Infinite SVM: a Dirichlet Process Mixture of Large-margin Kernel Machines

Jun Zhu[†] JUNZHU@CS.CMU.EDU
Ning Chen[†‡] CHENN07@MAILS.THU.EDU.CN
Eric P. Xing[†] EPXING@CS.CMU.EDU
[†]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA
[‡]Department of Computer Science and Technology, Tsinghua University, Beijing, 100084 China

Abstract

We present *Infinite SVM* (iSVM), a Dirichlet process mixture of large-margin kernel machines for multi-way classification. An iSVM enjoys the advantages of both Bayesian nonparametrics in handling the unknown number of mixing components, and large-margin kernel machines in robustly capturing local nonlinearity of complex data. We develop an efficient variational learning algorithm for posterior inference of iSVM, and we demonstrate the advantages of iSVM over Dirichlet process mixture of generalized linear models and other benchmarks on both synthetic and real Flickr image classification datasets.

1. Introduction

Large-margin methods such as SVMs are among the most widely used approaches to learn classifiers. Their power and popularity stem in part from the ability to handle diverse forms of inputs (e.g., vectors and strings) by using kernel functions (Schölkopf & Smola, 2001). However, learning a large-margin classifier using all the input data can be expensive and may also be limited without considering the underlying (e.g., clustering) structures of complex data. To overcome such limitations, recent progress has been made on developing mixture-of-experts (Collobert et al., 2002; Fu et al., 2010), which split the input space into a number of subregions and learn an SVM classifier within each region.

However, an open problem for finite mixture-of-experts is how to choose the number of experts, which is hard to be specified *a priori*. Classical remedies usually resort to post-processing procedures such as cross-validation or likelihood ratio test (Liu & Shao, 2003). The recent success of nonparametric Bayesian techniques such as the Dirichlet process (DP) mixtures in dealing with similar challenges in clustering (i.e., unknown number of clusters) and density estimation (i.e., unknown number of modalities), offers a promising direction to bypass the model selection problem and automatically resolve the unknown number of experts. Representative applications of DP mixture for prediction (e.g., classification) include the infinite mixture of Gaussian processes (Rasmussen & Ghahramani, 2001) and DP mixture of generalized linear models (GLMs) (Shahbaba & Neal, 2009; Hannah et al., 2010). However, these methods are largely restricted in two aspects. First, although mixture of linear experts can produce a *globally* nonlinear classifier (Shahbaba & Neal, 2009), it would lead to the creation of unnecessary extra experts in order to capture *local* nonlinearities underlying complex data. Second, in order to perform Bayesian inference, a likelihood model (e.g., GLMs) must be defined for response variables, which may involve a hard-to-compute partition function that can substantially complicate posterior inference. To the best of our knowledge, there has been no successful attempt to integrate Bayesian nonparametrics with large-margin kernel machines to obtain an infinite mixture of nonlinear large-margin decision surfaces, potentially more suitable for complex multi-way classification of high dimensional data.

In this paper, we present *Infinite SVM* (iSVM), a Dirichlet process mixture of large-margin kernel machines that conjoins the ideas of large-margin learning, kernel machine, and Bayesian nonparametrics. iSVM models a mixture of classifiers in a similar way as DP mixture models a mixture of density components; and employs the maximum entropy discrimination (MED) (Jaakkola et al., 1999) framework to integrate the large-margin principle with Bayesian pos-

Appearing in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

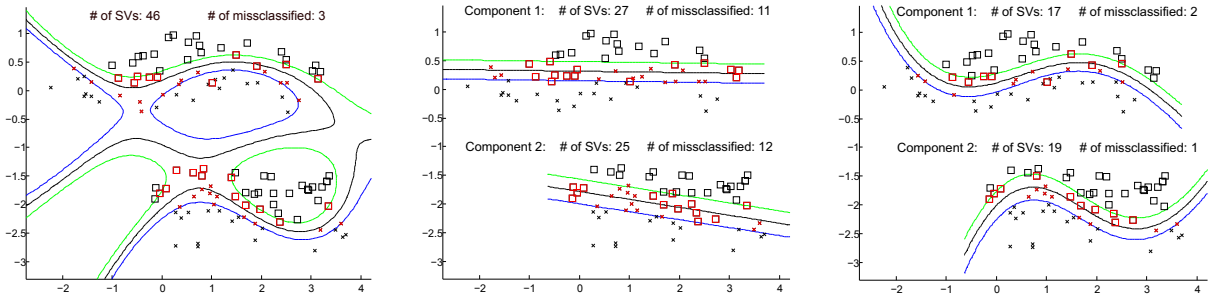

Figure 1. An illustration of iSVM for binary classification. The two classes are denoted by "squares" and "stars". We show the decision boundaries in black lines of: (L) the global SVM using radial basis function (RBF) kernel; (M) iSVM mixture of linear SVMs; and (R) iSVM mixture of SVMs using RBF kernels. Data points in red are support vectors.

terior inference. In such a way, iSVM combines the advantages of Bayesian nonparametrics to resolve the *unknown* number of mixing experts, and large-margin kernel machines to capture local nonlinearity without creating unnecessary experts, as well as handle rich forms of inputs. In addition, by avoiding employing a normalized distribution of response variables, iSVM can be learned efficiently based on existing high-performance techniques for DP and SVM. Our work represents the first attempt toward combining large-margin kernel machines with Bayesian nonparametrics, and our empirical results appear to verify the expected advantages inherited from both ideas.

Fig. 1 offers a direct illustration of these ideas using binary data whose features are sampled from a mixture of two Gaussians. Unlike the global SVM that learns a single nonlinear classifier on all the data (Left), iSVM automatically identifies the two components and learns a local large-margin classifier within each component (Middle and Right). For a particular data point, the final prediction is a weighted voting according to the mixing distributions over the components. As shown in Fig. 1 (and Table 2), the global RBF-SVM usually has more support vectors (SVs) than the component classifiers in RBF-iSVM or even more SVs than the entire RBF-iSVM. Fewer SVs will lead to more efficient testing for kernel methods. Finally, the mixture-of-linear-experts model (Middle) apparently needs extra (more than two in total) experts in order to capture the local nonlinearity within each component.

2. Infinite Large-margin Machines

We begin with a brief introduction of the MED large-margin machines and DP mixtures; followed by the iSVM which builds on these two lines of thoughts.

2.1. MED and Finite Mixture of SVMs

Let $\mathbf{x} \in \mathbb{R}^M$ be an input feature vector. We consider the general multi-way classification, where the response variable Y takes values from a finite set $\{1, \cdots, L\}$. Let $F(y, \mathbf{x}; \eta)$ be a discriminant function parameterized by η. Unlike standard SVMs, which perform point-estimate of η and typically lack a direct probabilistic interpretation, MED (Jaakkola et al., 1999) learns a distribution $q(\eta)$ by solving an entropic regularized risk minimization problem with prior $p_0(\eta)$

$$\min_{q(\eta)} \ \mathrm{KL}(q(\eta)\|p_0(\eta)) + C_1\mathcal{R}(q(\eta)), \tag{1}$$

where C_1 is a positive constant and $\mathrm{KL}(p\|q)$ is the Kullback-Leibler divergence; $\mathcal{R}(q(\eta)) = \sum_d \max_y(\ell_d^{\Delta}(y) + \mathbb{E}_{q(\eta)}[F(y, \mathbf{x}_d; \eta) - F(y_d, \mathbf{x}_d; \eta)])$ is the hinge-loss that captures the large-margin principle underlying the MED prediction rule

$$y^* = \arg\max_y \mathbb{E}_{q(\eta)}[F(y, \mathbf{x}; \eta)] \tag{2}$$

on training data $\mathcal{D} = \{(\mathbf{x}_d, y_d)\}_{d=1}^D$; and $\ell_d^{\Delta}(y)$ measures how y differs from the true label y_d.

Due to the Bayesian-style formulation, MED provides an elegant way to integrate the ideas of large-margin learning, kernel machines, and Bayesian generative modeling. In fact, MED subsumes SVM as a special case, and it has been extended to incorporate latent variables (Lewis et al., 2006) and perform structured output prediction (Zhu & Xing, 2009). However, as we have stated, learning a single global SVM/MED classifier with all the data could be expensive because the complexity scales with the training-set size, and it may also be limited without considering underlying data structures. *Finite* mixture of SVM tries to address this problem by splitting the input data into several subgroups *a priori* or using some heuristics and learning a simpler SVM/MED classifier within each group (Collobert et al., 2002; Fu et al., 2010).

The *Infinite SVM* (iSVM) is a novel extension of MED by using Bayesian nonparametric techniques to resolve the unknown number of experts in mixture models. Unlike existing DP mixtures, iSVM inherits the advantages of large-margin learning and kernel methods.

2.2. Dirichlet Process Mixtures

Ferguson (1973) first introduced the Dirichlet process (DP), and Sethuraman (1994) provided a stick-breaking representation, that is, the random

measure G distributed according to a Dirichlet process $\mathcal{DP}(G_0, \alpha)$ with base distribution G_0 and concentration parameter α can be written as

$$G = \sum_{i=1}^{\infty} \pi_i(\mathbf{v})\delta_{\theta,\theta_i},\ \pi_i(\mathbf{v}) = v_i \prod_{j=1}^{i-1}(1-v_j)$$

where $\theta_i \sim G_0$, $v_i \sim \text{Beta}(1, \alpha)$, and $\delta_{a,b}$ is the Kronecker delta function (i.e., $\delta_{a,b} = 1$ if $a = b$; otherwise 0). It is clear from this formulation that G is discrete almost surely (Blackwell & MacQueen, 1973), that is, the support of G consists of a countably infinite set of atoms, which are drawn independently from G_0.

Antoniak (1974) first introduced the idea of using a DP as the prior for the mixing proportions of simple distributions (e.g., Gaussian). Due to the fact that the distributions sampled from a DP are discrete almost surely, data generated from a DP mixture can be partitioned according to the distinct values of the sampled distributions. Therefore, DP mixture is a flexible mixture model, in which the number of components is random and grows as new data are observed.

2.3. Infinite Large-margin Machines

Now, we formally present iSVM, which consists of a DP mixture of large-margin classifiers on Y and a DP mixture of input features X.

DP mixture of large-margin classifiers: We define the DP mixture of large-margin classifiers as the MED that uses a Dirichlet process prior, that is, the prior in problem (1) follows a DP

$$p_0(\eta) \sim \mathcal{DP}(G_0, \alpha).$$

Due to the discrete nature of DP, there is a positive probability that the sampled $\eta_i \sim p_0(\eta)$ have identical values. Therefore, we can assign data samples to different classifiers for training or testing. We formalize this idea using the stick-breaking representation. Let Z be an assignment variable of the component with which data $\mathbf{x}$ is associated. The process of determining which classifier from a countably infinite set of candidates is used to classify a sample can be described as

1. draw $V_i|\alpha \sim \text{Beta}(1, \alpha),\ i \in \{1, 2, \cdots\}$.
2. draw $\eta_i|G_0 \sim G_0,\ i \in \{1, 2, \cdots\}$.
3. for the dth data point:
   (a) draw $Z_d|\{v_1, v_2, \cdots\} \sim \text{Mult}(\pi(\mathbf{v}))$

Once Z has been given, we define the component-wise discriminant function

$$F(y, \mathbf{x}; z, \boldsymbol{\eta}) = \eta_z^\top \mathbf{f}(y, \mathbf{x}) = \sum_{i=1}^{\infty} \delta_{z,i} \eta_i^\top \mathbf{f}(y, \mathbf{x}), \quad (3)$$

where $\mathbf{f}(y, \mathbf{x})$ is a ML-dim vector stacking L subvectors of which the yth is $\mathbf{x}$ and all the other subvectors are zero. As in MED, we consider learning a distribution of η for each component classifier and take the average over all possible η according to the learned distribution. Since Z is a latent variable, we define the overall discriminant function

$$\begin{aligned} F(y, \mathbf{x}) &= \mathbb{E}_{q(z,\boldsymbol{\eta})}[F(y, \mathbf{x}; z, \boldsymbol{\eta})] = \mathbb{E}_{q(z,\boldsymbol{\eta})}[\eta_z]^\top \mathbf{f}(y, \mathbf{x}) \\ &= \sum_{i=1}^{\infty} q(z = i)\mathbb{E}_q[\eta_i]^\top \mathbf{f}(y, \mathbf{x}), \end{aligned} \quad (4)$$

where the expectation is taken over the posterior distribution (or its approximation) of Z and $\boldsymbol{\eta}$. Then, a natural prediction rule for classification is

$$y^* = \arg\max_y F(y, \mathbf{x}). \quad (5)$$

Now, we need to devise an appropriate loss function of the prediction rule (5) to learn an optimal posterior distribution $p(\mathbf{z}, \boldsymbol{\eta}|\mathcal{D})$ (or an approximation), where $\mathbf{z}$ denotes the component assignment of all samples. Following the large-margin idea, we can show that

$$\mathcal{R}(q(\mathbf{z}, \boldsymbol{\eta})) = \sum_d \max_y (\ell_d^{\Delta}(y) + F(y, \mathbf{x}_d) - F(y_d, \mathbf{x}_d))$$

is an upper bound of the training error of rule (5) on $\mathcal{D}$ for any distribution $q(\mathbf{z}, \boldsymbol{\eta})$. Then, similar as in MED, we define the learning problem as to find the optimal distribution $q(\mathbf{z}, \boldsymbol{\eta})$ by solving the entropic regularized risk minimization problem

$$\min_{q(\mathbf{z},\boldsymbol{\eta})} \text{KL}(q(\mathbf{z}, \boldsymbol{\eta}) \| p_0(\mathbf{z}, \boldsymbol{\eta})) + C_1 \mathcal{R}(q(\mathbf{z}, \boldsymbol{\eta})). \quad (6)$$

From the generating process, it is clear that the prior $p_0(\mathbf{z}, \boldsymbol{\eta}) = \prod_d p(z_d|\alpha) \prod_i p(\eta_i|G_0)$, where $p(z|\alpha) = \int_{\mathbf{v}} p(z|\mathbf{v})p(\mathbf{v}|\alpha)$.

We have defined problem (6) as an extension of the MED principle to the flexible DP mixtures. Notice that our formulation is fundamentally different from the work (Lewis et al., 2006) that performs large-margin learning of finite mixture models by defining the discriminant function as the log-likelihood ratio, which is highly nonlinear and hard to deal with.

DP mixture of input features: for the observed features, they can be described as arising from a similar generative process as above. Let $\boldsymbol{\gamma}$ be the set of all $\gamma_i \sim G_0$. Once Z_d has been sampled, the observed features are generated as $X_d|z_d, \boldsymbol{\gamma} \sim p(\mathbf{x}_d|\gamma_{z_d})$.

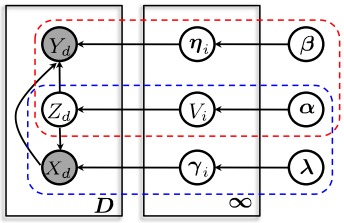

Figure 2. Graphical model representation of iSVM.

Similar as in (Blei & Jordan, 2006), we consider the broad exponential family of distributions for $\mathbf{x}$

$$p(\mathbf{x}|z,\boldsymbol{\gamma}) = \prod_{i=1}^{\infty}(h(\mathbf{x})\exp\{\gamma_i^\top \mathbf{x} - A(\gamma_i)\})^{\delta_{z,i}},$$

where A is the log-partition function, and the base distribution for $\boldsymbol{\gamma}$ is the corresponding conjugate prior

$$p_0(\boldsymbol{\gamma}|\lambda) = \prod_{i=1}^{\infty} h(\gamma_i)\exp\{\lambda_1^\top\gamma_i + \lambda_2(-A(\gamma_i)) - A(\lambda)\}.$$

Hybrid learning and inference: As illustrated graphically in Fig. 2, the two parts of iSVM as in "red" and "blue" boxes are closely coupled by sharing the same Z and the same infinite mixing proportion vector $\pi(\mathbf{v})$. Now, we need to devise a hybrid objective to learn an optimal distribution $q(\mathbf{z},\boldsymbol{\eta})$ for the DP mixture of large-margin classifiers and infer $p(\mathbf{z},\boldsymbol{\gamma},\mathbf{v}|\mathcal{D})$ (or an approximation) for the DP mixture of input features to uncover the underlying structures.

We have defined the learning problem (6) for the DP mixture of classifiers. For the other part of DP mixture of inputs, exact inference for $p(\mathbf{z},\boldsymbol{\gamma},\mathbf{v}|\mathcal{D})$ is generally intractable. Here, we choose to use variational methods which will yield an optimization problem that can be integrated with problem (6). Specifically, we solve the following problem for the optimal variational approximation $q(\mathbf{z},\boldsymbol{\gamma},\mathbf{v})$

$$\min_{q(\mathbf{z},\boldsymbol{\gamma},\mathbf{v})} \mathrm{KL}(q(\mathbf{z},\boldsymbol{\gamma},\mathbf{v})\|p(\mathbf{z},\boldsymbol{\gamma},\mathbf{v}|\mathcal{D})). \tag{7}$$

Putting the two parts together, we define the hybrid learning and inference problem as to find the optimal distribution $q(\mathbf{z},\boldsymbol{\eta},\boldsymbol{\gamma},\mathbf{v})$ by solving the problem

$$\min_{q\in\mathbb{Q}} f(q(\mathbf{z},\boldsymbol{\eta})) + C_2\mathrm{KL}(q(\mathbf{z},\boldsymbol{\gamma},\mathbf{v})\|p(\mathbf{z},\boldsymbol{\gamma},\mathbf{v}|\mathcal{D})), \tag{8}$$

where $f(q(\mathbf{z},\boldsymbol{\eta}))$ is the objective of problem (6); $\mathbb{Q}$ is the family of the joint distribution $q(\mathbf{z},\boldsymbol{\eta},\boldsymbol{\gamma},\mathbf{v})$; and C_2 is another non-negative constant.

By minimizing the hybrid learning and inference objective in problem (8), all the local experts are closely coupled, and the prediction rule is also coupled with the DP mixture of observed features. These strong couplings will cause different components of iSVM to cooperate nicely. Therefore, we can expect to learn a DP mixture model that could discover the underlying structures and can predict well on unseen data.

Notice that for the ease of understanding, we have defined the component-wise discriminant function $F(y,\mathbf{x};z,\boldsymbol{\eta})$ as linear in terms of the parameters $\boldsymbol{\eta}$. However, this linearity is local, and the weighted prediction rule (5) is nonlinear because of the data dependent mixing weights $q(Z)$. Moreover, using kernel functions on $\mathbf{x}$ to learn nonlinear component-wise classifiers can be easily extended, as explained later.

3. Opt. with Coordinate Descent

To make the problem tractable, we make the mean field assumption (Blei & Jordan, 2006) with truncated stick-breaking representations, that is, $q(\mathbf{z},\boldsymbol{\eta},\boldsymbol{\gamma},\mathbf{v}) = \prod_{d=1}^{D} q(z_d)\prod_{t=1}^{T} q(\eta_t)\prod_{t=1}^{T} q(\gamma_t)\prod_{t=1}^{T-1} q(v_t)$, where T is a variational parameter that can be freely set and $q(v_T = 1) = 1$ meaning that for any $t > T$, $\pi_t(\mathbf{v}) = 0$. Furthermore, we assume that $q(z_d)$ is a multinomial distribution with parameter $\phi_d = (\phi_d^1,\cdots,\phi_d^T)$ and $q(v_t) = \mathrm{Beta}(\nu_{t,1},\nu_{t,2})$ is a Beta distribution[1].

Since the regularizer $\mathrm{KL}(q(\mathbf{z},\boldsymbol{\eta})\|p_0(\mathbf{z},\boldsymbol{\eta}|\alpha,\beta))$ is hard to evaluate, we first relax it by using its upper bound

$$\begin{aligned}\mathcal{L}^{\mathrm{KL}}(q(\mathbf{z},\boldsymbol{\eta},\mathbf{v})) = {} & \mathrm{KL}(q(\boldsymbol{\eta})\|p_0(\boldsymbol{\eta}|\beta)) + \mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z})] \\ & -\mathbb{E}_{q(\mathbf{z},\mathbf{v})}[\log p(\mathbf{z},\mathbf{v}|\alpha) - \log q(\mathbf{v})].\end{aligned}$$

Then, we solve problem (8) with the above relaxed bound by using alternating minimization. Due to space limitation, we omit the solutions for $q(\mathbf{v})$ and $q(\boldsymbol{\gamma})$, which are in the same closed-form as in standard DP mixtures (Blei & Jordan, 2006). Below, we briefly present how to solve for $q(\mathbf{z})$ and $q(\boldsymbol{\eta})$ and provide more insights of how iSVM works.

For $q(\mathbf{z})$ and $q(\boldsymbol{\eta})$, we solve the simplified problem

$$\min_{q(\mathbf{z},\boldsymbol{\eta})} \mathcal{L}^{\mathrm{KL}}(q(\mathbf{z},\boldsymbol{\eta},\mathbf{v})) + C_1\mathcal{R}(q(\mathbf{z},\boldsymbol{\eta})) + C_2\mathcal{L}_{[q(\mathbf{z})]},$$

where $\mathcal{L}_{[q(\mathbf{z})]} = \mathbb{E}_{q(\mathbf{v})}[\log q(\mathbf{z})] - \mathbb{E}_{q(\mathbf{z},\mathbf{v})}[\log p(\mathbf{z}|\mathbf{v})] - \sum_d \mathbb{E}_q[\log p(\mathbf{x}_d|z_d,\boldsymbol{\gamma})]$. Since the discriminant function $F(y,\mathbf{x})$ is linear, this problem is convex over $q(\mathbf{z},\boldsymbol{\eta})$. We use the Lagrangian method to solve the constrained re-formulation using slack variables $\boldsymbol{\xi}$. By introducing lagrange multipliers ω_d^y, $\forall d, y$, we have the Lagrangian functional $L(q(\mathbf{z},\boldsymbol{\eta}),\boldsymbol{\omega},\boldsymbol{\xi})$.

Optimizing L over $q(\boldsymbol{\eta})$, we have $\forall 1 \le t \le T$: $q(\eta_t) \propto p_0(\eta_t|\beta)\exp\{\eta_t^\top(\sum_d \phi_d^t \sum_y \omega_d^y \mathbf{f}_d^\Delta(y))\}$, where $\mathbf{f}_d^\Delta(y) = \mathbf{f}(y_d,\mathbf{x}_d) - \mathbf{f}(y,\mathbf{x}_d)$. If we assume that the base distribution for η is $\mathcal{N}(\mu_0,\Sigma_0)$, we will have that $q(\eta_t) \sim \mathcal{N}(\mu_t,\Sigma_0)$ with shifted mean

$$\mu_t = \mu_0 + \Sigma_0(\sum_d \phi_d^t \sum_y \omega_d^y \mathbf{f}_d^\Delta(y)). \tag{9}$$

Another choice of the base distribution is Laplace distribution which is useful to identify important explanatory factors of data (Raman et al., 2010; Zhu & Xing, 2009). Here, we choose the standard normal base distribution $\mathcal{N}(0,I)$. Substituting the solution of $q(\eta_t) = \mathcal{N}(\mu_t,I)$ into L, we get the dual QP problem of a multi-way classification SVM

$$\begin{aligned}\max_{\boldsymbol{\omega}} \quad & -\frac{1}{2}\sum_t \mu_t^\top\mu_t + \sum_d\sum_y \omega_d^y \ell_d^\Delta(y) \\ \text{s.t..:} \quad & 0 \le \sum_y \omega_d^y \le C_1,\ \forall d.\end{aligned} \tag{10}$$

[1] We can infer $q(\eta_t)$ and $q(\gamma_t)$ without explicitly assuming their parametric forms by using conjugate priors.

Minimizing L over $q(\mathbf{z})$ and letting $\rho = \frac{C_2}{1+C_2}$, we have

$$\phi_d^t \propto \exp\big\{(\mathbb{E}[\log v_t] + \sum_{i=1}^{t-1}\mathbb{E}[\log(1-v_i)]) \quad (11)$$
$$+\rho(\mathbb{E}[\gamma_t]^\top \mathbf{x}_d - \mathbb{E}[A(\gamma_t)]) + (1-\rho)\sum_y \omega_d^y \mu_t^\top \mathbf{f}_d^\Delta(y)\big\},$$

where $\mathbb{E}[\log v_i] = \psi(\nu_{i,1}) - \psi(\nu_{i,1}+\nu_{i,2})$; $\mathbb{E}[\log(1-v_i)] = \psi(\nu_{i,2}) - \psi(\nu_{i,1}+\nu_{i,2})$; and ψ is the digamma function.

From Eq. (11), we can see that the last term weighted by $(1-\rho)$ plays a role of regularizing the mixing proportions ϕ. For those data points that are at the decision boundary (i.e., support vectors) of the component classifiers, this term biases them toward being re-allocated into the components where they can receive better predictions. Moreover, since the prediction rule (5) only relies on the expectation of $\boldsymbol{\eta}$, which has a linear form (9), and the QP problem (10) only depends on the dot product of data points, the above results can be directly generalized to nonlinear case by using kernel functions on the input $\mathbf{x}$. The same extension applies to the following efficient relaxation.

3.1. An Efficient Relaxation

In the above formulation, all the component classifiers are learned jointly by solving the dual problem (10) or its primal form. Thus, all the training data are involved in solving the QP problems and the number of parameters will be very large if T is large. An easier and more efficient formulation of iSVM is to minimize the following upper bound of $\mathcal{R}$

$$\mathcal{R}(q(\mathbf{z},\boldsymbol{\eta})) \le \sum_d \sum_t \phi_d^t \big[\max_y (\ell_d^\Delta(y) + \mathbb{E}_q[\eta_t]^\top \mathbf{f}_d^\Delta(y))\big],$$

in problem (8), which will lead to the similar optimization procedure as above. For the component-wise classifiers, we are learning T cost-sensitive SVMs (Morik et al., 1999)

$$\min_{\mu_t, \boldsymbol{\xi}^t} \quad \frac{1}{2}\mu_t^\top \mu_t + C_1 \sum_d \phi_d^t \xi_d^t$$
$$\text{s.t.. :} \quad \mu_t^\top \mathbf{f}_d^\Delta(y) \ge \ell_d^\Delta(y) - \xi_d^t, \forall d, y. \quad (12)$$

Although T can be large, each component classifier can be efficiently learned because on average only a few data samples have large ϕ_d^t in component t. If ϕ_d^t is small enough (e.g., less than $1e^{-10}$), the document d can be safely discarded in learning component classifier t. Another nice property of this error bound is that it is linear of ϕ. Thus, we can easily solve for ϕ in a closed-form. Finally, besides computational simplicity and efficiency, this relaxation allows local experts to compete with each other to make their individual predictions for a data point[2], while the formulation with $\mathcal{R}$ encourages all the experts to cooperate to make a final prediction. The competing mechanism could potentially result in sparser mixing proportions (Jacobs et al., 1991).

Finally, the SVMs (12) that uses multiple slack variables can be equivalently reformulated using a single slack variable, which can be more efficiently learned using a cutting plane algorithm (Joachims et al., 2009). In our experiments, we solve the so called 1-slack formulation of SVMs under the above efficient relaxation.

3.2. Testing with Variational Methods

When a new data example $\mathbf{x}$ comes in, we need to infer the posterior distribution

$$p(z, \boldsymbol{\eta}|\mathbf{x}, \mathcal{D}) = p(z|\mathbf{x}, \mathcal{D}, \boldsymbol{\eta}) p(\boldsymbol{\eta}|\mathbf{x}, \mathcal{D})$$

in order to apply the prediction rule (5). We consider the inductive setting[3] and assume that the prediction model (i.e., distribution of $\boldsymbol{\eta}$) only depends on labeled training data $\mathcal{D}$, that is, $p(\boldsymbol{\eta}|\mathbf{x}, \mathcal{D}) = p(\boldsymbol{\eta}|\mathcal{D})$ and $p(z|\mathbf{x}, \mathcal{D}, \boldsymbol{\eta}) = p(z|\mathbf{x}, \mathcal{D})$. We can use the variational distribution $q(\boldsymbol{\eta})$ as inferred in the training phase to approximate $p(\boldsymbol{\eta}|\mathcal{D})$. For $p(z|\mathbf{x}, \mathcal{D})$, we again use variational methods to approximate it, by solving

$$\min_{q(z)} \quad \mathrm{KL}(q(z) \| p(z|\mathbf{x}, \mathcal{D})), \quad (13)$$

where $q(z)$ is a variational distribution. We further relax the objective using its variational upper bound

$$\mathcal{L}^{test}(q(z)q(\mathbf{v},\boldsymbol{\gamma})) = -\mathbb{E}_{q(z)q(\mathbf{v},\boldsymbol{\gamma})}[\log p(z, \mathbf{x}, \mathbf{v}, \boldsymbol{\gamma}|\mathcal{D})]$$
$$-H(q(z)q(\mathbf{v},\boldsymbol{\gamma})) + \log p(\mathbf{x}|\mathcal{D}),$$

where $q(\mathbf{v},\boldsymbol{\gamma})$ is an arbitrary distribution and H is the entropy. We can solve this problem using alternating minimization between $q(\mathbf{v},\boldsymbol{\gamma})$ and $q(z)$. Here, we approximate this procedure by using only one iteration, that is, we fix $q(\mathbf{v},\boldsymbol{\gamma})$ at the distribution that approximates $p(\mathbf{v},\boldsymbol{\gamma}|\mathcal{D})$ inferred at training phase and solving problem (13) (using the upper bound) for $q(z)$ only.

4. Experiments

In this section, we provide empirical studies on both synthetic and real datasets.

4.1. Synthetic Data

We generate synthetic datasets in three different settings. For each setting, we randomly generate 50 datasets. Similar as in (Shahbaba & Neal, 2009), each dataset has 10000 samples, of which 100 samples are

[2] In fact, the upper bound of $\mathcal{R}$ is the expected loss of a stochastic predictor, which draws a component according to $q(Z)$ and predicts using the associated classifier.

[3] It is also interesting to investigate transductive inference (Joachims, 1999), in which both labeled and unlabeled data are present to do joint inference.

Table 1. (L) classification accuracies (%) and (R) F1 scores (%) for different models in three different synthetic situations.

	Setting 1	Setting 2	Setting 3	Setting 1	Setting 2	Setting 3
MNL	71.4 ± 10.1	71.0 ± 8.8	62.0 ± 5.7	66.4 ± 12.3	65.5 ± 8.6	61.3 ± 6.2
linear-SVM	71.8 ± 9.9	71.8 ± 8.2	62.9 ± 5.4	65.1 ± 14.3	65.5 ± 9.9	62.2 ± 5.9
rbf-SVM	74.0 ± 8.7	75.5 ± 5.3	74.4 ± 4.2	69.0 ± 11.8	71.8 ± 5.7	72.3 ± 4.4
dpMNL	73.1 ± 9.9	76.3 ± 6.3	74.2 ± 5.5	68.8 ± 11.9	71.7 ± 8.0	71.1 ± 6.0
linear-iSVM	73.5 ± 8.8	76.9 ± 5.7	75.1 ± 4.7	68.9 ± 11.3	72.1 ± 7.7	72.2 ± 5.1
rbf-iSVM	**74.7** ± 8.6	**78.4** ± 5.2	**76.8** ± 4.1	**69.9** ± 11.7	**74.0** ± 7.8	**74.6** ± 4.2

training data and the rest are testing data. We use Gaussian distributions for $\mathbf{x}$. The three settings are:

Setting 1: the first test is on a binary classification problem with 4 covariates. Data are generated from two Gaussian components. The priors for the two component parameters are $\mu_1 \sim \mathcal{N}(1,1)$ and $\mu_2 \sim \mathcal{N}(2,1)$; $\log(\sigma_i^2) \sim \mathcal{N}(0, 2^2)$, $i = 1, 2$. We draw 5000 examples from each component and sample true labels from a Bernoulli model with $p(y = 1|\mathbf{x}) = \frac{1}{1+e^{-\vartheta}}$ and

$$\vartheta = -a\sin(x_1^3 + 1.2) - x_1\cos(bx_2 + 0.7) - cx_3 + 2,$$

where a, b and c are sampled from $\mathcal{N}(1, 0.5^2)$. Note that the last covariate is not used in the Bernoulli model, which could introduce additional complexity.

Setting 2: the second test is the same as the Simulation 2 in (Shahbaba & Neal, 2009). Specifically, 10000 3-dimensional samples are sampled from uniform distributions in the cubic $[0, 5]^3$ and the true labels are sampled from a similar Bernoulli model as above, except that the term x_1^3 is replaced by $x_1^{1.04}$ in order to produce a balanced distribution on two classes.

Setting 3: the last synthetic test is on datasets generated from a two component mixture of uniform distributions in cubic $[0, 5]^3$ and $[-5, 0]^3$. The true labels are sampled from a Bernoulli model which has the same form as in Setting 2. We sample a, b and c from $\mathcal{N}(1, 0.5^2)$ for the first component and from $\mathcal{N}(-1, 0.5^2)$ for the second component.

We compare with multinomial logit (MNL), DP mixture of MNL (dpMNL) (Shahbaba & Neal, 2009) and SVMs. For SVM and iSVM, we compare their performance using linear and RBF kernels. We set $T = 20$ in inference[4]. iSVM is insensitive to C_2, and we fix it at 64 in all experiments. We use standard normal base distributions for η and γ. For each setting, we generate a validation set for selecting the rest hyperparameters.

[4]Most datasets have 2 or 3 components with large ϕ values. The results don't change when T is set at a larger value. Note that it's hard to determine the true number of components. Take Setting 1 as an example, since the means and variances of two Gaussian components are randomly generated, they can overlap much and the true number of components is larger or smaller than 2.

Table 2. Number of support vectors for RBF-SVM and RBF-iSVM in three different synthetic situations.

	Setting 1	Setting 2	Setting 3
rbf-SVM	16.3 ± 3.4	13.0 ± 1.6	16.5 ± 2.6
rbf-iSVM	**8.5** ± 2.5	**5.9** ± 0.8	**7.5** ± 1.4

Table 1 shows the average accuracy and F1 scores (Shahbaba & Neal, 2009) over all the datasets in each setting, together with standard deviation. We can see that in general linear methods such as MNL and linear SVM are insufficient in modeling the complex data which are not linearly separable. In contrast, mixture models can effectively improve the performance. Furthermore, using large-margin learning and nonlinear experts such as SVM with RBF kernels can potentially further improve. The results in Setting 2 suggest that even when we know that the data are generated from a single cluster, the global classifiers (e.g., SVM with a linear or RBF kernel) might still not be rich enough to capture the complex nonlinearity.

Table 2 compares the average number of support vectors (SVs) of the global RBF-SVM and the component classifiers in RBF-iSVM. We can see that on average the component classifier in iSVM has fewer SVs than RBF-SVM. This suggests that the component classifiers are simpler and more efficient in testing[5]. Moreover, the linear-iSVM, which can be very efficient in training and testing (See Table 3), performs comparably (a bit better) with the global RBF-SVM.

The reason for the large standard deviation of all the tested models is that the randomly generated datasets are very diverse. Fig. 3 compares the performance between RBF-iSVM and dpMNL on all the 50 datasets in Setting 2. We can see that on most datasets, iSVM outperforms dpMNL. Considering the data diversity, our improvements on average performance are quite significant.

[5]In all experiments, we approximate the weighted prediction rule (5) by using the single component classifier with which an input is most likely associated. This approximation doesn't affect the results because the inferred ϕ usually has one element that is much larger than others.

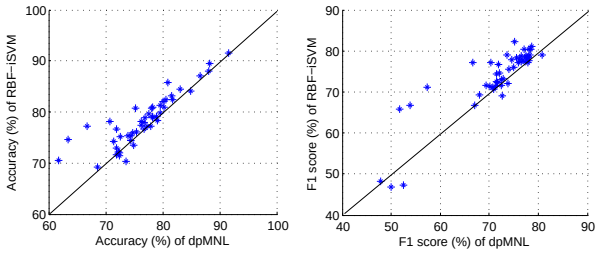

Figure 3. Performance comparison between RBF-iSVM and dpMNL on the 50 datasets in Setting 2.

4.2. High-dimensional Real Data

Now, we study the empirical performance of iSVM on the public dataset[6], which is a subset of the large NUS-WIDE dataset (Chua et al., 2009). It contains 3411 natural scene images downloaded from Flickr website. The images are about 13 types of animals, including tiger, cat and etc. Two types of features are considered in (Chen et al., 2010) for multiview analysis, including 500-dimension SIFT bag-of-words and 634-dimension real-valued features (e.g., color histogram, edge direction histogram, and block-wise color moments). We consider the real-valued features in iSVM by using normal distributions for $\mathbf{x}$.

4.2.1. Prediction Performance

We compare with MNL, dpMNL, SVM and the multi-view method MMH (Chen et al., 2010) that uses both SIFT and real-valued features. We also compare with de-coupled approaches that use either KMeans or DP mixtures to cluster the data and then learn a separate linear SVM classifier for each cluster. We denote the de-coupled approaches as *KMeans+SVM* and *DPM+SVM*, respectively. Since sampling methods are usually slow, previous studies (Shahbaba & Neal, 2009; Hannah et al., 2010) are largely limited to low-dimensional (e.g., tens of dimensions) datasets. For this high-dimensional dataset, the sampling algorithm[7] for dpMNL doesn't work properly on the original features. To make dpMNL efficient and robust, we project the original features into a low-dimensional space. We try both PCA and exponential family harmonium (EFH) (Welling et al., 2004) that uses normal distributions to model real-valued inputs. Similar as in the synthetic experiments, we fix C_2 and use standard normal base distributions for η and γ. We run 5-fold cross-validation on 2054 randomly sampled training data to select the other hyperparameters.

Table 3 shows the average performance of iSVM over five runs with randomly initialized ϕ. Similarly, we present the average performance for dpMNL with five runs of the stochastic sampling algorithm. For dpMNL, we show the best results of using the low-dimensional features produced by EFH and PCA[8]. In general, we have similar observations as in the synthetic experiments: 1) mixture-of-experts can improve the performance; 2) de-coupled methods perform worse than joint methods; and 3) using large-margin and nonlinear local experts can potentially further improve. Finally, the linear-iSVM, which can be very efficient in training (about 200 seconds) and testing, obtains comparable results with the global RBF-SVM, which is about 2 order of magnitude slower in training and 1 order of magnitude slower in testing because of expensive kernel computation. RBF-iSVM has comparable training time with RBF-SVM.

Table 3. Classification accuracy (%), F1 score (%), and test time (sec) for different models on the Flickr image dataset. All methods except dpMNL are implemented in C.

	Accuracy	F1 score	Test time
MNL	49.8 ± 0.0	48.4 ± 0.0	**0.02** ± 0.00
MMH	51.7 ± 0.0	50.1 ± 0.0	0.33 ± 0.01
RBF-SVM	56.0 ± 0.0	52.8 ± 0.0	7.58 ± 0.06
DPMNL-EFH70	51.2 ± 0.9	49.9 ± 0.8	42.1 ± 7.39
DPMNL-PCA50	51.9 ± 0.7	49.9 ± 0.8	27.4 ± 2.08
DPM+SVM	54.9 ± 0.5	52.4 ± 1.0	0.21 ± 0.01
KMEANS+SVM	54.1 ± 0.0	52.1 ± 0.0	**0.02** ± 0.00
LINEAR-ISVM	55.7 ± 0.1	**53.7** ± 0.1	0.22 ± 0.01
RBF-ISVM	**56.4** ± 0.5	53.5 ± 0.7	6.67 ± 0.05

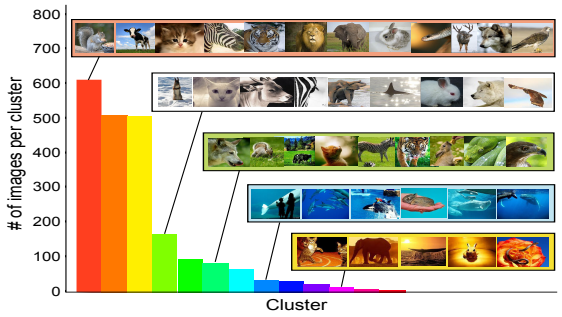

Figure 4. Representative images within different subgroups discovered by iSVM using RBF kernels.

4.2.2. Clustering Structures

Fig. 4 shows the sizes of the subgroups discovered on Flickr data. For each cluster, we present several representative images. We can see that the images in the same cluster tend to have similar background color. One possible reason is that color features are the main features used. We also observe that some clusters (e.g., the 8th) tend to have a few images from a small number of categories while some clusters (e.g, the 1st) tend

[6] http://www.cs.cmu.edu/∼junzhu/subspace_learning.htm

[7] We use the authors' matlab code downloaded from: http://www.ics.uci.edu/∼babaks/Homepage/Codes.html

[8] For PCA, the features are first-k principal components. We tried $k = 10, 20, \cdots, 130$ and reported the best results.

to contain many images from different categories. In general, a simple (but might be nonlinear) classifier is likely to be sufficient to classify the images within each cluster. Also, as explained in Sec 3, the error term in problem (8) could bias an image to be allocated to a cluster in which it can be well classified.

5. Conclusions and Future Work

We have presented *Infinite SVM* (iSVM), a Dirichlet process (DP) mixture of large-margin kernel machines. iSVM integrates the advantages of Bayesian nonparametrics to resolve the unknown number of mixing experts and large-margin kernel methods to capture local nonlinearities. Unlike existing DP mixtures, iSVM learns local experts using large-margin principle without requiring to define a normalized distribution of response variables. We present a variational learning algorithm and demonstrate the advantages of iSVM over existing DP mixtures and other benchmarks on both synthetic and Flickr image classification datasets.

For future work, we can incorporate advances in large-margin kernel methods and Bayesian nonparametrics to extend and improve iSVM. For example, we can use efficient kernel methods (Fine & Scheinberg, 2001; Rahimi & Recht, 2007) to do large-scale applications, and combine multiple kernels (Lewis et al., 2006; Dunson & Bhattacharya, 2010) to deal with heterogenous data; and we can extend the basic Dirichlet process mixtures to model structured inputs (e.g., sequences with Markov property (Beal et al., 2002)) and learn predictors with a structured output space (Zhu et al., 2008).

Acknowledgments

This work is supported by AFOSR FA95501010247, ONR N000140910758, NSF Career DBI-0546594, and an Alfred P. Sloan Research Fellowship to EPX. NC is supported by National Key Project for Basic Research of China (No. G2007CB311003, 2009CB724002).

References

Antoniak, C.E. Mixture of Dirichlet process with applications to Bayesian nonparametric problems. *Annals of Stats*, (273):1152–1174, 1974.

Beal, M., Ghahramani, Z., & Rasmussen, C. The infinite hidden Markov model. In *NIPS*, 2002.

Blackwell, D. & MacQueen, J. Ferguson distributions via polya urn scheme. *Annals of Stats*, (1):353–355, 1973.

Blei, D. & Jordan, M. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1:121–144, 2006.

Chen, N., Zhu, J., & Xing, E. Predictive subspace learning for multiview data: a large margin approach. In *NIPS*, 2010.

Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., & Zheng, Y. NUS-WIDE: a real-world web image database from national university of singapore. In *CIVR*, 2009.

Collobert, R., Bengio, S., & Bengio, Y. A parallel mixture of SVMs for very large scale problems. In *NIPS*, 2002.

Dunson, D. & Bhattacharya, A. Nonparametric Bayes regression and classification through mixtures of product kernels. *Bayesian Stats*, 2010.

Ferguson, T. A Bayesian analysis of some nonparametric problems. *Annals of Stats*, (1):209–230, 1973.

Fine, S. & Scheinberg, K. Efficient SVM training using low-rank kernel representations. *JMLR*, (2):243–264, 2001.

Fu, Z., Robles-Kelly, A., & Zhou, J. Mixing linear SVMs for nonlinear classification. *IEEE Trans. on Nueral Networks*, 21(12):1963 – 1975, 2010.

Hannah, L., Blei, D., & Powell, W. Dirichlet process mixtures of generalized linear models. In *AISTATS*, 2010.

Jaakkola, T., Meila, M., & Jebara, T. Maximum entropy discrimination. In *NIPS*, 1999.

Jacobs, R., Jordan, M., Nowlan, S., & Hinton, G. Adaptive mixtures of local experts. *Neural Computation*, 3 (1):79–87, 1991.

Joachims, T. Transductive inference for text classification using support vector machines. In *ICML*, 1999.

Joachims, T., Finley, T., & Yu, C-N. Cutting-plane training of structural SVMs. *Machine Learning*, 2009.

Lewis, D., Jebara, T., & Noble, W. Nonstationary kernel combination. In *ICML*, 2006.

Liu, X. & Shao, Y. Asymptotics for likelihood ratio tests under loss of identifiability. *Annals of Stats*, 31(3): 807–832, 2003.

Morik, K., Brockhausen, P., & Joachims, T. Combining statistical learning with a knowledge-based approach – a case study in intensive care monitoring. In *ICML*, 1999.

Rahimi, A. & Recht, B. Random features for large-scale kernel machines. In *NIPS*, 2007.

Raman, S., Fuchs, T., Wild, P., Dahl, E., Buhmann, J., & Roth, V. Inifinite mixture-of-experts model for sparse survival regression with application to breast cancer. *BMC Bioinformatics*, 11:S8, 2010.

Rasmussen, C. & Ghahramani, Z. Infinite mixtures of Gaussian process experts. In *NIPS*, 2001.

Schölkopf, B. & Smola, A. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press Cambridge, 2001.

Sethuraman, J. A constructive definition of Dirichlet priors. *Statistica Sinica*, (4):639–650, 1994.

Shahbaba, B. & Neal, R. Nonlinear models using Dirichlet process mixtures. *JMLR*, 10:1829–1850, 2009.

Welling, M., Rosen-Zvi, M., & Hinton, G. Exponential family harmoniums with an application to information retrieval. In *NIPS*, 2004.

Zhu, J. & Xing, E. Maximum entropy discrimination Markov networks. *JMLR*, 10:2531–2569, 2009.

Zhu, J., Xing, E., & Zhang, B. Partially observed maximum entropy discrimination Markov networks. In *NIPS*, 2008.