DISCRIMINATIVE EXEMPLAR CLUSTERING

Yingzhen Yang[1], *Feng Liang*[2], *Thomas S. Huang*[1]

Department of Electrical and Computer Engineering[1], Department of Statistics[2]
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA

ABSTRACT

Exemplar-based clustering methods partition the data space and identify the representative, or the exemplar, of each cluster. With the number of clusters adaptively determined, exemplar-based clustering methods are appealing since they avoid or alleviate the difficult task of estimating the latent parameters in case of complex models and high dimensionality of the data. Most exemplar-based clustering methods are based on generative models, where the exemplars serve as the parameters of the generative models. However, generative models do not consider the discriminative capability of the cluster boundaries explicitly described in discriminative models. In this paper, we present Discriminative Exemplar Clustering (DEC), that improves the discriminative power of exemplar-based clustering method by minimizing the misclassification error of the nonparametric unsupervised plug-in classifier while maintaining the appealing property of exemplar-based clustering. The optimization of DEC is performed in a pairwise Markov Random Field. Experimental results on synthetic and real data demonstrate the effectiveness of our method compared to other exemplar-based clustering methods.

Index Terms— Exemplar-based Clustering, Pairwise Markov Random Fields

1. INTRODUCTION

Clustering is an important data analysis method which partitions data space into a set of self-similar clusters. Exemplar-based clustering methods, mostly based on generative models, identify the representative, or the exemplar of each cluster while partitioning the data. The exemplars always serve as parameters of the parametric distributions, e.g. the means of the mixture components. Therefore, exemplar-based clustering methods are appealing since they either avoid the difficult task of estimating the latent parameters of generative models in case of complicated parametric model and high dimensionality of the data, such as Affinity Propagation [1] and its succeeding algorithms [2, 3]; or alleviate this problem by only estimating the mixture coefficients of the mixture models such as [4] and its accelerated version [5].

On the other hand, discriminative clustering methods explicitly search for the cluster boundaries for optimal data partition, and many of them seek for the cluster boundaries by explicitly learning a classifier from unlabeled data. [6] learns a max-margin two-class classifier in an unsupervised manner, and their method is known as unsupervised SVM whose theoretical property is further analyzed in [7]. Also, [8] and [9] learn the kernelized Gaussian classifier and the kernel logistic regression classifier respectively. and adopt the entropy of the posterior distribution of the class label by the classifier to measure the quality of the learned classifier.

[10] shows the effectiveness of combining generative and discriminative models in classification tasks, and recent work [11] further demonstrates the convincing performance of coupled generative and discriminative models for clustering. Most exemplar-based clustering methods innately lacks the advantages of discriminative models which explicitly maximize the gap between different clusters, so that we propose to enhance the exemplar-based clustering with discrimination capability aiming to improve the clustering performance. To achieve this goal, we formulate a novel discriminative clustering model by minimizing the misclassification error of unsupervised classification (MEUC) using the plug-in classifier. MEUC is incorporated into the exemplar-based clustering scheme to form a new clustering algorithm called Discriminative Exemplar Clustering (DEC), while maintaining the appealing property of exemplar-based clustering.

2. THE MODEL

Let (X, Y) be a random couple with joint distribution P_{XY}, where $X \in \mathcal{X} \subset \mathbb{R}^d$ is a vector of d features and $Y \in \{1, 2, ..., Q\}$ is a label indicating the class to which X belongs. Q is finite and can be unknown. The sample $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are independent copies of (X, Y), $\{y_l\}_{l=1}^n$ are missing, and we aim to perform clustering on observed data $\{\mathbf{x}_l\}_{l=1}^n$.

This research is supported in part by ONR Grant N00014-12-1-0122.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

2.1. Exemplar-Based Clustering

In this subsection we introduce Affinity Propagation (AP) [1], which is a representative of nonparametric exemplar-based clustering methods. In AP, each $\mathbf{x}_l$ is associated with a cluster indicator e_l ($l \in \{1, 2, ...n\}, e_l \in \{1, 2, ...n\}$), indicating that $\mathbf{x}_l$ takes $\mathbf{x}_{e_l}$ as the cluster exemplar. Data from the same cluster share the same cluster exemplar. We define $\mathbf{e} \triangleq \{e_l\}_{l=1}^n$. Moreover, a configuration of the cluster indicators $\mathbf{e}$ is consistent iff $e_l = l$ when $e_m = l$ for any $l, m \in 1..n$, meaning that $\mathbf{x}_l$ should take itself as its exemplar if any $\mathbf{x}_m$ take $\mathbf{x}_l$ as its exemplar. It is required that the cluster indicators $\mathbf{e}$ should always be consistent.

Affinity Propagation [1], a representative of the exemplar-based clustering methods, solves the following optimization problem

$$\min_{\mathbf{e}} \sum_{l=1}^{n} S_{l,e_l} \quad s.t. \quad \mathbf{e} \text{ is consistent} \tag{1}$$

S_{l,e_l} is the dissimilarity between x_l and x_{e_l}, and note that $S_{l,l}$ is set to be nonzero to avoid the trivial minimizer of (1). The objective function (1) ensembles that of K-means clustering, and AP is based on generative models. In order to improve the discriminative power of such exemplar-based clustering, we propose a novel discriminative clustering model by minimizing the misclassification error of unsupervised classification (MEUC) using the plug-in classifier below.

2.2. Discriminative Clustering by Minimizing the Misclassification Error of Unsupervised Classification (MEUC)

2.2.1. Learning Unsupervised Classifier from Unlabeled Data

The training scheme via hypothetical labeling introduced by the unsupervised SVM [6, 7] forms the basis for learning a classifier from unlabeled data in a principled way. With any hypothetical labeling $\hat{\mathbf{y}} = \{\hat{y}_l\}_{l=1}^n$, we can build the corresponding training data $S_{\hat{\mathbf{y}}} = \{\mathbf{x}_l, \hat{y}_l\}$ for a potential classifier. Let $S_{\hat{\mathbf{y}},i} = \{\mathbf{x}_l : \hat{y}_l = i, 1 \le l \le n\}$ be the data with label i, then $\{S_{\hat{\mathbf{y}},i}\}_{i=1}^Q$ is a partition of the data, and two labelings are equivalent in the sense of clustering if they produce the same data partition. In this way, the quality of a labeling $\hat{\mathbf{y}}$, or equivalently a data partition, can be evaluated by the misclassification error of the classifier learned from the corresponding training data $S_{\hat{\mathbf{y}}}$. Unsupervised SVM [6, 7] perform clustering by searching for the hypothetical labeling with minimum associated misclassification error of SVM. We use the same training scheme as unsupervised SVM to learn the unsupervised plug-in classifier, aiming to find the hypothetical labeling with minimum misclassification error of the plug-in classifier. Note that it is difficult to adapt unsupervised SVM and other popular discriminative clustering methods based on information maximization [8, 9, 12] to the exemplar-based clustering scheme since all such methods estimate the parameters of the classifier by continuous optimization without exemplar finding. In contrast, the misclassification error bound for the unsupervised plug-in classifier is expressed in terms of pairwise similarities between data points, which can be straightforwardly incorporated into the nonparametric exemplar-based clustering scheme.

2.2.2. Objective Function of MEUC

By the training scheme for unsupervised classifier, the misclassification error (or the generalization error) of the unsupervised classifier $F_{S_{\hat{\mathbf{y}}}}$ learned from the training data $S_{\hat{\mathbf{y}}}$ is:

$$R\left(F_{S_{\hat{\mathbf{y}}}}\right) \triangleq \Pr\left[(X,Y) : F_{S_{\hat{\mathbf{y}}}}(X) \neq Y\right] \tag{2}$$

$F_{S_{\hat{\mathbf{y}}}}(X)$ is the classification function which returns the class label of a sample X. In this paper we investigate the case when $F_{S_{\hat{\mathbf{y}}}}$ is the plug-in classifier $\mathrm{PI}_{S_{\hat{\mathbf{y}}}}$:

$$\mathrm{PI}_{S_{\hat{\mathbf{y}}}}(X) = \arg\max_{1 \le i \le Q} \hat{\eta}_{n,h_n}^{(i)}(X) \tag{3}$$

where $\hat{\eta}_{n,h_n}^{(i)}$ is the nonparametric kernel estimator of the regression function $\eta^{(i)}$. To evaluate the quality of a given hypothetical labeling $\hat{\mathbf{y}}$, we assume that $\{\hat{y}_l\}$ is the missing latent labeling, i.e. $\{\hat{y}_l\} = \{y_l\}$, and then estimate (2) with respect to a collection of possible joint distributions P_{XY}. Before stating the generalization error bound theorem, we introduce notations and assumptions used in the new discriminative clustering model. Suppose P_X is the induced marginal distribution of X, and f is the probabilistic density function of P_X which is a mixture of class-conditional densities. $\left(f^{(i)}, \pi^{(i)}\right)$ is the class-conditional density function and the prior of class i ($f = \sum_{i=1}^{Q} \pi^{(i)} f^{(i)}$). $\eta^{(i)}(x)$ is the regression function of Y on $X = x$, i.e. $\eta^{(i)}(x) = \Pr[Y = i \,|X = x] = \frac{\pi^{(i)} f^{(i)}(x)}{f(x)}$. For the sake of the consistency of the kernel density estimators, there are further assumptions on the marginal density and class-conditional densities:

(A) f is bounded, i.e. $0 < f_{\min} \le f \le f_{\max}$, and $f \in \Sigma_{\gamma,c}$

(B) $\{f^{(i)}\}$ is bounded, i.e. $0 < f_{\min}^{(i)} \le f^{(i)} \le f_{\max}^{(i)}$, and $f^{(i)} \in \Sigma_{\gamma,c_i}$, $1 \le i \le Q$. where $\Sigma_{\gamma,c}$ is the class of Hölder-γ smooth functions with Hölder constant c:

$$\Sigma_{\gamma,c} \triangleq \{f : \mathbb{R}^d \to \mathbb{R} \mid \ |f(x) - f(y)| \le c\|x - y\|^{\gamma}\}, \gamma > 0$$

and we denote by $\mathcal{P}_X$ the the collection of marginal distributions that satisfy assumption (A), and denote by $\mathcal{P}_{X|Y}$ the collection of class-conditional distributions that satisfy assumption (B). We then define the collection of joint distributions $\mathcal{P}_{XY,\hat{\mathbf{y}}}$ that P_{XY} belongs to, which specifies the prior of the classes and requires $\{\mathbf{x}_l\}$ be generated according to P_X, $S_{\hat{\mathbf{y}},i}$ be generated according to $P_{X|Y=i}$ and the marginal density and class-conditional densities satisfy assumption (A)-(B):

$$\mathcal{P}_{XY,\hat{\mathbf{y}}} \triangleq \{P_{XY} \mid P_X \in \mathcal{P}_X, \{P_{X|Y=i}\} \in \mathcal{P}_{X|Y}, \{\mathbf{x}_l\} \overset{i.i.d.}{\sim} P_X, \\ S_{\hat{\mathbf{y}},i} \overset{i.i.d.}{\sim} P_{X|Y=i} \text{ and } \pi^{(i)} = \tfrac{|S_{\hat{\mathbf{y}},i}|}{n} \text{ for } 1 \le i \le Q\} \tag{4}$$

The objective of our discriminative clustering model is to find the optimal labeling $\hat{\mathbf{y}}$ that minimizes the supremum of the associated misclassification error over the collection of the joint distributions $\mathcal{P}_{XY,\hat{\mathbf{y}}}$:

$$\min_{\hat{\mathbf{y}}\in\hat{\mathcal{Y}}} \sup_{\mathcal{P}_{XY,\hat{\mathbf{y}}}} R\left(F_{S_{\hat{\mathbf{y}}}}\right) \tag{5}$$

where $\hat{\mathcal{Y}} = \{\hat{\mathbf{y}} \mid \frac{\#\{\hat{y}_l=i\}}{n} \geq \pi^{(0)}, 1 \leq i \leq Q\}$ is the set of balanced labelings (avoiding cluster imbalance), $\pi^{(0)}$ is a positive constant. Theorem 1 gives the tight generalization bound for the error of the unsupervised plug-in classifier:

Theorem 1. *(Generalization Error of the Plug-In Classifier) Let the kernel bandwidth be $h_n = \Theta(n^{-\frac{(1-\varepsilon)}{d+2\gamma}})$ for any $\varepsilon \in (0,1)$. With probability greater than $1 - QLh_n^E$, the generalization error of the plug-in classifier (3) satisfies*

$$\sup_{\mathcal{P}_{XY}} R\left(\mathrm{PI}_S\right) \leq \frac{1}{n^2}\sum_{l,m}\theta_{lm}G_{lm,h_n} + \mathcal{O}\left(n^{-\frac{(1-\varepsilon)\gamma}{d+2\gamma}}\right) \tag{6}$$

where $\theta_{lm} = \mathbb{1}_{\{\mathbf{y}_l\neq\mathbf{y}_m\}}$ is a class indicator function and

$$G_{lm,h_n} = G_{h_n}\left(\mathbf{x}_l,\mathbf{x}_m\right),\ G_h\left(x,y\right) = \frac{K_h\left(x-y\right)}{\hat{f}_{n,h}^{\frac{1}{2}}\left(x\right)\hat{f}_{n,h}^{\frac{1}{2}}\left(y\right)}, \tag{7}$$

L, E are constants, $\hat{f}_{n,h_n}$ is the kernel density estimator of f:

$$\hat{f}_{n,h}\left(x\right) = \frac{1}{n}\sum_{l=1}^{n} K_h\left(x-\mathbf{x}_l\right), \tag{8}$$

where $K_h(\cdot)$ is the isotropic Gaussian kernel with bandwidth h: $K_h\left(x\right) = \frac{1}{2\pi^{d/2}h^d}e^{-\frac{\|x\|^2}{2h^2}}$, $\hat{\eta}_{n,h_n}$ in (3) is $\hat{\eta}_{n,h_n}^{(i)}\left(x\right) = \frac{\sum_{l=1}^{n} K_{h_n}\left(x-\mathbf{x}_l\right)\mathbb{1}_{\{\mathbf{y}_l=i\}}}{n\hat{f}_{n,h_n}\left(x\right)}$, and the bound in (6) is tight.

Based on Theorem 1, by omitting the scalar $\frac{1}{n^2}$ and the term $\mathcal{O}\left(n^{-\frac{(1-\varepsilon)\gamma}{d+2\gamma}}\right)$ that is infinitesimally small with sufficiently large n, we relax the objective of MEUC (5) to

$$\min_{\hat{\mathbf{y}}\in\hat{\mathcal{Y}}} \sum_{l,m}\theta_{lm}G_{lm,h_n} \tag{9}$$

G_{lm,h_n} can be interpreted as the similarity between x_l and x_m, and the optimization of (9) encourages minimum sum of similarity between data points from different clusters.

2.3. Discriminative Exemplar Clustering

In this section we formulate Discriminative Exemplar Clustering (DEC), which improves the discriminative capability of exemplar-based clustering (Section 2.1) by incorporating the discriminative clustering model MEUC which minimizes the misclassification error of the unsupervised classification using the plug-in classifier (Section 2.2). Since $\sum_{l,m}\theta_{lm}G_{lm,h_n}$ in (9) can be rewritten as $\sum_{l,m}\tilde{\theta}_{lm}G_{lm,h_n}$ where $\tilde{\theta}_{lm} = \mathbb{1}_{\{e_l\neq e_m\}}$ and $\{e_l\}$ are the cluster indicators, it is straightforward to combine MEUC (9) and exemplar-based clustering (1) as

$$\min_{\hat{\mathbf{y}}\in\hat{\mathcal{Y}}} \sum_{l=1}^{n} S_{l,e_l} + \lambda\sum_{l,m}\left(\tilde{\theta}_{lm}G_{lm,h_n}\right) \quad s.t. \quad \mathbf{e} \text{ is consistent}$$

To avoid imbalanced data partition, we use the within-cluster sum of dissimilarities to control the size of clusters by setting $S_{l,e_l} = \exp\left(-G_{le_l,h_n}\right)$. Moreover, to ensure consistent cluster indicators, we design the penalty function ρ_{lm}:

$$\rho_{lm}\left(e_l,e_m\right) = \begin{cases} \infty & e_m = l, e_l \neq l \text{ or } e_l = m, e_m \neq m \\ 0 & \text{otherwise} \end{cases}$$

DEC minimizes the following relaxed objective function

$$\Psi\left(e\right) = \sum_{l=1}^{n} S_{l,e_l} + \lambda\sum_{l,m}\left(\tilde{\theta}_{lm}G_{lm,h_n} + \rho_{lm}\left(e_l,e_m\right)\right) \tag{10}$$

where λ is a balancing parameter. Due to the form of (10), we construct a pairwise Markov Random Field (MRF) representing the unary term u_l and the pairwise term $\tilde{\theta}_{lm}G_{lm,h_n} + \rho_{lm}$ as the data likelihood and prior respectively. The variables e are modeled as nodes and the unary term and pairwise term in (10) are modeled as potential functions in the pairwise MRF. The minimization of the objective function is then converted to a MAP (Maximum a Posterior) problem in the pairwise MRF. (10) is minimized by Max-Product Belief Propagation (BP) [13] in two steps:

Message Passing: BP iteratively passes messages along each edge according to

$$m_{lm}^t\left(e_m\right) = \min_{e_l}\left(M_{lm}^{t-1}\left(e_l\right) + \tilde{\theta}_{lm}G_{lm,h_n} + \rho_{lm}\left(e_l,e_m\right)\right) \tag{11}$$

$$M_{lm}^t\left(e_l\right) \triangleq \sum_{k\in\mathcal{N}(l)\setminus m} m_{kl}^t\left(e_l\right) + u_l\left(e_l\right) \tag{12}$$

where m_{lm}^t is the message sent from node l to node m in iteration t, $\mathcal{N}\left(l\right)$ is the set of neighbors of node l.

Inferring the optimal label: After the message passing converges or the maximal number of iterations is achieved, the final belief for each node is $b_l\left(e_l\right) = \sum_{k\in\mathcal{N}(l)} m_{kl}^T\left(e_l\right) + u_l\left(e_l\right)$, T is the number of iterations of message passing. The resultant optimal e_l^* is $e_l^* = \arg\min_{e_l} b_l\left(e_l\right)$.

3. EXPERIMENTAL RESULTS

We demonstrate the performance of DEC on synthetic and real data sets in this section. The default value for the kernel bandwidth h_n in (10) is h_n^*, which is set as the variance of the pairwise distance between data points $\left\{\|x_i - x_j\|_{i<j}\right\}$, and the default value for the balancing parameter λ in the objective function (10) is 1. DEC produces different number of clusters by varying both λ and h_n. We let $h_n = \alpha h_n^*$, where α is called the bandwidth ratio controlling the kernel bandwidth.

λ varies between $[0.2, 1]$ and the bandwidth ratio α varies between $[0.2, 1.9]$ with step 0.2 and 0.05 respectively. We use fixed parameter setting throughout all the experiments. For the clustering methods that depend on random initialization such as K-means, we run them 30 times and report the average performance.

Data Sets

We conduct experiments on two synthetic data sets and four real data sets. For both synthetic data sets, 300 points are randomly generated in R^2 whose distribution is a mixture of 5 Gaussians with equal weight and different scales. In the first data set, the means and covariance matrices for the Gaussian components are randomly generated. The means for the 5 Gaussian components are generated from $\mathcal{N}\left(\left(-6 \quad -6\right), \mathrm{I}\right)$, $\mathcal{N}\left(\left(-6 \quad 6\right), \mathrm{I}\right), \mathcal{N}\left(\left(6 \quad 6\right), \mathrm{I}\right), \mathcal{N}\left(\left(6 \quad -6\right), \mathrm{I}\right)$ and $\mathcal{N}\left(\left(0 \quad 0\right), \mathrm{I}\right)$ respectively. The covariance matrices for the first four Gaussian components are generated from $W\left(\mathrm{I}, 2\right)$ and that for the last Gaussian component is generated from $W\left(2\mathrm{I}, 2\right)$, where $W\left(\Sigma, d\right)$ indicates Wishart distribution with covariance matrix Σ and d is the degree of freedom. The second data set is almost the same as the first one except that the covariance matrices for the last two Gaussian components are generated from $W\left(2\mathrm{I}, 2\right)$. We choose four real data sets from UCI repository [14], i.e. Iris, Parkinsons, Vertebral Column (VC), and Breast Tissue (BT). We use the popular adjusted rand index (ARI) [15] for evaluating the performance of the clustering methods. ARI has been widely used as a measure of agreement between the inferred cluster labels and the ground truth cluster assignments. It ranges from -1 to 1, and achieves the maximum 1 when the inferred label is identical to the ground truth.

Clustering Results

We compare DEC to K-means, spectral clustering (SC) [16], Gaussian Mixture Model (GMM) [17], Affinity Propagation (AP) [1], Convex Clustering with Exemplar-Based Model (CEB) [4] for the task of clustering on the synthetic data. We report the same experimental results for the two synthetic data sets, Parkinsons, VC and BT data sets as in [18], which is a preliminary version of this paper. To reveal the effectiveness of default parameters, we require all the three exemplar-based clustering methods, i.e. AP, CEB, DEC, run with default parameters. The clustering results in terms of ARI on the two synthetic data sets are reported in in Table 1(a). We report the average ARI (Avg ARI) and the standard deviation (SD) of ARI for all the clustering methods. T is the number of times when exemplar-based clustering methods choose the correct cluster number, AC is the average number of clusters they produce on each data set. We observe that DEC achieves the highest average ARI, and the number of times it chooses the correct cluster number is more than other exemplar-based clustering methods. DEC outperforms AP and CEB by virtue of its discriminative capability from the MEUC model and modeling the data distribution more accurately by kernel density estimation. We also observe that CEB always performs better than GMM since it finds the global minimum of a convex likelihood function of a mixture model. AP tends to split the data into many small clusters by its default setting, and it produces the largest average number of clusters. Moreover, we compare DEC to representative exemplar-based clustering methods, i.e. AP and CEB, for clustering on real data sets. All the exemplar-based clustering methods produces different cluster numbers by varying their parameters. AP controls the number of clusters by a parameter called preference. We first estimate the lower bound and upper bound for the preference using routine functions provided by the authors, then evenly sample 170 (the number of parameter settings for DEC) preference values between its upper bound and lower bound, and then run AP with each sampled preference value. CEB partitions the data by varying the scale β which controls the shape of the mixture components. Likewise, we evenly sample 170 values between $[0.1, 3]$ for β. we record the average ARI and the standard deviation of ARI for all the exemplar-based clustering methods when they produce the correct number of clusters for each data set (shown in Table 1(b)). Coupled with the discriminative clustering model MEUC by minimizing the misclassification error of the unsupervised plug-in classifier, DEC combines the advantages of both exemplar-based clustering and discriminative clustering to achieve better performance.

Table 1. Clustering Results

(a) Clustering on the synthetic data

Data		K-means	SC	GMM	AP	CEB	DEC
1	Avg ARI	0.8625	0.8917	0.9143	0.7548	0.9215	**0.9421**
	SD	0.0445	0.0341	0.0528	0.0992	0.0471	**0.0317**
	T(AC)	-	-	-	0(8.7)	7(5.8)	**8**(4.5)
2	Avg ARI	0.8472	0.8116	0.8874	0.6996	0.9224	**0.9389**
	SD	0.0598	0.0304	0.0625	0.0882	0.0475	0.0383
	T(AC)	-	-	-	0(9.3)	6(6)	**8**(4.5)

(b) Clustering on the real data

Methods	Iris	Parkinsons	VC	BT
AP	0.7297 ± 0.0301	0.0626 ± 0	0.2937 ± 0	0.2691 ± 0
CEB	0.5432 ± 0.0199	-	-0.0059 ± 0.0012	0.0796 ± 0
DEC	$\mathbf{0.7619} \pm 0.0068$	$\mathbf{0.1595} \pm 0.2011$	$\mathbf{0.3380} \pm 0.0292$	$\mathbf{0.4618} \pm 0.0104$

4. CONCLUSION

We propose Discriminative Exemplar Clustering to improve the discrimination capability of the exemplar-based clustering scheme, which employs a novel discriminative clustering model by minimizing the misclassification error of the unsupervised plug-in classifier. Coupling the discriminative clustering model MEUC and the exemplar-based clustering scheme, we build the objective function of DEC and optimize it in a Pairwise MRF. Experimental results show that DEC compares favorably to other exemplar-based clustering methods on synthetic and real data sets.

5. REFERENCES

[1] Brendan J. Frey and Delbert Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, pp. 972–977, 2007.

[2] Inmar E. Givoni and Brendan J. Frey, "A binary variable model for affinity propagation," *Neural Computation*, vol. 21, no. 6, pp. 1589–1600, 2009.

[3] Inmar E. Givoni and Brendan J. Frey, "Semi-supervised affinity propagation with instance-level constraints," *Journal of Machine Learning Research - Proceedings Track*, vol. 5, pp. 161–168, 2009.

[4] Danial Lashkari and Polina Golland, "Convex clustering with exemplar-based models," in *NIPS*, 2007.

[5] Rikiya Takahashi, "Sequential minimal optimization in adaptive-bandwidth convex clustering," in *SDM*, 2011, pp. 896–907.

[6] Linli Xu, James Neufeld, Bryce Larson, and Dale Schuurmans, "Maximum margin clustering," in *NIPS*, 2004.

[7] Zohar Karnin, Edo Liberty, Shachar Lovett, Roy Schwartz, and Omri Weinstein, "Unsupervised svms: On the complexity of the furthest hyperplane problem," *Journal of Machine Learning Research - Proceedings Track*, vol. 23, pp. 2.1–2.17, 2012.

[8] Felix V. Agakov and David Barber, "Kernelized infomax clustering," in *NIPS*, 2005.

[9] Ryan Gomes, Andreas Krause, and Pietro Perona, "Discriminative clustering by regularized information maximization," in *NIPS*, 2010, pp. 775–783.

[10] Tommi Jaakkola and David Haussler, "Exploiting generative models in discriminative classifiers," in *NIPS*, 1998, pp. 487–493.

[11] Liwei Wang, Xiong Li, Zhuowen Tu, and Jiaya Jia, "Discriminative clustering via generative feature mapping," in *AAAI*, 2012.

[12] Masashi Sugiyama, Makoto Yamada, Manabu Kimura, and Hirotaka Hachiya, "On information-maximization clustering: Tuning parameter selection and analytic solution," in *ICML*, 2011, pp. 65–72.

[13] Yair Weiss and William T. Freeman, "On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 736–744, 2001.

[14] D.J. Newman A. Asuncion, "UCI machine learning repository," 2007.

[15] L. Hubert and P. Arabie, "Comparing Partitions," *Journal of Classification*, 1985.

[16] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss, "On spectral clustering: Analysis and an algorithm," in *NIPS*, 2001, pp. 849–856.

[17] Chris Fraley and Adrian E. Raftery, "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 611–631, June 2002.

[18] Yingzhen Yang, Xinqi Chu, and Thomas S. Huang, "Nonparametric pairwise similarity for discriminative clustering," in *NIPS 2013 Workshop on Modern Nonparametric Methods in Machine Learning,*, 2013.