Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence

Discriminative Clustering via Generative Feature Mapping

Liwei Wang
The Chinese University of Hong Kong
lwwang@cse.cuhk.edu.hk

Xiong Li
Shanghai Jiao Tong University
lixiong@sjtu.edu.cn

Zhuowen Tu
Microsoft Research Asia
University of California, Los Angeles
zhuowent@microsoft.com

Jiaya Jia
The Chinese University of Hong Kong
leojia@cse.cuhk.edu.hk

Abstract

Existing clustering methods can be roughly classified into two categories: generative and discriminative approaches. Generative clustering aims to explain the data and thus is adaptive to the underlying data distribution; discriminative clustering, on the other hand, emphasizes on finding partition boundaries. In this paper, we take the advantages of both models by coupling the two paradigms through feature mapping derived from linearizing Bayesian classifiers. Such the feature mapping strategy maps nonlinear boundaries of generative clustering to linear ones in the feature space where we explicitly impose the maximum entropy principle. We also propose the unified probabilistic framework, enabling solvers using standard techniques. Experiments on a variety of datasets bear out the notable benefit of our method in terms of adaptiveness and robustness.

Introduction

Clustering is a fundamental problem in machine learning. Given a set of unlabeled data, the goal is to group the data samples/points into different clusters, representing the intrinsic structure and underlying membership. In this paper, we study clustering from two angles: *generative* and *discriminative* models.

Generative models, e.g. mixture of Gaussians, are widely used in clustering. Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin 1977) and its variants consist of two iterative estimation procedures: E-step and M-step. The E-step infers the posterior probability of samples belonging to the clusters; the M-step updates the model parameters based on the expectation of the inferred posterior. In addition, the K-means algorithm can be seen as a deterministic version of EM for Gaussian mixtures models (Bishop 2006; Gomes, Krause, and Perona 2010). The advantage of generative methods is two-fold: First, the explicit data distribution is studied in generative models. In particular, the cluster boundaries (though implicitly studied) can be nonlinear and adaptively determined by data distribution. Second, generative models provide a principal way to deal with structured data. However, generative methods do not explicitly parameterize the partition boundaries and maximize their discrimination capability.

Discriminative clustering approaches perform clustering in a more explicit way that directly seek for boundaries among the clusters. Among them, spectral clustering e.g. normalized cuts (Shi and Malik 1997) uses eigendecomposition of affinity to partition data into disjoint clusters. Max-margin clustering (MMC) is proposed in (Xu et al. 2004) inspired by the success of max-margin criterion of support vector machine (SVM) (Vapnik 2000) in discriminative classification. Relaxed from integer optimization, (Xu et al. 2004) formulates the problem as a semi-definite programming(SDP) problem. (Xu 2005) extends the two-class MMC to the multi-cluster case and shows its efficiency in semi-supervised learning for classification.

Solving SDPs is often time-consuming. To make it more practical and adapt to large scale data, (Zhang, Tsang, and Kwok 2009) takes alternating optimization techniques by sequentially solving various support vector regression problems (iterSVR). (Li et al. 2009) proposes the Label-Generating MMC (LG-MMC), that maximizes the margin by label generation, which is formulated as a tighter relaxation convex optimization of MMC problem compared with (Xu et al. 2004). DIFFRAC (Bach and Harchaoui 2007) obtains a closed form expression on this max margin problem by choosing a square loss instead of hinge loss, which is more efficient and adaptive to large scale data. (Gomes, Krause, and Perona 2010) proposes a probabilistic approach to discriminative clustering by formulating the problem as a regularized maximizing mutual information (RIM) between the empirical distribution and the induced label distribution. Note that the capability of these max-margin related methods and RIM to tackle the nonlinear clustering problem – that is, the partition boundaries among clusters are nonlinear – depends on the choice of kernels. A single kernel is usually not sufficient for different datasets. In addition, noisy data make the partition boundary sensitive to outliers.

We aim to couple generative and discriminative clustering to model the partition boundaries explicitly as well as make the method adaptive to data and robust against outliers. This type of combination has demonstrated high usefulness in supervised learning problems (Jaakkola and Haussler 1999; Tu 2007). Feature mapping is one common way, where generative models provide feature mappings (Perina et al. 2009;

Li, Lee, and Liu 2011; 2012) while discriminative models explicitly maximize the discrimination in the feature space. In this regard, generative models together with feature mappings take place of implicit kernel selecting in discriminative methods. However, we find the strategy that learns generative and discriminative models separately is still not perfect because feature mappings from generative models can not be tuned for discriminative boundaries.

In this paper, we investigate the way to build a unified probabilistic framework for discriminative clustering via generative feature mapping. The feature mapping is derived from linearizing Bayesian classifiers based on the observation that sufficient statistics of generative model forms the natural feature mapping from data space to the feature space. This hybrid approach releases discriminative methods from the burden of implicitly selecting kernels given that our generative feature mapping plays the role of kernels but would be more adaptive and explicit. By embedding feature mapping into the maximum entropy framework, the resulting system inherits both the generation and discrimination power, and makes the problem solvable using standard techniques. In various experiments to be reported later, our method exhibits consistent improvements over state-of-the-arts.

Generative Feature Mapping

We begin with the derivation of generative feature mapping from the MAP classifier, where its efficiency can be verified by analyzing the error rate of classifiers with it (Jaakkola and Haussler 1999; Perina et al. 2009; Li, Lee, and Liu 2011). Then we construct the clustering framework using the derived feature mapping in the next section.

First we consider the binary classification problem that assigns labels $y \in \{-1, +1\}$ to samples $\mathbf{x} \in \mathbb{R}^d$. Let $P(\mathbf{x} \mid \theta)$ be the class-conditional distributions for $y = +1$. The decision rule of a MAP/Bayesian classifier is $\hat{y} = \text{sign}(P(\mathbf{x} \mid \theta) - \frac{1}{2})$, which can be noted as $\hat{y} = \text{sign}(\mathcal{L}(\mathbf{x}))$ where $\text{sign}(a) = +1$ for $a > 0$ and $\text{sign}(a) = -1$ otherwise, and

$$\mathcal{L}(\mathbf{x}) \triangleq \log P(\mathbf{x} \mid \theta) + b \tag{1}$$

where $b = -\log \frac{1}{2}$. We consider a general case where the marginal distribution $P(\mathbf{x} \mid \theta)$ is modeled by a hierarchical generative model. Let $P(\mathbf{x}, \mathbf{h} \mid \theta)$ be its joint distribution with random hidden variables $\mathbf{h}$. In this case, it is difficult to obtain the closed form of $P(\mathbf{x} \mid \theta)$ since the integration is usually intractable. A practical method (Jaakkola, Meila, and Jebara 1999) is to use a lower bound of $\log P(\mathbf{x} \mid \theta)$ to estimate $L(\mathbf{x})$. Here we use the lower bound derived by variational inference (Neal and Hinton 1998; Jordan et al. 1999),

$$\log P(\mathbf{x} \mid \theta) \geq -\text{KL}(Q(\mathbf{h}) \,\|\, P(\mathbf{x}, \mathbf{h})) \triangleq F(\mathbf{x}, \theta) \tag{2}$$

where $Q(\mathbf{h}) = \prod_i Q(h_i)$ is the approximate posterior distribution. Then $\mathcal{L}(\mathbf{x})$ can be approximated by following practical discriminant function $\hat{\mathcal{L}}$:

$$\hat{\mathcal{L}}(\mathbf{x}) = F(\mathbf{x}, \theta) + b \tag{3}$$

It is easy to validate that the residual error of Eq. (2) can be written as $\log P(\mathbf{x} \mid \theta) - F(\mathbf{x}) = \text{KL}(Q(\mathbf{h}) \,\|\, P(\mathbf{h} \mid \mathbf{x}))$, so how tight the lower bound is depends on how well $Q(\mathbf{h})$ approximates the real posterior $P(\mathbf{h} \mid \mathbf{x})$.

In this paper, we assume that the joint distributions of adopted generative models belong to exponential family,

$$P(\mathbf{x}, \mathbf{h} \mid \theta) = \exp\{a(\theta)^T T(\mathbf{x}, \mathbf{h}) + S(\mathbf{x}, \mathbf{h}) + d(\theta)\} \tag{4}$$

where θ is the vector of parameters; $T(\mathbf{x}, \mathbf{h})$ is the vector of sufficient statistics; $a(\theta)$ is a vector-valued function; $S(\mathbf{x}, \mathbf{h})$ and $d(\theta)$ are scalar functions. This assumption covers most generative models. As a part of the generative model, $P(\mathbf{h})$ is also belong to exponential family.

$$P(\mathbf{h}; \theta_h) = \exp\{c(\theta_h)^T T(\mathbf{h}) + S(\mathbf{h}) + f(\theta_h)\}$$

It is reasonable to assume, as in (Jordan et al. 1999), that, for each sample $\mathbf{x}$, the approximate posterior $Q(\mathbf{h}; \hat{\theta})$ shares the same form as the real posterior, but with different parameters:

$$Q(\mathbf{h} \mid \mathbf{x}, \theta'_h) = \exp\{c(\theta'_h)^T T(\mathbf{h}) + S(\mathbf{h}) + f(\theta'_h)\} \tag{5}$$

where $c(\theta'_h)$ and $f(\theta'_h)$ depend on each sample $\mathbf{x}$. For different samples $\mathbf{x}$, $Q(\mathbf{h} \mid \mathbf{x}, \theta'_h)$ shares the same form but with corresponding parameters $c(\theta'_h)$ and $f(\theta'_h)$. Denoting $Q(\mathbf{h}) \triangleq Q(\mathbf{h} \mid \mathbf{x}, \theta'_h)$, it can be verified that

$$\begin{aligned} F(\mathbf{x}, \theta) = &\, \text{E}_{Q(\mathbf{h})}[a(\theta)^T T(\mathbf{x}, \mathbf{h}) + S(\mathbf{x}, \mathbf{h}) + d(\theta) \\ &\, -\mathbf{1}^T \text{diag}(c(\theta'_h)) T(\mathbf{h}) - S(\mathbf{h}) - f(\theta'_h)] \\ = &\, \alpha^T \text{E}_{Q(\mathbf{h})}[\tilde{T}(\mathbf{x}, \mathbf{h})] + \beta \end{aligned} \tag{6}$$

where $\tilde{T}(\mathbf{x}, \mathbf{h}) \triangleq (T(\mathbf{x}, \mathbf{h})^T, S(\mathbf{x}, \mathbf{h}), (\text{diag}(\mathbf{c}(\theta'_h)) T(\mathbf{h}))^T, S(\mathbf{h}), f(\theta'_h))^T$; constants $\alpha \triangleq (a(\theta)^T, 1, -\mathbf{1}^T, -1, -1)^T$, $\beta \triangleq d(\theta)$. It is worth noting that $\tilde{T}(\mathbf{x}, \mathbf{h})$ is independent with the parameters of the adopted generative models.

Substituting the lower bound Eq. (6) into Eq. (3), we have:

$$\hat{\mathcal{L}}(\mathbf{x}) = F(\mathbf{x}, \theta) + b = \tilde{\alpha}^T \text{E}_{Q(\mathbf{h})}[\phi(\mathbf{x}, \mathbf{h})]$$

where $\phi(\mathbf{x}, \mathbf{h}) \triangleq (\tilde{T}(\mathbf{x}, \mathbf{h})^T, 1)^T$ and $\tilde{\alpha} = (\alpha^T, \beta + b)^T$. This is the linear form of MAP classifier, where $\text{E}_Q[\phi(\mathbf{x}, \mathbf{h})]$ is considered to be feature mapping derived from generative models, where the sufficient statistic of $P(\mathbf{x}, \mathbf{h}; \theta)$ and $P(\mathbf{h}; \theta_h)$ are the most informative entities. For simplicity, we note this feature mapping as:

$$\Phi(\mathbf{x}) = \text{E}_Q[\phi(\mathbf{x}, \mathbf{h})] \tag{7}$$

Compared with (Jaakkola and Haussler 1999; Perina et al. 2009; Li, Lee, and Liu 2011), the form of the derived feature mapping is simpler, essentially dominated by the sufficient statistics of the adopted generative models. Compared with (Li, Lee, and Liu 2012), the feature mapping in Eq. (7) takes a deterministic form (that maps an input sample to be a real-valued vector) instead of the stochastic one (that maps an input sample to be a random vector with some distribution), which further makes the following theoretical justification feasible.

Theorem 1. ***(Comparison with MAP classifiers)*** *With the feature mapping $\Phi(\mathbf{x})$ (Eq. (7)) derived from the probabilistic distribution $P(\mathbf{x} \mid \theta)$, the error rate of a linear classifier*

$R(\Phi)$ is at least as low as that of the practical MAP classifier (Eq. (3)) $R_{\hat{\mathcal{L}}}$:

$$R(\Phi) \leq \mathrm{E}_{\mathbf{x},y} L\left\{-y\hat{\mathcal{L}}(\mathbf{x})\right\} \triangleq R_{\hat{\mathcal{L}}},$$

where $L(a)$ is a zero-one loss function, outputting 1 if $a > 0$ and 0 otherwise.

Proof. For the linear classifier $\mathbf{w} \in \mathbb{R}^n$, the following inequality always holds:

$$\min_{\mathbf{w}} \mathrm{E}_{\mathbf{x},y} L[-y(\mathbf{w}^T\Phi(\mathbf{x}))] \leq \mathrm{E}_{\mathbf{x},y} L[-y(\mathbf{w}^T\Phi(\mathbf{x}))].$$

Such inequality also holds when $\mathbf{w}$ in the right side of inequality is taken replace by a certain vector, i.e., the weight $\tilde{\alpha}$ derived from Eq.(7). Thus we have

$$\begin{aligned} R(\Phi) &= \min_{\mathbf{w}} \mathrm{E}_{\mathbf{x},y} L\{-y(\mathbf{w}^T\Phi(\mathbf{x}))]\} \\ &\leq \mathrm{E}_{\mathbf{x},y} L\{-y\tilde{\alpha}^T \mathrm{E}_Q[\phi(\mathbf{x},\mathbf{h})]\} \\ &= \mathrm{E}_{\mathbf{x},y} L\{-y[F(\mathbf{x},\theta)+b]\} \\ &= \mathrm{E}_{\mathbf{x},y} L\{-y\hat{\mathcal{L}}(\mathbf{x})\}, \end{aligned}$$

□

The theorem suggests that, with this feature mapping, the linear classifier inherits MAP classifier's ability in adaptively inferring nonlinear partition boundaries. Although the feature mapping is derived from class-conditional case, it can be straightforwardly used when the class label is unavailable. In this case, the feature mapping is still $\Phi(\mathbf{x}) = \mathrm{E}_Q[\phi(\mathbf{x},\mathbf{h})]$ where $\phi(\mathbf{x},\mathbf{h}) = (\tilde{T}(\mathbf{x},\mathbf{h})^T, 1)^T$, and $P(\mathbf{x}\,|\,\theta)$ is the distribution of the all data points instead of data points from a class.

Generative-Discriminative Clustering

As proved in Theorem 1, our feature mapping can map the nonlinear boundaries to linear ones. In this section, we construct a unified framework for discriminative clustering coupled with generative feature mapping.

We make use of the maximum entropy principle based classification framework (Jaakkola, Meila, and Jebara 1999), which is written as

$$\begin{aligned} &\min_{Q(\mathbf{w},\gamma^t)} \mathrm{KL}(Q(\mathbf{w},\gamma^t)\,\|\,P(\mathbf{w},\gamma^t)) \\ &\text{s.t. } \mathrm{E}_Q[y^t\mathbf{w}^T\mathbf{x}^t] > \mathrm{E}_Q[\gamma^t],\ \forall\, t, \end{aligned} \tag{8}$$

where $y = \{-1,+1\}$ is the label for two-classes case; γ^t is the margin for sample $\mathbf{x}^t$. The KL term measures the divergence of the posterior and prior distribution. When we specify the prior $P(\mathbf{w}) = N(0,\mathrm{I})$, minimizing the term encourages $\mathbf{w}$ to have a short length.

The problem in Eq.(8) takes the general form of maximum entropy (Jaakkola, Meila, and Jebara 1999) or posterior regularization (Graça, Ganchev, and Taskar 2007). This well studied problem can be solved via its dual form (Csiszar 1975; Bertsekas 1999) with the general solution:

$$Q(\mathbf{w},\gamma^t) = \frac{1}{Z(\lambda)} P(\mathbf{w},\gamma^t) \exp\left\{\sum_t \lambda_t[y^t\mathbf{w}^T\mathbf{x}^t - \gamma^t]\right\} \tag{9}$$

where $\lambda = \{\lambda_1, \cdots, \lambda_T\}$ are the non-negative Lagrange multipliers, one for each constraint, and $Z(\lambda)$ is the partition function. λ can be determined by finding the unique maximum of the concave function $J(\lambda) = -\log Z(\lambda)$.

The Proposed Approach

We extend the above implementation to multi-class and unsupervised case. The point is to treat the label as a hidden random variable following Multinomial distribution, and constrain the margin of every two classes instead of maximizing those margins (Xu 2005). We couple the discriminative principle and the generative models via the derived feature mapping and express discriminative clustering under the maximum entropy principle as

$$\begin{aligned} &\min_{\mathcal{Q},\theta} \eta \sum\nolimits_t \mathrm{KL}(Q(\mathbf{h}^t)\,\|\,P(\mathbf{x}^t,\mathbf{h}^t;\theta)) + \\ &\qquad\qquad \mathrm{KL}(Q(y^t,\gamma^t,\mathbf{w})\,\|\,P(y^t,\gamma^t,\mathbf{w})) \\ &\text{s.t. } \mathrm{E}_{\mathcal{Q}^t}\left[\sum_{i=1}^K y_i^t\mathbf{w}_i^T\phi^t - (1-y_j^t)\mathbf{w}_j^T\phi^t - \gamma_j^t\right]\mathrm{I}(y_j^t \neq 1) \geq 0\ \ \forall\, t,j, \end{aligned} \tag{10}$$

where $\mathcal{Q} = \{Q(\mathbf{h}^t), Q(\gamma^t), Q(y^t), Q(\mathbf{w})\}_t$; K is the number of clusters; $P(\mathbf{x},\mathbf{h};\theta)$ is the joint distribution of the chosen generative model; $P(\mathbf{w}) = N(0,\mathbf{I})$ is the prior of weight; $P(\gamma_j^t) = c\exp\{-c(d-\gamma_j^t)\}$ for $\gamma_j^t \leq d$ is the prior of margin; $P(\mathbf{y})$ is the Multinomial prior of cluster label. The two KL terms are objective functions for generative model (Eq.(2)) and discrimination power (Eq.(8)) respectively, with a parameter $\eta > 0$ tuning their balance.

To simplify the inference and parameter estimation procedures, we simplify the inequality constraints and formulate them as

$$\begin{aligned} &\mathrm{E}_{\mathcal{Q}^t}\left[\sum_{i=1}^K y_i^t\mathbf{w}_i^T\phi^t - (1-y_j^t)\mathbf{w}_j^T\phi^t - \gamma_j^t\right]\mathrm{I}(y_j^t \neq 1) \geq 0,\ \ \forall\, t,j, \\ &\Leftrightarrow \mathrm{E}_{\mathcal{Q}^t}\left[\sum_i \tilde{y}_{ij}^t\mathbf{w}_i^T\phi^T - \gamma_j^t\right]\mathrm{I}(y_j^t \neq 1) \geq 0,\ \ \forall\, t,j \end{aligned} \tag{11}$$

where $\tilde{y}_{ij}^t = y_i^t - (1-y_i^t)\mathrm{I}(i=j)$ measures, in comparison with class j, the extent of the sample t belonging to class i. Note that we exclude the case $y_j^t = 1$ through $\mathrm{I}(y_j^t \neq 1)$. In this case, $\tilde{y}_{jj} = 1$ and rest elements of $\tilde{\mathbf{y}}^t$ are 0. Therefore the excluded inequalities are $\mathrm{E}_{\mathcal{Q}^t}\left[\mathbf{w}_j^T\phi^T - \gamma_j^t\right] \geq 0, \forall j,t$ which simply constrains the 'margin' of single class instead of two classes. We claim that including these cases would make the model simple without causing degeneration. Thus the constraints used in following formulation are:

$$\mathrm{E}_{\mathcal{Q}^t}\left[\sum_i \tilde{y}_{ij}^t\mathbf{w}_i^T\phi^T - \gamma_j^t\right] \geq 0,\ \ \forall\, t,j \tag{12}$$

The effectiveness of these simplified constraints will be validated in our experiments.

Inference and Parameter Estimation

Note that the forms in Eqs. (10) and (12) are consistent with that in Eq. (8). It is easy to derive the solution of Eq. (10) analogous to Eq. (9). Because the inference procedure involves estimating multiple variables and parameters, we employ an iteration strategy (Friedman, Hastie, and Tibshirani 2008). Also, since $P(\mathbf{w})$ follows a Gaussian distribution in our problem, it is possible to integrate over $\mathbf{w}$ when calculating the marginal distribution for other variables.

Solution of $Q(\mathbf{h}^t)$ Fixing $Q(\gamma_j^t)$, $Q(\mathbf{y}^t)$, and λ, we apply the standard solver of Eq. (9), which yields

$$
\begin{aligned}
&Q(\mathbf{h}^t, \mathbf{w}) \\
&\propto P(\mathbf{x}^t, \mathbf{h}^t, \mathbf{w}) \exp\left\{\sum_{t,j} \frac{\lambda_j^t}{\eta}\left[\sum_j \mathrm{E}[\tilde{y}_{ij}^t]\mathbf{w}_i^T\phi^t - \mathrm{E}[\gamma_j^t]\right]\right\} \\
&\propto P(\mathbf{x}^t, \mathbf{h}^t)\prod_i P(\mathbf{w}_i)\exp\left\{\sum_{t,j}\frac{\lambda_j^t}{\eta}\left[\mathrm{E}[\tilde{y}_{ij}^t]\mathbf{w}_i^T\phi^t - \frac{\mathrm{E}[\gamma_j^t]}{K}\right]\right\}.
\end{aligned}
$$

To get $Q(\mathbf{h}^t) = \int Q(\mathbf{h}^T, \mathbf{w})d\mathbf{w}$, we consider the integral $\int\prod_i P(\mathbf{w}_i)\exp(\cdot)d\mathbf{w} = \prod_i \int P(\mathbf{w}_i)\exp(\cdot)d\mathbf{w}_i$, which can be expanded as

$$
\begin{aligned}
&\prod_i \int P(\mathbf{w}_i)\exp\left\{\sum_{t,j}\frac{\lambda_j^t}{\eta}\left[\mathrm{E}[\tilde{y}_{ij}^t]\mathbf{w}_i^T\phi^t - \frac{\mathrm{E}[\gamma_j^t]}{K}\right]\right\}d\mathbf{w}_i \\
&= \prod_{i=1}^K \exp\left\{\frac{1}{2}\sum_{t,t'}\frac{\lambda_j^t\lambda_{j'}^{t'}}{\eta^2}\mathrm{E}[\tilde{y}_{ij}^t]\mathrm{E}[\tilde{y}_{ij'}^{t'}](\phi^t)^T\phi^{t'} - \sum_{t,j}\frac{\lambda_j^t\mathrm{E}[\gamma_j^t]}{\eta K}\right\} \\
&= \exp\left\{\frac{1}{2}\sum_{t,t',j,j'}\frac{\lambda_j^t\lambda_{j'}^{t'}}{\eta^2}\Big(\sum_i \mathrm{E}[\tilde{y}_{ij}^t]\mathrm{E}[\tilde{y}_{ij}^{t'}]\Big)(\phi^t)^T\phi^{t'} - \sum_{t,j}\frac{\lambda_j^t\mathrm{E}[\gamma_j^t]}{\eta}\right\}.
\end{aligned}
$$

This form gives us the posterior distribution

$$
\begin{aligned}
&Q(\mathbf{h}^t) \propto P(\mathbf{x}^t, \mathbf{h}^T) \qquad (13)\\
&\cdot\exp\left\{\frac{1}{2}\sum_{t,t',j,j'}\frac{\lambda_j^t\lambda_{j'}^{t'}}{\eta^2}\Big(\sum_i \mathrm{E}[\tilde{y}_{ij}^t]\mathrm{E}[\tilde{y}_{ij}^{t'}]\Big)(\phi^t)^T\phi^{t'} - \sum_{t,j}\frac{\lambda_j^t\mathrm{E}[\gamma_j^t]}{\eta}\right\}.
\end{aligned}
$$

Now $\mathrm{E}_{Q(\mathbf{h}^t)}[\phi^t]$ can be easily estimated using samples drawn from $Q(\mathbf{h}^t)$.

Solution of $Q(\gamma_j^t)$ Fixing $Q(y^t)$, $Q(\mathbf{h}^t)$ and λ, we derive

$$
\begin{aligned}
Q(\gamma_j^t) &= \int Q(\gamma_j^t, \mathbf{w})d\mathbf{w} \\
&= P(\gamma_j^t)\int P(\mathbf{w})\exp\Big\{\lambda_j^t\Big[\sum\nolimits_i \mathrm{E}[\tilde{y}_{ij}^t]\mathbf{w}_i^T\mathrm{E}[\phi^t] - \gamma_j^t\Big]\Big\}d\mathbf{w} \\
&\propto P(\gamma_k^t)\exp\left\{-\lambda_j^t\gamma_j^t\right\} \\
&\propto \exp\left\{-(c-\lambda_j^t)(d-\gamma_j^t)\right\}.
\end{aligned}
$$

This yields the expectation with respect to the posterior

$$\mathrm{E}_{Q(\gamma_j^t)}[\gamma_j^t] = d - (c-\lambda_j^t)^{-1}. \qquad (14)$$

Solution of $Q(\mathbf{y}^t)$ Fixing $Q(\gamma^t)$, $Q(\mathbf{h}^t)$, and λ, we derive

$$
\begin{aligned}
&Q(\mathbf{y}^t) = \int Q(\mathbf{y}^t, \mathbf{w})d\mathbf{w} \\
&= P(\mathbf{y}^t)\int P(\mathbf{w})\exp\left\{\sum_j \lambda_j^t\left[\sum_i \tilde{y}_{ij}^t\mathbf{w}_i\mathrm{E}[\phi^t] - \mathrm{E}[\gamma_j^t]\right]\right\}d\mathbf{w} \\
&\propto P(\mathbf{y}^t)\exp\left\{\frac{1}{2}\sum_{t,t',j,j'}\lambda_j^t\lambda_{j'}^{t'}\left(\sum_i \tilde{y}_{ij}^t\tilde{y}_{ij}^{t'}\right)\mathrm{E}[\phi^t]^T\mathrm{E}[\phi^{t'}] - \sum_{t,j}\lambda_j^t\mathrm{E}[\gamma_j^t]\right\}.
\end{aligned}
$$

Since $\mathbf{y}^t$ takes discrete values, the expectation $\mathrm{E}_{Q(\mathbf{y}^t)}[\mathbf{y}^t]$ can be directly calculated given $Q(\mathbf{y}^t)$.

Solution of λ Fixing $Q(\gamma^t, \mathbf{y}^t, \mathbf{h}^t)$, we finally get

$$
\begin{aligned}
Q(\mathbf{w}) &= \frac{1}{Z(\lambda)}P(\mathbf{w})\exp\left\{\sum_{t,j}\lambda_j^t\left[\sum_i \mathrm{E}[\tilde{y}_{ij}^t]\mathbf{w}_i^t\mathrm{E}[\phi^t] - \mathrm{E}[\gamma_j^t]\right]\right\} \\
&= \frac{1}{Z(\lambda)}\prod_i P(\mathbf{w}_i)\exp\left\{\sum_{t,j}\lambda_j^t\left[\tilde{y}_{ij}^t\mathbf{w}_i\mathrm{E}[\phi^t] - \frac{\mathrm{E}[\gamma_j^t]}{K}\right]\right\}.
\end{aligned}
$$

The non-negative λ can be obtained by maximizing the objective function $J_\lambda = -\log Z(\lambda)$ (Jaakkola, Meila, and Jebara 1999), written as

$$J_\lambda = \sum_{j,t}\lambda_j^t\mathrm{E}[\gamma_j^t] - \frac{1}{2}\sum_{t,t',j,j'}\lambda_j^t\lambda_j^t\mathrm{E}[\tilde{y}_j^t]\mathrm{E}[\tilde{y}_{j'}^{t'}]\mathrm{E}[\phi^t]^T\mathrm{E}[\phi^{t'}].$$

It refers to a standard quadratic programming problem.

Parameters θ In the objective function Eq. (10), only the first two KL terms depend on parameters θ. So minimizing the objective function with respect to θ equals minimizing KL with respect to θ, subject to no inequality constraint. *The resulting update rules for θ are the same as those for original generative models*.

Balance of Generative and Discriminative Models

The discriminative model determines the form of the posterior $Q(\mathbf{y}^t)$ here. But, in quantity, the dominated factor of $Q(\mathbf{y}^t)$ is the feature mapping $\mathrm{E}_{Q(\mathbf{h}^t)}[\phi(\mathbf{x}^t, \mathbf{h}^t)]$ and therefore, $Q(\mathbf{h}^t)$ is jointly determined (Eq.(13)) by the generative and discriminative models. Eq.(13) naturally implements a mechanism balancing the two different paradigms, and the balance can be tuned by the parameter η.

Experiments

We conduct comprehensive evaluations on the UCI, MNIST digit and Multi-PIE face data.

Clustering Accuracy

Clustering accuracy is a commonly adopted measure in assessing clustering methods. We first take a set of data with label y, then remove the label and run each clustering methods output with the predicted label $\hat{y}$. We adopt the definition of the maximal classification accuracy among all possible permutation mappings (Wu and Scholkopf 2007; Chen et al. 2011):

$$Accuracy = \frac{\sum_{i=1}^N \delta(y_i, f(\hat{y}_i))}{N}$$

where N is the number of samples. $f(\cdot)$ maps each predicted cluster label to a class label and the optimal matching can be found by Hungarian algorithm (Wu and Scholkopf 2007; Chen et al. 2011). $\delta(\cdot)$ outputs 1 for $y_i \equiv f(\hat{y}_i)$ and 0 otherwise. Since the accuracy of clustering approaches sometimes depends on the random initialization, we run all evaluated methods for 20 times and report both average accuracy and standard deviation.

Data	#class	K-means	Spectral	ITERSVR	LG-MMC	DIFFRAC	Ours
Sonar	2	53.37 ± 0.00	50.48 ± 0.00	55.29 ± 0.00	$\mathbf{70.67 \pm 0.00}$	55.76 ± 0.00	55.77 ± 0.00
Credit	2	57.53 ± 0.00	66.67 ± 0.00	62.32 ± 0.00	54.35 ± 0.00	56.73 ± 0.00	$\mathbf{72.03 \pm 0.00}$
SpHeart	2	59.92 ± 0.00	53.50 ± 0.00	71.54 ± 0.00	71.91 ± 0.00	79.40 ± 0.00	$\mathbf{79.78 \pm 0.00}$
Cancer	2	85.41 ± 0.00	86.12 ± 0.00	83.66 ± 0.00	89.98 ± 0.00	89.10 ± 0.00	$\mathbf{90.51 \pm 0.00}$
Wine	3	80.90 ± 0.00	72.47 ± 0.00	-	-	66.29 ± 0.00	$\mathbf{92.62 \pm 0.00}$
Iris	3	85.33 ± 4.79	$\mathbf{90.00 \pm 0.00}$	-	-	74.67 ± 0.00	89.60 ± 0.34
Tissue	6	37.74 ± 0.00	40.56 ± 0.00	-	-	40.57 ± 0.00	$\mathbf{50.00 \pm 0.00}$

Table 1: Summary of clustering accuracy (%±std) on UCI dataset

Data	#class	K-means	Spectral	ITERSVR	LG-MMC	DIFFRAC	Ours
Digits 3-8	2	79.63 ± 0.55	$\mathbf{92.43 \pm 0.00}$	80.43 ± 0.00	69.86 ± 0.00	87.86 ± 0.00	85.57 ± 0.32
Digits 1-7	2	96.14 ± 0.00	96.52 ± 0.00	94.43 ± 0.00	95.57 ± 0.00	96.86 ± 0.00	$\mathbf{97.97 \pm 0.08}$
Digits 2-7	2	95.28 ± 0.00	$\mathbf{97.57 \pm 0.00}$	95.29 ± 0.00	95.86 ± 0.00	95.43 ± 0.00	95.11 ± 0.53
Digits 8-9	2	90.97 ± 0.09	$\mathbf{97.71 \pm 0.00}$	93.43 ± 0.00	90.14 ± 0.00	87.86 ± 0.00	91.64 ± 0.71
Digits 3-8-9	3	73.99 ± 0.03	75.36 ± 0.16	-	-	70.76 ± 0.00	$\mathbf{75.47 \pm 0.02}$
Digits 1-2-7	3	90.38 ± 0.02	86.87 ± 0.19	-	-	63.33 ± 0.00	$\mathbf{92.86 \pm 0.08}$
Digits 1-8-7	3	92.86 ± 0.00	89.46 ± 0.64	-	-	61.24 ± 0.00	$\mathbf{94.41 \pm 0.06}$
Digits 3-2-8	3	80.89 ± 0.09	$\mathbf{91.14 \pm 0.00}$	-	-	44.76 ± 0.00	80.85 ± 0.36

Table 2: Summary of clustering accuracy (%±std) on MNIST dataset.

Specification of Evaluated Methods

Here we compare the proposed discriminative clustering approach with state-of-the-art ones, including Spectral Clustering (Ng, Jordan, and Weiss 2001), iterSVR (Zhang, Tsang, and Kwok 2009), LG-MMC (Li et al. 2009), and DIFFRAC (Bach and Harchaoui 2007). For Spectral Clustering, we use the implementation released by (Chen et al. 2011) and enable the self-tuning parameters mode. For LG-MMC and iterSVR, we use the author's implementations and follow the strategy in (Li et al. 2009), where the parameter C is chosen from $\{0.1, 0.5, 1.5, 10, 100\}$ and report the best results over linear and RBF kernels. The bandwidth σ of RBF kernel is chosen from $\{0.25, 0.5, 1, 2, 4\}\sqrt{\gamma}$ where γ is the average distance from all pairs of instances. Note that, the implementations of these two methods are for binary clustering, thus the results on multi-class datasets would be absent. For DIFFRAC, we use the author's implementation, and report the best results over linear and RBF kernels. K-means is also involved since it is widely used, with competitive performance.

The proposed approach in Eq.(10) requires to specify a generative model. Here we adopt Gaussian Mixture Model (GMM), a simple yet effective generative one in all our experiments. We also note that our approach is capable of coupling other generative models and is not restricted to GMM. Let $\mathbf{x}$ be the observed variable and $\mathbf{h} = (h_1, \cdots, h_M)^T$ be the hidden variable drawn from Multinomial distribution, that is $P(\mathbf{h}) = \prod_{i=1}^{M} {a_i}^{h_i}$, where $h_i = \{0, 1\}$, $\sum_i h_i = 1$; and $a_i \geq 0$, $\sum_i a_i = 1$. Then the joint distribution of this model can be expressed as $P(\mathbf{x}, \mathbf{h}; \theta) = \prod_{i=1}^{M} N(\mathbf{x}; \mu_i, \Sigma_i)^{h_i} \prod_{i=1}^{M} {a_i}^{h_i}$ where the model parameter $\theta = \{\mu_i, \Sigma_i, a_i\}_{i=1}^{M}$. μ_i is the mean value and $\Sigma_i = \mathrm{diag}(\sigma_{i1}, \cdots, \sigma_{iM})$ is the diagonal covariance matrix of the i-th mixture center. Let $\mathbf{a}' = \{a'_i\}_i$ be the parameter of the approximate posterior of $\mathbf{h}$. Following Eq.(7), feature mapping $\Phi(\mathbf{x}) = \mathrm{E}_Q[\phi(\mathbf{x}, \mathbf{h})] = \mathrm{E}_Q[(\tilde{T}(\mathbf{x}, \mathbf{h})^T, 1)^T]$ is formed with,

$$\tilde{T}(\mathbf{x}, \mathbf{h}) = \mathrm{vec}(\{h_i(\mathbf{x}^T, \mathrm{diag}(\mathbf{x}\mathbf{x}^T), 1), h_i \log a'_i\}_{i=1}^{M})$$

where the posterior $Q(\mathbf{h})$ could be updated according to Eq.(13), and the update rules of θ again are identical with those of GMM model. The cluster label can be decided by

$$\hat{y}^t = \max_i \mathrm{E}_{Q(\mathbf{y}^t)}(\mathbf{y}^t)$$

The number of mixture components usually should satisfy $M \geq K$. The parameters of margin prior (Eq.(10) and Eq.(14)) are set to $c = 10, d = 1$. The balance parameter (Eq.(10)) is set to $\eta = 2$.

UCI Datasets

UCI datasets (Frank and Asuncion 2010) contain real data from different areas and are widely used in machine learning. To evaluate clustering methods, we select 7 popular UCI datasets where the number of classes varies from 2 to 6, and the number of samples of each class varies from 14 to 673.

As shown in Table (1), K-means performs well on *wine* and *Iris*. Spectral Clustering and LG-MMC work the best on *Iris* and *Sonar* respectively, while DIFFRAC outperforms K-means and Spectral Clustering on *SpHeart*. At the same time, our method works consistently well on these data and produces the best results in 5 out of 7 datasets.

MNIST Digits

MNIST (LeCun et al. 1998) is a handwritten digits database widely used in clustering. It comprises 60000 samples for

Data	#class	K-means	Spectral	DIFFRAC	Ours
Faces group 1	5	**68.49 ± 8.87**	51.11 ± 1.51	61.71 ± 0.00	64.05 ± 3.45
Faces group 2	5	42.30 ± 3.31	34.71 ± 0.15	31.14 ± 0.00	**44.37 ± 3.02**
Faces group 3	5	60.26 ± 8.47	46.09 ± 0.14	59.14 ± 0.00	**67.57 ± 4.23**
Faces group 4	5	66.71 ± 6.81	46.23 ± 1.74	61.71 ± 0.00	**69.18 ± 3.12**
Faces group 5	5	33.03 ± 4.79	29.14 ± 0.00	27.71 ± 0.00	**41.00 ± 2.26**
Faces group 6	5	55.49 ± 6.02	40.43 ± 4.37	58.00 ± 0.00	**59.11 ± 3.91**
Faces group 7	5	61.20 ± 4.60	38.97 ± 0.87	**68.86 ± 0.00**	65.43 ± 2.90
Faces group 8	5	63.49 ± 8.12	52.97 ± 0.15	65.71 ± 0.00	**72.46 ± 3.79**
Faces group 9	5	68.86 ± 5.45	51.11 ± 0.82	64.29 ± 0.00	**71.40 ± 4.21**
Faces group 10	5	**68.54 ± 4.74**	35.71 ± 2.72	65.14 ± 0.00	66.51 ± 2.24
Faces group 11	7	70.67 ± 6.73	60.14 ± 0.14	**76.73 ± 0.00**	71.22 ± 4.39
Faces group 12	7	61.43 ± 4.25	49.96 ± 2.20	63.27 ± 0.00	**69.53 ± 3.44**
Faces group 13	7	38.43 ± 5.97	37.76 ± 0.14	39.59 ± 0.00	**39.71 ± 3.21**
Faces group 14	7	44.49 ± 4.04	31.34 ± 0.10	43.88 ± 0.00	**46.13 ± 3.46**
Faces group 15	7	57.63 ± 4.66	31.84 ± 0.69	**57.96 ± 0.00**	56.40 ± 2.14
Faces group 16	7	55.22 ± 3.66	42.22 ± 1.55	51.02 ± 0.00	**57.06 ± 1.32**
Faces group 17	7	62.16 ± 4.21	49.59 ± 3.47	58.98 ± 0.00	**67.32 ± 3.01**
Faces group 18	7	55.30 ± 5.24	41.78 ± 2.39	**70.20 ± 0.00**	60.02 ± 3.61
Faces group 19	7	41.57 ± 4.10	32.92 ± 0.97	**56.12 ± 0.00**	54.64 ± 2.02
Faces group 20	7	43.47 ± 2.89	32.76 ± 0.38	41.84 ± 0.00	**44.07 ± 1.24**

Table 3: Summary of clustering accuracy (%±std) on Multi-PIE face dataset.

digits 0-9, about 6000 samples for each digit. Each sample is a 28×28 image and here reduced to a 50-dimension vector using PCA. Similar with (Zhang, Tsang, and Kwok 2009), we randomly select 350 samples for each digit and form a subset for evaluation. As claimed in (Zhang, Tsang, and Kwok 2009; Li et al. 2009), *Digits* 3-8, 1-7, 2-7, 8-9, are the most challenging pairs. We adopt these pairs and further construct four triplets, *Digits* 3-8-9,1-2-7,1-8-7,3-2-8, based on the above pairs for multi-class clustering. There triplets would be more challenging than pairs.

Experimental results are summarized in Table 2. Spectral clustering and our method exhibit satisfying performance. It is noteworthy that the performance of IterSVR and LG-MMC is close to that of DIFFRAC. Particularly, K-means works fairly well on triplets 1-2-7, 3-2-8, 1-8-7, and 3-8-9.

Face Clustering

The face clustering experiment is conducted on Multi-PIE dataset (Gross et al. 2010). We use the data of session #1 which contains 249 individuals. Each individual has 70 samples from 10 light conditions and 7 poses. After face images are normalized to 100×100, LBP features (Ahonen, Hadid, and Pietikäinen 2004) are extracted and then reduced to 50-dimension vectors using PCA. The clustering task here seeks to group face images from the identical individual into the same cluster (Xu et al. 2004; Xu 2005). From 249 individuals, we randomly select 10 groups, each containing 5-*individuals* and other 10 groups, each containing 7-*individuals*. For each individual, all 70 samples are used.

The accuracy results of face clustering are reported in Table (3). DIFFRAC outperforms others on 5 out of 20 groups, while our method shows the best performance on 13 out of 20 groups. As to Spectral Clustering, its performance seems not as satisfying as that in UCI and digit datasets. K-means works well on several datasets especially in *group* 2 and *group* 7 but with a much larger variance. The variances(or standard deviations) of K-means, Spectral Clustering, and our method in this experiment are much larger than those in UCI and Digits experiments. This might be due to the fact that face data here are taken from different poses and light conditions, where very large intra-class variances exist.

Results Analysis

To conclude this section, K-means performs well on average but with relatively large variances. The evaluated discriminative clustering methods, such as Spectral clustering, iterSVR, LG-MMC and DIFFRAC, accomplish reasonable accuracy with smaller variances. However, their performance seems not consistently advantageous over others on various datasets. Our method works averagely more stable and is fairly adaptive to different data and problems, with the support of the generative models and discriminative principles. Our experiments also indicate that the way to combine generative and discriminative models does inherit the profit of the two paradigms by considering data distribution via generative models and maximizing the discrimination capability via the discriminative principle.

Conclusion

We have proposed a probabilistic framework for discriminative clustering embedding generative feature mapping. The generative feature mapping is derived from linearizing Bayesian classifiers that efficiently map nonlinear boundaries to linear ones. This enables us to seek for nonlinear boundaries among clusters in the data space via computing linear boundaries in the feature space. The complete frame-

work is formulated under the principle of maximum entropy, which can be understood as the KL-projection onto the constrained posterior space and can be efficiently solved using standard techniques. In experiments, coupled with the Gaussian Mixture Model, our approach not only exhibits highly competitive performance on average but also shows consistent advantages across datasets.

Acknowledgment

This work is supported by office of Naval Research Award N000140910099 and NSF CAREER award IIS-0844566 and by a grant from the Research Grants Council of the Hong Kong Special Administrative Region (Project No. 412911).

References

Ahonen, T.; Hadid, A.; and Pietikäinen, M. 2004. Face recognition with local binary patterns. *ECCV*.

Bach, F., and Harchaoui, Z. 2007. Diffrac: a discriminative and flexible framework for clustering. *Annual Conference on Neural Information Processing Systems (NIPS)*.

Bertsekas, D. 1999. *Nonlinear programming*. Athena Scientific.

Bishop. 2006. *Pattern recognition and machine learning*, volume 4. springer New York.

Chen, W.-Y.; Song, Y.; Bai, H.; Lin, C.-J.; and Chang, E. Y. 2011. Parallel spectral clustering in distributed systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 33(3):568–586.

Csiszar, I. 1975. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability* 146–158.

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the em algorithm. *J. Royal Statist. Soc. Series B* 39(1):1–38.

Frank, A., and Asuncion, A. 2010. UCI machine learning repository.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2008. *The Elements of Statistical Learning*. Spinger.

Gomes, R.; Krause, A.; and Perona, P. 2010. Discriminative clustering by regularized information maximization. In *Annual Conference on Neural Information Processing Systems (NIPS)*.

Graça, J.; Ganchev, K.; and Taskar, B. 2007. Expectation maximization and posterior constraints. In *Annual Conference on Neural Information Processing Systems (NIPS)*.

Gross, R.; Matthews, I.; Cohn, J.; Kanade, T.; and Baker, S. 2010. Multi-pie. *Image and Vision Computing* 28(5):807–813.

Jaakkola, T., and Haussler, D. 1999. Exploiting generative models in discriminative classifiers. In *Annual Conference on Neural Information Processing Systems (NIPS)*.

Jaakkola, T.; Meila, M.; and Jebara, T. 1999. Maximum entropy discrimination. In *In Advances in Neural Information Processing Systems 12*.

Jordan, M.; Ghahramani, Z.; Jaakkola, T.; and Saul, L. 1999. An introduction to variational methods for graphical models. *Machine learning* 37(2):183–233.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11):2278–2324.

Li, Y.; Tsang, I.; Kwok, J.; and Zhou, Z. 2009. Tighter convex maximum margin clustering. In *International Conference on Artificial Intelligence and Statistics(AISTATS)*.

Li, X.; Lee, T.; and Liu, Y. 2011. Hybrid generative-discriminative classification using posterior divergence. In *CVPR*.

Li, X.; Lee, T.; and Liu, Y. 2012. Stochastic feature mapping for pac-bayes classification. *arXiv:1204.2609*.

Neal, R., and Hinton, G. 1998. A view of the em algorithm that justifies incremental, sparse, and other variants. *NATO ASI SERIES D BEHAVIOURAL AND SOCIAL SCIENCES* 89:355–370.

Ng, A.; Jordan, M.; and Weiss, Y. 2001. On spectral clustering: Analysis and an algorithm. *Annual Conference on Neural Information Processing Systems (NIPS)*.

Perina, A.; Cristani, M.; Castellani, U.; Murino, V.; and Jojic, N. 2009. Free energy score space. In *Annual Conference on Neural Information Processing Systems (NIPS)*.

Shi, J., and Malik, J. 1997. Normalized cuts and image segmentation. *IEEE Trans. on PAMI* 22:888–905.

Tu, Z. 2007. Learning generative models via discriminative approaches. In *CVPR*.

Vapnik, V. 2000. *The nature of statistical learning theory*. Springer Verlag.

Wu, M., and Scholkopf, B. 2007. A local learning approach for clustering. *Annual Conference on Neural Information Processing Systems (NIPS)*.

Xu, L.; Neufeld, J.; Larson, B.; and Schuurmans, D. 2004. Maximum margin clustering. *Annual Conference on Neural Information Processing Systems (NIPS)*.

Xu, L. 2005. Unsupervised and semi-supervised multi-class support vector machines. In *AAAI Conference on Artificial Intelligence(AAAI)*.

Zhang, K.; Tsang, I.; and Kwok, J. 2009. Maximum margin clustering made practical. *IEEE Transactions on Neural Networks* 20(4):583–596.