})}{q(\mathbf{x}|\boldsymbol{\lambda}')}d\mathbf{x} + \frac{1}{2}\int_{-\infty}^{+\infty} q(\mathbf{x}|\boldsymbol{\lambda}')\log\frac{q(\mathbf{x}|\boldsymbol{\lambda}')}{p(\mathbf{x}|\boldsymbol{\lambda})}d\mathbf{x}. \tag{7}$$

Another possibility is the *Jensen-Shannon* (JS) divergence (Chan, Vasconcelos, and Moreno 2004; Endres and Schindelin 2003) expressed by

$$D_{JS}(p||q) = \frac{1}{2}\text{KL}(p||r) + \frac{1}{2}\text{KL}(q||r), \tag{8}$$

where $\text{KL}(\cdot||\cdot)$ denotes *Kullback-Leibler* (KL) divergence and $r(\mathbf{x}) = \frac{1}{2}(p(\mathbf{x}) + q(\mathbf{x}))$ represents a mixture distribution. The JS divergence can be interpreted as the average KL divergence between each probability distribution and the average distribution, or equivalently as the diversity of two distributions with equal priors. Finally, a probabilistic kernel for tensors is defined as a product of mode-m factor kernels, which is given by

$$k(\mathcal{X},\mathcal{X}') = \alpha^2\prod_{m=1}^{M}\exp\left(-\frac{1}{2\beta_m^2}S_m(\mathcal{X}||\mathcal{X}')\right), \tag{9}$$

where α denotes a magnitude parameter and $[\beta_1,\ldots,\beta_M]$ play the role of characteristic length-scales which implement automatic relevance determination (ARD) (Rasmussen and Williams 2006). All kernel parameters are usually denoted by $\boldsymbol{\theta} = \{\alpha,\beta_m|m = 1,\ldots,M\}$. It can be shown that both sKL and JS divergences are non-negative and equal to zero when $p(\mathbf{x}) = q(\mathbf{x})$, while they do not fulfill the triangle inequality (i.e., we do not have $D(p||q) \leq D(p||r)+D(r||q)$). However, it has been proven in (Endres and Schindelin 2003; Chan, Vasconcelos, and Moreno 2004) that $[D_{sKL}(p||q)]^{\frac{1}{2}}$ and $[D_{JS}(p||q)]^{\frac{1}{2}}$ fulfill the triangle inequality thus is a metric, implying that *tensor kernel* defined in (9) is a metric kernel.

For simplicity, Gaussian model assumption is employed with model parameters including a mean vector and a full covariance matrix, i.e., $\boldsymbol{\lambda}_m = \{\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m\}$ that can be estimated by maximum likelihood from $\mathbf{X}_{(m)}$. The detailed algorithms of sKL and JS between two multivariate Gaussian are given in (Moreno, Ho, and Vasconcelos 2003; Abou-Moustafa and Ferrie 2012). In practice, because of the absence of closed-form solutions for probabilistic kernels, one may end up with a kernel matrix that is not positive-definite due to inaccuracies in the approximations.

The tensor kernels described here have some interesting properties. An intuitive interpretation for the operation performed is that Mth-order tensor observations are first mapped into M-dimensional probability density function space, then information divergence is applied as a similarity measure. Hence, such a kernel combines generative models

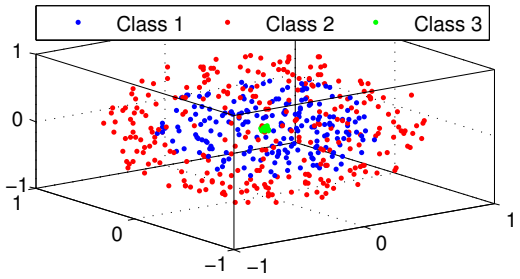

Figure 1: Three classes data points are shown in their first three dimensional space.

with discriminative ones when used in conjunction with a specific discriminative method such as the GP classifier. The probabilistic tensor kernel can handle tensors with missing values or different sizes. Furthermore, the number of kernel parameters in (9) is much smaller than that of the RBF kernel performed on unfolded tensors, implying that the tensor kernel is less prone to overfitting.

Variational Bayesian Inference

Based on the probabilistic tensor kernel described above, GP prior for cth latent function $f^c(\mathcal{X})$ can be rewritten as

$$f^c(\mathcal{X})|\boldsymbol{\mathcal{X}}, \boldsymbol{\theta}^c \sim \mathcal{GP}(0, k(\mathcal{X}, \mathcal{X}'|\boldsymbol{\theta}^c)), \qquad (10)$$

where the hyperparameters $\{\boldsymbol{\theta}^c\}_{c=1}^C$ are set to the same for all classes and the kernel parameter α is set to a constant, i.e., $\boldsymbol{\theta} = [\beta_1 \dots \beta_M]^T$ which consists of length-scales parameters. An hierarchic hyperprior is placed over $\boldsymbol{\theta}$ such that each hyperparameter has, for example, an independent exponential distribution given by $\varphi_m \sim \text{Exp}(\psi_m)$ where $\varphi_m = 1/2\beta_m^2$, and a Gamma prior is placed on the mean values of the exponential $\psi_m \sim \Gamma(\sigma, \tau)$ thus forming a conjugate pair. The associated hyperparameters σ, τ are simply set to 10^{-6}.

We now consider the variational Bayesian approximation for GP classification. The posterior $p(\mathbf{f}|\mathcal{D})$ is approximated by the variational posterior $q(\mathbf{f}|\mathcal{D})$ by minimizing the Kullback-Leibler divergence

$$\text{KL}(q(\mathbf{f}|\mathcal{D})||p(\mathbf{f}|\mathcal{D})) = \int q(\mathbf{f}|\mathcal{D}) \log \frac{q(\mathbf{f}|\mathcal{D})}{p(\mathbf{f}|\mathcal{D})} d\mathbf{f}. \qquad (11)$$

This is the difference between the log marginal likelihood $\log p(\mathbf{y})$ and a variational lower bound. Finally, the posterior for latent function is approximated by $q(\mathbf{f}) = \prod_{n=1}^N q(\mathbf{f}_n) = \prod_{n=1}^N \mathcal{N}_{f_n}^{y_n}(\widetilde{\mathbf{m}}_n, \mathbf{I})$. If we also consider the set of hyperparameters in this variational treatment then the approximate posterior for the covariance kernel hyperparameters takes the form of

$$q(\varphi) \propto \mathcal{N}(\mathbf{0}, \mathbf{K}_\varphi) \prod_{m=1}^M \text{Exp}(\varphi_m|\widetilde{\psi}_m), \qquad (12)$$

and the required posterior expectations can be estimated by employing importance sampling. The hyperparameters in hyperprior is approximated by $q(\psi_m) = \Gamma(\sigma + 1, \tau + \widetilde{\varphi}_m)$

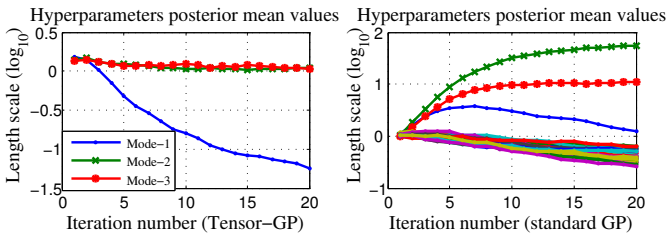

Figure 2: Evolution of estimated posterior means for the inverse squared length scale hyperparameters.

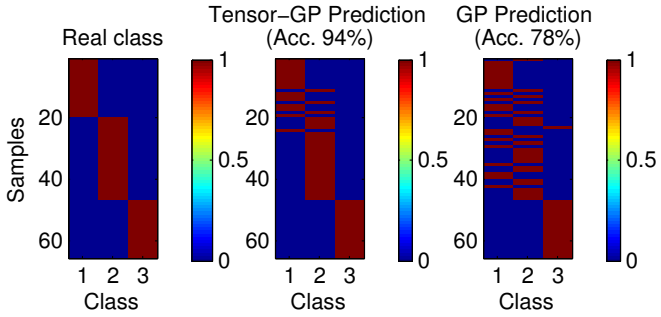

Figure 3: Performance comparison between Tensor-GP and the GP classifier with relatively small number of samples for model learning.

and the associated posterior mean is simply obtained by $\widetilde{\psi}_m = (\sigma + 1)/(\tau + \widetilde{\varphi}_m)$. The detailed procedure for approximation inference can be found in (Girolami and Rogers 2006).

Experimental Results

Illustrative simulations on synthetic data

In order to investigate the properties and advantages of tensor-based GP classification approach, two experiments have been performed on synthetic datasets under conditions with respect to data structure and number of obervations.

In the first simulation, 27-dimensional data vectors $\mathbf{x}$ were generated such that if $y = 1$ then $0.5 > x_1^2 + x_2^2 + x_3^2 > 0.1$, for $y = 2$ then $1.0 > x_1^2 + x_2^2 + x_3^2 > 0.6$ and for $y = 3$ then $[x_1, x_2, x_3]^T \sim \mathcal{N}(\mathbf{0}, 0.01\mathbf{I})$. The remaining dimensions $[x_3, \dots, x_{27}]^T$ are all distributed as $\mathcal{N}(0, 0.1)$. Hence, the first three dimensions are discriminative for the classification, as shown in Fig. 1, while the remaining 24 dimensions are irrelevant to the classification task. The three target values were sampled uniformly thus creating a balance of samples drawn from the three target classes. For comparison, the GP multiclass classifier (Girolami and Rogers 2006) was performed on the dataset $\{\mathbf{x}_n\} \in \mathbb{R}^{27}$ with labels $\{y_n\}$, while the proposed Tensor-GP was performed on the same dataset represented by tensors $\{\mathcal{X}_n\} \in \mathbb{R}^{3\times3\times3}$ with labels $\{y_n\}$ where $\mathcal{X}_n$ is obtained by tensorization of $\mathbf{x}_n$. One hundred samples drawn from the above distribution were used in the variational inference routine with a further 70 points being used for performance evaluation.

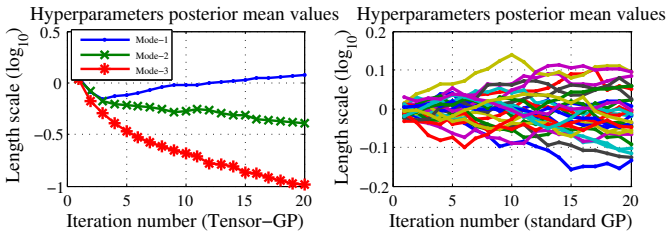

Figure 4: Evolution of estimated posterior means for the inverse squared length scale hyperparameters on a dataset generated by CP model.

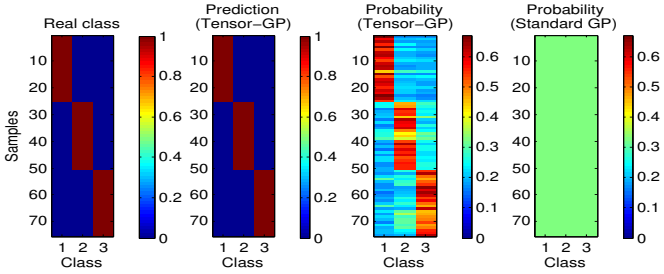

Figure 5: Performance comparison between Tensor-GP and the GP classifier on structured tensors generated by a multilinear CP model.

A common radial basis covariance function of the form $k_{ij} = \exp(\sum_d |x_{id} - x_{jd}|^2/2\beta_d^2)$ was employed by standard GP and the covariance function in (9) was employed by Tensor-GP. A hyperprior is placed on $\varphi = 1/2\beta^2$ for inference of hyperparameters in covariance functions. The variational iterations ran for twenty steps where the estimated posterior mean values for the covariance function parameters $\tilde{\varphi}_d$ shows automatic relevance detection (ARD) in progress. From Fig. 2, as would be expected that the 24 irrelevant features are effectively removed from the model in standard GP. For Tensor-GP, the mode-1 factor kernel is removed from the model, which depends on how significant the first three features affect the distributions in each mode. The predictive performance are compared in Fig. 3. Observe that on this dataset with a relative small number of samples for model learning, the proposed Tensor-GP classifier achieves a predictive performance of 94%, while the performance of standard GP is only 78%. This result illustrates that although multiway structures of original data are not discriminative for classification task, Tensor-GP still works well and outperforms standard GP in case of small sample sizes.

Since Tensor-GP model is assumed to be more suitable for multidimensional structured data which originally contains a multiway structure and the information carried by interaction among different modes are discriminative for the classification task. In the second simulation, multiway tensor data were generated according to the CP model, defined in (2), with $\{\mathbf{u}_r^{(m)}, r = 1, \ldots, R, m = 1, \ldots, M\}$ are drawn from $\mathcal{N}(\mathbf{0}, \sigma_m^2\mathbf{I})$ and $\{\lambda_r\}_{r=1}^R \sim \mathcal{N}(0, 1)$. The rank of tensors is set to $R = 20$, the order of tensors is set to $M = 3$ and the size in each mode is set to 3, thus $\mathbf{U}^{(m)} \in \mathbb{R}^{3\times 20}$ and $\mathcal{X} \in \mathbb{R}^{3\times 3\times 3}$. Three classes data $\{\boldsymbol{\mathcal{X}}, \mathbf{y}\}$ were generated such that if $y = 1$ then $\sigma = [1, 1, 1]$, for $y = 2$ then $\sigma = [1, 1.01, 1]$ and for $y = 3$ then $\sigma = [1.02, 1, 1]$. Thus, the distribution of the first two modes are discriminative while the third mode is irrelevant to the classification task. For comparison, the standard GP was also applied on dataset $\{\mathbf{X}, \mathbf{y}\}$ where $\mathbf{X}$ consists of vectorization of each tensor data point by $\mathbf{x}_n = \text{vec}(\mathcal{X}_n)$. As can be seen from Fig. 4, after 20 variational iterations, the posterior mean values of length-scale hyperparameters are well learned in Tensor-GP. As would be expected, mode-3 factor kernel is effectively removed from the model as it contains no discriminative information, and model-1 factor kernel is shown to be the most discriminative, which is consistent with the three classes data having larger discrepancy of distributions in mode-1 than mode-2. For GP classifier, as shown in Fig. 4, there are no features to be removed or enhanced significantly, which is also consistent with the dataset since the most discriminative information does not lie in any specific features but the global distribution of all features. The predictive performance are compared in Fig. 5, which shows real class labels, predictions by Tensor-GP, and predictive probability (or confidence) for each class by two methods. The advantages of proposed method is significant such that Tensor-GP achieves a predictive performance of 100% with high confidence, while GP obtains a predictive performance of 37%, implying that it completely fails to classify this dataset.

These two examples demonstrate that our proposed Tensor-GP is feasible for general classification tasks and shows advantages on large number of features with small number of observations. The most important point is that when data structure are discriminative for a classification task, Tensor-GP enables us to effectively capture the multimode structure of inputs represented by tensors, resulting in an enhanced predictive performance. In addition, the number of hyperparameters in Tensor-GP is relatively small and thus it is less prone to overfitting.

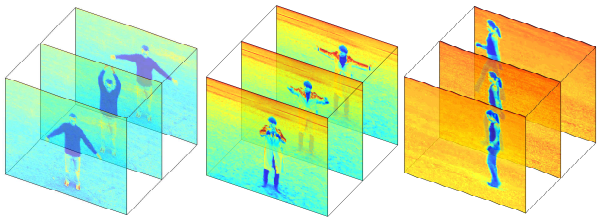

Figure 6: Three examples of video sequences for hand waving, hand clapping and walking actions, which are represented as third-order tensors.

Action classification in videos on KTH dataset

Human action recognition in videos is of high interest for a variety of applications such as video surveillance, human-computer interface and video retrieval, where the most competing methods are based on motion estimation (Ali and Shah 2010), local space-time interest points

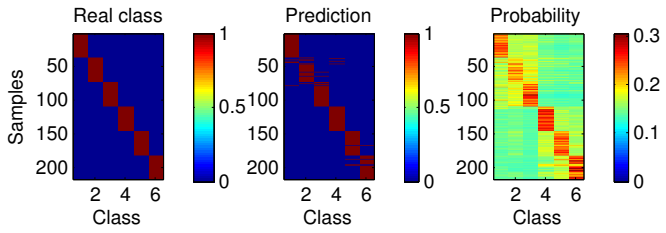

Figure 7: Classification results and probability of predictions on the test set.

and visual code words (Niebles, Wang, and Fei-Fei 2008; Holte et al. 2012), multiple classifiers (Song et al. 2011; Zhang et al. 2011), sparse representation (Guha and Ward 2012) and multiway tensor methods (Kim and Cipolla 2009; Lui, Beveridge, and Kirby 2010). Tensor representation enables us to directly analyze 3D video volume and encode global space-time structure information. To illustrate the advantages of tensor-based Gaussian process, we applied it for video classification on the largest public KTH human action database (Schuldt, Laptev, and Caputo 2004) that contains six types of actions (walking (W), running (R), jogging (J), boxing (B), hand-waving (H-W), and hand-clapping (H-C)) performed by 25 persons in four different scenarios (outdoors, outdoors with scale variation, outdoors with different clothes, and indoors). The total 600 video sequences are divided with respect to the persons into a training set (8 persons), a validation set (8 persons) and a test set (9 persons) according to the standard setting in (Schuldt, Laptev, and Caputo 2004). Each video is space-time aligned and uniformly resized to $20\times20\times32$, which are then be represented by a third-order tensor $\mathcal{X}_n$ (see Fig. 6). Since our model can infer hyperparameters without validation procedure, we simply use the training and validation sets for model learning and make classifications on the test set. The classification results on the test set are shown in Fig. 7 which also shows the probability or confidence of their predictions. Observe that misclassified videos always show less confidence as compared with correctly classified ones. The confusion matrix on test set are shown in Table 1, in which rows correspond to the ground truth, and columns correspond to the classification results. It can be seen that our method achieves average accuracy of 94% and the confusion mainly appears between running and jogging, and between hand clapping and waving, which is consistent with our intuition that these two pairs of actions are easily confused. In addition, the comparisons with the state-of-the-art methods on the KTH dataset are shown in Table 2 and our method achieves similar classification accuracy as TCCA(Kim and Cipolla 2009) and better than WX/SVM(Wong, Kim, and Cipolla 2007), MIL(Ali and Shah 2010), pLSA/LDA(Niebles, Wang, and Fei-Fei 2008), LF/SVM(Schuldt, Laptev, and Caputo 2004). As compared to TCCA, Tensor-GP does not require the precise space-time alignment and is able to naturally infer the hyperparameters. In addition, Tensor-GP can provide the uncertainty of the predictions and can handle tensor observations with missing values. These properties make Tensor-GP more interesting and promising for structured data classification. In summary, both global and local space-time information are discriminative and promising for action recognition and the results demonstrate the effectiveness of the proposed probabilistic tensor kernel on capturing global space-time structures of video volumes and the advantages of the proposed multiclass tensor-based GP for classification.

Table 1: Confusion matrix (average accuracy 94%)

	Walk	Run	Jog	Box	H-C	H-W
Walk	1.0	0	0	0	0	0
Run	.08	.78	.06	.08	0	0
Jog	.03	.03	.94	0	0	0
Box	0	0	0	1.0	0	0
H-C	0	0	0	0	.98	.02
H-W	0	0	0	0	.08	.92

Table 2: Comparisons on the KTH data set.

TCCA	WX/SVM	MIL	pLSA/LDA	LF/SVM
95.33%	91.6%	87.7%	83.33%	71.72%

Conclusions

We propose a multiclass Gaussian process classification framework with tensor-variate inputs, which brings together the advantages of GP model and tensor representation. The main contribution of this work is to introduce a new family of probabilistic kernels for higher-order tensors using information divergences, which can be employed to specify a GP prior for latent functions. Thus a tensor-variate GP classifier based on multinomial probit likelihood and a fully variational Bayesian treatment is developed, which has shown to be promising for classification of multidimensional structured data, especially when data structure is discriminative. The empirical comparisons with variational GP with vector inputs suggest that the proposed probabilistic kernel for tensors is an effective similarity measure with respect to predictive performance. Therefore a new perspective for the development of a range of machine learning methods that admit the underlying multilinear structure is provided. The effectiveness of tensor-based GP classifier is also demonstrated by a real-world application of human action recognition in videos.

Acknowledgments

This work was partially supported by JSPS KAKENHI (Grant No. 24700154), and the National Natural Science Foundation of China (Grant No. 61202155, 90920014, 91120305).

References

Abou-Moustafa, K., and Ferrie, F. 2012. A note on metric properties for some divergence measures: The Gaussian case. *The Journal of Machine Learning Research* 25:1–15.

Ali, S., and Shah, M. 2010. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(2):288–303.

Chai, K. 2012. Variational multinomial logit Gaussian process. *The Journal of Machine Learning Research* 98888:1745–1808.

Chan, A.; Vasconcelos, N.; and Moreno, P. 2004. A family of probabilistic kernels based on information divergence. *Univ. California, San Diego, CA, Tech. Rep. SVCL-TR-2004-1*.

Cichocki, A.; Zdunek, R.; Phan, A. H.; and Amari, S. I. 2009. *Nonnegative Matrix and Tensor Factorizations*. John Wiley & Sons.

De Sterck, H. 2012. A nonlinear GMRES optimization algorithm for canonical tensor decomposition. *SIAM Journal on Scientific Computing* 34(3):1351–1379.

Endres, D., and Schindelin, J. 2003. A new metric for probability distributions. *IEEE Transactions on Information Theory* 49(7):1858–1860.

Girolami, M., and Rogers, S. 2006. Variational Bayesian multinomial probit regression with Gaussian process priors. *Neural Computation* 18(8):1790–1817.

Guha, T., and Ward, R. 2012. Learning sparse representations for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(8):1576–1588.

Holte, M.; Chakraborty, B.; Gonzalez, J.; and Moeslund, T. 2012. A local 3-D motion descriptor for multi-view human action recognition from 4-D spatio-temporal interest points. *IEEE Journal of Selected Topics in Signal Processing* 6(5):553–565.

Kersting, K., and Xu, Z. 2009. Learning preferences with hidden common cause relations. In *Machine Learning and Knowledge Discovery in Databases*. Springer. 676–691.

Kim, T., and Cipolla, R. 2009. Canonical correlation analysis of video volume tensors for action categorization and detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(8):1415–1428.

Kim, H., and Ghahramani, Z. 2006. Bayesian Gaussian process classification with the EM-EP algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28(12):1948–1959.

Kolda, T., and Bader, B. 2009. Tensor decompositions and applications. *SIAM Review* 51(3):455–500.

Lui, Y.; Beveridge, J.; and Kirby, M. 2010. Action classification on product manifolds. In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 833–839. IEEE.

Moreno, P.; Ho, P.; and Vasconcelos, N. 2003. A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. *Advances in Neural Information Processing Systems* 16:1385–1393.

Nickisch, H., and Rasmussen, C. 2008. Approximations for binary Gaussian process classification. *Journal of Machine Learning Research* 9:2035–2078.

Niebles, J.; Wang, H.; and Fei-Fei, L. 2008. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision* 79(3):299–318.

Rasmussen, C., and Williams, C. 2006. *Gaussian processes for machine learning*. The MIT Press, Cambridge, MA, USA.

Riihimäki, J.; Jylänki, P.; and Vehtari, A. 2012. Nested expectation propagation for Gaussian process classification with a multinomial probit likelihood. *arXiv preprint arXiv:1207.3649*.

Saatci, Y. 2011. *Scalable Inference for Structured Gaussian Process Models*. Ph.D. Dissertation, University of Cambridge.

Schuldt, C.; Laptev, I.; and Caputo, B. 2004. Recognizing human actions: A local SVM approach. In *Proceedings of the 17th International Conference on Pattern Recognition, ICPR*, volume 3, 32–36. IEEE.

Signoretto, M.; De Lathauwer, L.; and Suykens, J. A. 2011. A kernel-based framework to tensorial data analysis. *Neural networks* 24(8):861–874.

Song, Y.; Zheng, Y.; Tang, S.; Zhou, X.; Zhang, Y.; Lin, S.; and Chua, T. 2011. Localized multiple kernel learning for realistic human action recognition in videos. *IEEE Transactions on Circuits and Systems for Video Technology* 21(9):1193–1202.

Tsuda, K.; Akaho, S.; Kawanabe, M.; and Müller, K. 2004. Asymptotic properties of the Fisher kernel. *Neural Computation* 16(1):115–137.

Williams, C., and Barber, D. 1998. Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(12):1342–1351.

Wong, S.; Kim, T.; and Cipolla, R. 2007. Learning motion categories using both semantic and structural information. In *Computer Vision and Pattern Recognition. CVPR'07*, 1–6. IEEE.

Xu, Z.; Yan, F.; and Qi, Y. 2012. Infinite Tucker decomposition: Nonparametric Bayesian models for multiway data analysis. In *Proceedings of the 29 th International Conference on Machine Learning*. ACM.

Zhang, T.; Liu, J.; Liu, S.; Xu, C.; and Lu, H. 2011. Boosted exemplar learning for action recognition and annotation. *IEEE Transactions on Circuits and Systems for Video Technology* 21(7):853–866.