JMLR: Workshop and Conference Proceedings 13: 47-62
2nd Asian Conference on Machine Learning (ACML2010), Tokyo, Japan, Nov. 8–10, 2010.

Minimum Conditional Entropy Clustering: A Discriminative Framework for Clustering

Bo Dai BDAI@NLPR.IA.AC.CN
NLPR/LIAMA
Institute of Automation, Chinese Academy of Sciences, Beijing, P.R.China

Baogang Hu HUBG@NLPR.IA.AC.CN
NLPR/LIAMA
Institute of Automation, Chinese Academy of Sciences, Beijing, P.R.China

Editor: Masashi Sugiyama and Qiang Yang

Abstract

In this paper, we introduce an assumption which makes it possible to extend the learning ability of discriminative model to unsupervised setting. We propose an *information-theoretic* framework as an implementation of the low-density separation assumption. The proposed framework provides a unified perspective of Maximum Margin Clustering (MMC), Discriminative k-means, Spectral Clustering and Unsupervised Renyi's Entropy Analysis and also leads to a novel and efficient algorithm, *Accelerated Maximum Relative Margin Clustering* (ARMC), which maximizes the margin while considering the spread of projections and affine invariance. Experimental results show that the proposed discriminative unsupervised learning method is more efficient in utilizing data and achieves the state-of-the-art or even better performance compared with mainstream clustering methods.

Keywords: Clustering, Discriminative, Information-Theoretic Framework, Low-density Separation Assumption, Accelerated Maximum Relative Margin Clustering.

1. Introduction

Clustering, as an important subject in unsupervised learning, has been studied extensively. The basic target of clustering can be informally described as determining how data are organized. Distinguished from supervised learning, the learner is given only unlabeled examples in clustering. Most generative clustering algorithms assume that data are generated from certain distributions either explicitly or implicitly. Hence these clustering algorithms label the instances by fitting such joint distribution for the unlabeled data as their underlying mechanism.

However, there are some limitations of these approaches. Generally speaking, estimation of a joint distribution is much harder and merely an intermediate step; it does not solve the problem directly, e.g., assigning cluster label to each instance. It is possible that the information provided by data is sufficient for clustering but insufficient for fitting the generative model. Mathematically speaking, given $\mathbf{x}$ representing the samples and $\mathbf{y}$ denoting the associated labels, the cluster assignments, maximizing the joint likelihood may not lead to the best performance because the involved optimization improves the fit of $p(\mathbf{x})$ rather than $P(\mathbf{y}|\mathbf{x})$, which is directly related to the final task. Moreover, the clustering results

may be incorrect when the assumption of the generative model is violated. These reasons make the algorithms rooted in generative models inefficient sometimes.

On the other hand, discriminative models, which are capable of learning the conditional distribution $P(\mathbf{y}|\mathbf{x})$ directly, avoid the disadvantages of generative models. This school of methods provide us a more powerful and efficient approach to extract useful information from data. Nevertheless, as (Seeger (2006)) pointed out, standard discriminative model cannot learn any knowledge from *unlabeled* data. Extending the learning ability of discriminative model to unsupervised setting for clustering is the motivation of our work.

In this paper, we establish a novel *information-theoretical* framework that can learn knowledge from *unlabeled* data in a *discriminative* way. Exploiting (Ben-David et al. (2009)) in which the novel low-density separators learning problem was defined and some theoretic results had been obtained, we introduce a similar generalized smoothness assumption that connects $P(\mathbf{y}|\mathbf{x})$ and $p(\mathbf{x})$ into unsupervised learning problems and propose a framework using conditional entropy to describe the clustering information. To deal with different problems, different forms of $P(\mathbf{y}|\mathbf{x})$ can be substituted into the framework straightforwardly. The proposed framework is important because it provides a unified view for understanding many existing clustering methods and a powerful tool to develop new algorithms. More specifically, we show that the well-known algorithms as Maximum Margin Clustering (MMC) (Xu et al. (2004)), Discriminative k-means (Ye et al. (2008)), Spectral Clustering (Shi and Malik (2000); Ng et al. (2001)) and Unsupervised Renyi's Entropy Analysis (Yuan and Hu (2009)) connect closely to the proposed framework. Finally, we have presented the *Accelerated Maximum Relative Margin Clustering* (ARMC) as a special case of our framework. We show that this special case is the unsupervised extension of Relative Margin Machine (RMM) (Shivaswamy and Jebara (2010)), and can be solved efficiently.

The remainder of this paper is organized as follows. In the following part, we review related work to provide a panorama and the position of our framework. The *Unsupervised Smoothness Assumption* is introduced in Section 2. We present our framework, including variations, optimization algorithm and some analyses in Section 3. Section 4 specifies *Accelerated Maximum Relative Margin Clustering*. To validate our framework and evaluate the efficiency, experimental results are reported in Section 5. The last section gives the conclusion.

1.1 Related Work

There have been a few works that developed discriminative models to extract information from unlabeled data (Chapelle et al. (2006)). Some semi-supervised learning algorithms such as semi-supervised support vector machine (Joachims (1999)), graph-based methods (Belkin et al. (2005)), extend the learning ability of discriminative model through different data dependent regularizations which are the implementations of semi-supervised smoothness assumption.

Similar with our framework, entropy regularization (Grandvalet and Bengio (2004)) uses conditional entropy in the same way. However, it just plays as a regularization and focuses on semi-supervised learning problem.

Recently, maximum margin clustering, discriminative k-means and spectral clustering are proposed to discover the underlying structure of data by utilizing the connection between

$P(\mathbf{y}|\mathbf{x})$ and $p(\mathbf{x})$ directly instead of Bayes's theorem. Discriminative k-means separates the instances and executes LDA subspace selection simultaneously. Spectral clustering can be viewed from several points, one of which is finding a labelling of instances so that to separate the data graph smoothly. Maximum margin clustering wants to label the instances so that if one runs an SVM, the margin obtained would be largest over all candidate labellings. By complicated transformations, (Xu et al. (2006)) extended MMC to multi-class and sequential data. To reduce the computational cost of MMC, many optimization methods have been proposed (Zhang et al. (2007); Zhao et al. (2008); Li et al. (2009)). We can see later that all of these three families of algorithms are the implementations of *Unsupervised Smoothness Assumption* implicitly and connect closely to our framework. As a special case of the proposed framework, Maximum Relative Margin Clustering maximizes the relative margin (Shivaswamy and Jebara (2010)) which considering the spread of projections and affine invariance besides margin.

Although there were some works extending the learning ability of discriminative model to unsupervised setting, no systematic studies have been proposed to the best of our knowledge. By connecting clustering with the novel low-density separator learning problem defined in (Ben-David et al. (2009)), we propose this framework, *Minimum Conditional Entropy Clustering*, to solidify the foundation of clustering via *discriminative* models.

1.2 Our Contributions

The main contributions of this paper are three folds:

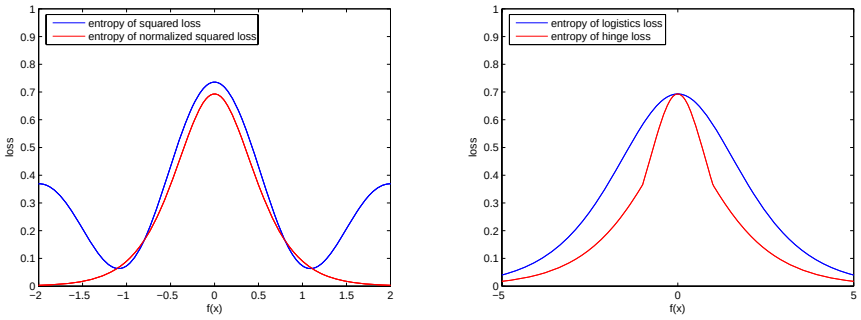
- Firstly, we introduce the *Unsupervised Smoothness Assumption* which extends the discriminative models to unsupervised setting.
- Secondly, a *information-theoretic* framework is adopted, which provides us a unified perspective to existing algorithms and a systematic tool to develop new algorithms.
- Finally, we design a new algorithm, *Accelerated Maximum Relative Margin Clustering*, as an unsupervised extension of Relative Margin Machine.

2. Unsupervised Smoothness Assumption

As we introduced in Section 1, the standard discriminative models turn blind eyes to the unlabeled data because of the independence assumption about $p(\mathbf{x})$ and $P(\mathbf{y}|\mathbf{x})$. To utilize the unlabeled data to recover a discriminative model, we should hold the other kind of assumptions which is able to connect $P(\mathbf{y}|\mathbf{x})$ with $p(\mathbf{x})$ reasonably. By connecting clustering with low-density separators learning problem, we extend the generalization of smoothness assumption (Chapelle et al. (2006)) that combines $p(\mathbf{x})$ and $P(\mathbf{y}|\mathbf{x})$ together to unsupervised learning similar to semi-supervised learning.

Unsupervised Smoothness Assumption: *If two instances* $\mathbf{x_1}$ *and* $\mathbf{x_2}$ *are close in a high-density region, then the difference between their corresponding outputs should be small with high probability.*

This assumption implies that the discriminant boundary should not be located in a high-density region. In other words, the clusters are separated through low-density re-



(a) (b)

Figure 1: Entropy loss of different forms of $P(\mathbf{y}|\mathbf{x})$

gions (similar with low-density separation). Note that this assumption can be applied to both discrete and continuous unsupervised problems.

This assumption is reasonable for clustering. Consider a dense region that contains massive instances; it seems unlikely to distinguish the instances in this region into different clusters. Even though we may face the conditions that the clusters overlap heavily, kernel method can help the assumption holding. Recently, a theoretic analysis of relationship between the clustering stability and data density along the cluster boundary has been proposed; the lower these densities the more stable the clustering (Ben-David and von Luxburg (2008)). Moreover, the existence of a universally consistent algorithm for finding the low-density separator also has been proved (Ben-David et al. (2009)). However, as semi-supervised smoothness assumption may hurt the performance, when unsupervised smoothness assumption benefits to unsupervised learning is still needed to consider carefully.

3. Minimum Conditional Entropy Clustering

In this section, we first define the clustering problem and present the notations formally. Then, we introduce the information-theoretic framework, *Minimum Conditional Entropy Clustering*, including some special variations and general optimization method. Finally, we present some analyses of the framework.

3.1 Problem Formulation

Given a training data set $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ drawn *i.i.d* according to a certain distribution $p(\mathbf{x})$. Here $\mathbf{x}_i \in \mathbb{R}^d$ is the input feature vector. The task of clustering can be informally described as to find a labeling $\mathbf{y}$ of dataset or to seek a hypothesis $P(\mathbf{y}|\mathbf{x})$ to divide the dataset so that the data in the same group (with the same label) are similar and the data in different groups (with the different labels) are dissimilar.

3.2 Framework

The framework is presented as follows. We first introduce the *Minimum Conditional Entropy Principle* as an implementation of *Unsupervised Smoothness Assumption*. Different forms

of $P(\mathbf{y}|\mathbf{x})$ for different problems are specified. Finally, we employ deterministic annealing EM as a general optimization algorithm [1].

3.2.1 Minimum Conditional Entropy Principle

To describe the unsupervised smoothness assumption, there are many criteria which can be candidates. However, we choose conditional entropy of labellings defined as (Grandvalet and Bengio (2004)):

$$H(Y|X) = -\int \sum_{c} P(\mathbf{y}_c|\mathbf{x}) \log P(\mathbf{y}_c|\mathbf{x}) \mathrm{d}P(\mathbf{x}) \tag{1}$$

to be the foundation of our framework. Proposed as a measurement of disorder from information theory, conditional entropy measures the discriminative ability of $P(\mathbf{y}|\mathbf{x})$ and is weighted by $p(\mathbf{x})$ simultaneously. When the decision function is smooth, this criterion leads the phenomenon that the decision boundary, where the posterior probability is ambiguous, locates far away from a high-density region. There are some other studies concluded that the conditional entropy is decreasing while reducing the classes overlap (Grandvalet and Bengio (2004); O'Neill (1978)). From this perspective, minimum conditional entropy principle is appropriate to reflect the unsupervised smoothness assumption. This property of conditional entropy can also be viewed from Figure.1. Although different forms of $P(\mathbf{y}|\mathbf{x})$ generate different special criteria from (1), all of these criteria push the decision boundary, where all of the posterior probabilities are equal and the differences of them $f(\mathbf{x})$ are zero, away from the dense unlabeled points. The other important property of conditional entropy is that this criterion is raised from information theory, making it natural and convenient when our discriminative model is expressed as a probabilistic model.

A problem that the optimum of the criterion may be a degenerate solution which separates all of the data into one cluster comes from the minimum conditional entropy principle. Some literatures illustrate the importance of the class-balance constraints (Xu et al. (2004); Zhang et al. (2007)), we also add a penalty $\mathscr{D}(P(\mathbf{y}|\mathbf{x}))$ to avoid this problem and keep the decision boundary smooth without loss of generality. To avoid computing integral and the assumption about $p(\mathbf{x})$ in (1), we apply the plug-in principle that replaces the integral in (1) to finite samples expectation to get an empirical estimation of conditional entropy and the criterion transforms as:

$$-\sum_{i=1}^{n}\sum_{j=1}^{c} P(\mathbf{y}_i=j|\mathbf{x}_i) \log P(\mathbf{y}_i=j|\mathbf{x}_i) + \lambda \mathscr{D}(P(\mathbf{y}|\mathbf{x})) \tag{2}$$

3.3 Variations of Framework

The *Minimum Conditional Entropy Principle* plays as a general criterion for clustering. In different problem settings, we can assume that $P(\mathbf{y}|\mathbf{x})$ appears in different forms to capture the different properties[2]. For different $P(\mathbf{y}|\mathbf{x})$ plugged into the framework, the algorithms

1. This algorithm can solve most of the optimization problems derived from our framework. Particular $P(\mathbf{y}|\mathbf{x})$ may have more efficient optimization algorithm.
2. We utilize $\mathbf{w}^T\mathbf{w}$ as $\mathscr{D}(P(\mathbf{y}|\mathbf{x}))$ in below several different models ($\mathbf{w}$ is the parameters). However, in some case, this is not enough; we could add other constraints when necessary.

derived from the framework can handle both two-class clustering problems and multi-class clustering problems, even unsupervised structure problems. We specify some conventional forms as follows, but the principle can be applied to other probabilistic discriminative models to design new clustering algorithms.

Logistics regression, which assumes that

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_{\mathbf{w}}(\mathbf{x})} \exp\big(\langle \mathbf{w}_y, \varphi(\mathbf{x})\rangle\big) \tag{3}$$

where $Z_{\mathbf{w}}(\mathbf{x}) = \sum_{y\in\{1,2,\ldots,c\}} \exp\big(\langle \mathbf{w}_y, \varphi(\mathbf{x})\rangle\big)$, is one of the most popular classification algorithms due to its robustness and the close relation to large margin classifiers. Moreover, the logistics regression can deal with multi-class problems easily compared with SVM. Substituting $P(\mathbf{y}|\mathbf{x})$ into the minimum conditional entropy principle, we get the algorithm, which we named *Logistics Clustering*, as:

$$\min_{\mathbf{w}} -\sum_{i=1}^{n}\sum_{y_i} \frac{1}{Z_{\mathbf{w}}(\mathbf{x})}\langle \mathbf{w}_{y_i}, \varphi(\mathbf{x}_i)\rangle \exp\big(\langle \mathbf{w}_{y_i}, \varphi(\mathbf{x}_i)\rangle\big) + \lambda \sum_{y} \mathbf{w}_y^T\mathbf{w}_y + \sum_{i=1}^{n} \log Z_{\mathbf{w}}(\mathbf{x}) \tag{4}$$

Logistics clustering can be easily extended to non-linear form based on standard kernelization trick. As proof we present later, a modified logistics clustering converges to MMC and our framework can provide an upper bound of the modified logistics clustering.

Conditional Random Fields(CRF) (Lafferty et al. (2001)), the state-of-the-art discriminative model in capturing the relations and structured dependencies among input elements and outputs, are a family of undirected graphical models. The posterior distribution of conditional random fields can be viewed as a generalization of logistics regression:

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_{\mathbf{w}}(\mathbf{x})} \exp\big(\langle \mathbf{w}, \mathcal{F}(\mathbf{x}, \mathbf{y})\rangle\big) \tag{5}$$

$$Z_{\mathbf{w}}(\mathbf{x}) = \sum_{\mathbf{y}} \exp\big(\langle \mathbf{w}, \mathcal{F}(\mathbf{x}, \mathbf{y})\rangle\big) \tag{6}$$

where $\mathbf{w} \in \mathbb{R}^K$, $\mathcal{F} = [f_1, f_2, \ldots, f_K]^T$ is a K-dimensional vector of the feature function: $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. This form of $P(\mathbf{y}|\mathbf{x})$ extends our framework to *Unsupervised Structured Prediction*:

$$\min_{\mathbf{w}} -\sum_{i=1}^{n}\sum_{\mathbf{y}_i} \frac{\langle \mathbf{w}, \mathcal{F}(\mathbf{x}_i, \mathbf{y}_i)\rangle}{Z_{\mathbf{w}}(\mathbf{x})} \exp\big(\langle \mathbf{w}, \mathcal{F}(\mathbf{x}_i, \mathbf{y}_i)\rangle\big) + \lambda \mathbf{w}^T\mathbf{w} + \sum_{i=1}^{n} \log Z_{\mathbf{w}}(\mathbf{x}) \tag{7}$$

The above criterion extends the learning ability of CRF to unlabeled data and provides a novel unsupervised structured prediction method. Its efficiency in utilizing unlabeled data has been proved by (Jiao et al. (2006)) which employs it as a data-dependent regularization for semi-supervised structured prediction.

Next, we introduce hinge loss to exponential family to get a new form of $P(\mathbf{y}|\mathbf{x})$ as the probabilistic interpretation of SVM (Sollich (2000)),

$$P(y|\mathbf{x}) = \frac{1}{Z_{\mathbf{w}}(\mathbf{x})} \exp\big(-[1 - y(\mathbf{w}^T\varphi(\mathbf{x}) + b)]_+\big) \tag{8}$$

where $Z_{\mathbf{w}}(\mathbf{x}) = \sum_{y=\pm 1} \exp\big(-[1 - y(\mathbf{w}^T\varphi(\mathbf{x}) + b)]_+\big)$. We can obtain a maximum margin style clustering criterion.

Denoting $f(\mathbf{x}, \mathbf{w}) = (\mathbf{w}^T\varphi(\mathbf{x}) + b)$, based on the above definition and the minimum conditional entropy principle, we obtain the variant, *Maximum Relative Margin Clustering* (MRMC), as follows,

$$\min_{\mathbf{w},b} \sum_{i=1}^{n} \sum_{y_i=\pm 1} P(y_i|\mathbf{x}_i)[1 - y_i f(\mathbf{x}_i, \mathbf{w}, b))]_+ + \sum_{i=1}^{n} \log Z_{\mathbf{w}}(\mathbf{x}_i) + \lambda \mathbf{w}^T\mathbf{w} \tag{9}$$

Instead of optimizing labels individually in MMC, we associate the labels of instances with the hyperplane probabilistically and multiply the loss of each instance with its posterior probability. Meanwhile, the second term in (9) is compressing the range of $\mathbf{w}^T\mathbf{x}+b$ actually[3]. The constraints with the same effect has been discovered from the affine invariance perspective (Shivaswamy and Jebara (2010)), the projected data compact perspective (Dai and Niu (2010)) and the preventing premature convergence perspective (Zhang et al. (2007)). However, ours is more natural. The details of this variant will be studied in Section 4. We will prove that MMC is a lower bound of MRMC. Although this $P(\mathbf{y}|\mathbf{x})$ just defined for two-class clustering problems, it can be extended to multi-class clustering by hierarchical clustering (Zhang et al. (2007)) heuristically or revising the model following (Xu et al. (2006)).

At last, Gaussian distribution is taken as conditional probability and the criterion transforms as:

$$\min_{\mathbf{Y}} \sum_{i=1}^{n} \sum_{j=1}^{c} \frac{1}{Z(\mathbf{x}_i)} exp\big(-\frac{\|\varphi(\mathbf{x}_i) - \mathbf{y}_j\|^2}{\sigma_j^2}\big) \frac{\|\varphi(\mathbf{x}_i) - \mathbf{y}_j\|^2}{\sigma_j^2} + \sum_{i=1}^{n} \log Z(\mathbf{x}_i) \tag{10}$$

Remove the second term from the criterion without any effect and assume σ_j^2 the same:

$$\min_{\mathbf{Y}} \sum_{i=1}^{n} \sum_{j=1}^{c} P(\mathbf{y}_j|\mathbf{x}_i)\|\varphi(\mathbf{x}_i) - \mathbf{y}_j\|^2 \tag{11}$$

It is easy to see that this optimization problem is equivalent to fuzzy k-means clustering. It seems confusing that fuzzy k-means can be viewed as Gaussian Mixture Model which is a generative model, however, appears as a discriminative model here. The key to this phenomenon is the symmetry of 2-norm loss function, w.r.t $\mathbf{x}$ and $\mathbf{y}$, which makes the conditional posterior probability and joint probability different by a coefficient. This may be the intrinsic interpretation of (Xu et al. (2009)).

3.4 General Optimization Algorithm

In this section, we specify a general optimization algorithm for the proposed framework. Gradient algorithms can be used here. However, deterministic annealing EM (Yuille et al. (1994)) balances the computational complexity and accuracy. Almost all of the algorithms that derived from our framework can be solved by this method. Deterministic annealing

3. We will prove this property later and this is the reason that we named the algorithm as *Maximum Relative Margin Clustering*.

Algorithm 1 Deterministic Annealing EM

Input: data $\mathcal{D}$, number of clusters c and annealing parameter η
Output: $P(\mathbf{y}|\mathbf{x})$
Initialize $P_{init}(\mathbf{y}|\mathbf{x})$ and λ_0
repeat
M-step: Solve

$$min_{P(\mathbf{y}|\mathbf{x})} \sum_{i=1}^{n} \sum_{j=1}^{c} P_{t-1}(\mathbf{y}|\mathbf{x}_i) \log P(\mathbf{y}|\mathbf{x}_i) + \lambda_{t-1} \mathscr{D}(P(\mathbf{y}|\mathbf{x}))$$

E-step: Estimate $P_t(\mathbf{y}|\mathbf{x}_i)$
Decrease $\lambda_t = \eta \lambda_{t-1}$
until converge

EM is a simple generalization of the EM algorithm. Starting from a high temperature, the path of optimization is solved by gradually deceasing the temperature. At each iteration, the solution is obtained by the standard EM algorithm. In E-step, posterior probabilities are estimated by the last iterative solution. In M-step, a suitable optimization method will be applied depending on situation.

3.5 Justifications

Discussions above clarify the details of our framework motivated by extending the ability of discriminative model to unlabeled data. In this section, we present some theoretical analyses and justifications from other perspectives for our framework.

3.5.1 Spectral Clustering Perspective

Proposition 1 *Fuzzy k-means can be derived from our framework and k-means is a special case.*

Proof Substituting Gaussian distribution into (2), the fuzzy k-means is derived from our framework as (11). When approximate the $P(\mathbf{y}_j|\mathbf{x}_i)$ in E-step as

$$P(\mathbf{y}_j|\mathbf{x}_i) = \begin{cases} 1, & \|\mathbf{x}_i - \mathbf{y}_j\| < \|\mathbf{x}_i - \mathbf{y}'_j\|, j \neq j' \\ 0, & others \end{cases}$$

we obtain the k-means algorithm. ■

Corollary 2 *Spectral Clustering and Discriminative k-means can be obtained based on our framework.*

Proof As the result in (Ye et al. (2008)) that Discriminative k-means is a variant of spectral clustering with special kernel, we just need to prove that spectral clustering can be obtained from our framework. Based on the result of (Dhillon et al. (2005)), k-means is equivalent to $\max_{\mathbf{Y}} tr(\mathbf{Y}^T K \mathbf{Y})$ subject to $\mathbf{Y}^T \mathbf{Y} = \mathbf{I}_n$ where K is the kernel matrix. And the binary clustering matrix $\mathbf{Z} = W^{-\frac{1}{2}} \mathbf{Y}$, $W = diag[1/n_i]_{i=1,\ldots,k}$. Obviously, this is the criterion of spectral clustering. ■

Proposition 3 *Unsupervised Renyi's Entropy Discriminant Analysis(Unsupervised REDA), robust version of Locality Preserving Projections, can be derived from our framework.*

Proof By assuming $c = n$, $\varphi(\mathbf{x}_i) = A\mathbf{x}_i$ and ignoring normalizer, the (11) transforms as:

$$\max_{\mathbf{A}} \sum_{i=1}^{n}\sum_{j=1}^{n} -exp\big(-\frac{\|A\mathbf{x}_i - A\mathbf{x}_j\|^2}{\sigma_j^2}\big)\frac{\|A\mathbf{x}_i - A\mathbf{x}_j\|^2}{\sigma_j^2}$$

When employing EM algorithm to update the posterior probability, the solution path is the same as Unsupervised REDA. ■

3.5.2 Max-Margin Perspective

Proposition 4 *The term* $\log Z_{\mathbf{w}}(\mathbf{x}_i)$ *in (9) is compressing the range of* $\mathbf{w}^T\mathbf{x} + b$ *actually and (9) is an implementation of the Relative Margin Clustering.*

Proof The objective function of *Maximum Relative Margin Clustering* is expressed as,

$$\min_{\mathbf{w},b} \sum_{i=1}^{n}\sum_{y_i=\pm 1} P(y_i|\mathbf{x}_i)[1 - y_i f(\mathbf{x}_i, \mathbf{w}, b))]_+ + \sum_{i=1}^{n} \log Z_{\mathbf{w}}(\mathbf{x}_i) + \lambda \mathbf{w}^T\mathbf{w}$$

For saving space, we denote the hinge loss $\mathcal{L}(y, f(\mathbf{x}, \mathbf{w}, b))$ as ξ_y. Based on the fact that $\log Z_{\mathbf{w}}(\mathbf{x}_i)$ is *soft maximum functions* which is an approximation of $\max(-\xi_1, -\xi_{-1})$ where $\xi \geq 0$, we transform the above formulation approximately as

$$\begin{aligned} \min\nolimits_{\mathbf{w},b,B} \quad & \sum_{i=1}^{n}\sum_{y_i=\pm 1} P(y_i|\mathbf{x}_i)[1 - y_i f(\mathbf{x}_i, \mathbf{w}, b))]_+ + \mu B + \lambda \mathbf{w}^T\mathbf{w} \\ s.t. \quad & |\mathbf{w}^T\mathbf{x}_i + b| < B \quad i = 1, \ldots, n \end{aligned}$$

Noticing that $-P(y_i|\mathbf{x}_i)$ in the first term can be written as the derivate of $\log Z_{\mathbf{w}}(\mathbf{x}_i)$ w.r.t. corresponding loss, then the first term is an approximation of $\min(\xi_1, \xi_{-1})$.

$$\begin{aligned} & \sum_{y_i=\pm 1} P(y_i|\mathbf{x}_i)[1 - y_i f(\mathbf{x}_i, \mathbf{w}, b))]_+ \\ = \quad & -\bigtriangledown_{\vec{\xi}} (\log Z_{\mathbf{w}}(\mathbf{x}_i))^T \vec{\xi} \approx \bigtriangledown_{\vec{\xi}} \min(\xi_1, \xi_{-1})^T \vec{\xi} = \min(\xi_1, \xi_{-1}) \end{aligned}$$

Hence, the original objective function can approximately transform as

$$\begin{aligned} \min \mathcal{L}(\mathbf{w}, b) \quad \approx \quad & \min_{\mathbf{w},b,B} \sum_{i=1}^{n} (\delta(sgn(f(\mathbf{x}_i, \mathbf{w}, b)))[1 - f(\mathbf{x}_i, \mathbf{w}, b)]_+ \\ + \quad & (1 - \delta(sgn(f(\mathbf{x}_i, \mathbf{w}, b))))[1 + f(\mathbf{x}_i, \mathbf{w}, b)]_+) + \mu B + \lambda \mathbf{w}^T\mathbf{w} \\ s.t. \quad & |\mathbf{w}^T\mathbf{x}_i + b| < B \quad i = 1, \ldots, n \end{aligned} \tag{12}$$

where $\delta(1) = 1$ and $\delta(-1) = 0$. Obviously, (9) is another unsupervised implementation of the Relative Margin Machine (Shivaswamy and Jebara (2010)) which maximizes the margin while considering the spread of projection data and affine invariance. ■

Corollary 5 *When ignoring the second term in (9), Maximum Relative Margin Clustering provides an upper bound of Maximum Margin Clustering.*

Proof Based on Proposition 4 and setting $\mu = 0$, it is trivial that

$$\min_{\mathbf{w},b} \mathcal{L}(\mathbf{w},b) \geq \min_{\mathbf{y}} \min_{\mathbf{w},b} \sum_{i=1}^{n} ((1-\delta(y_i))[1+f(\mathbf{x}_i,\mathbf{w},b)]_+ + \delta(y_i)[1-f(\mathbf{x}_i,\mathbf{w},b)]_+) + \lambda \mathbf{w}^T\mathbf{w} \quad \blacksquare$$

Proposition 6 *Modified Logistics Clustering converges to Maximum Margin Clustering. Substituting the form of Modified Logistics Regression into our framework, we have an upper bound.*

Proof Based on the result in (Zhang et al. (2003)), the loss function of Modified Logistics Regression

$$\min_{\mathbf{w}} g_r(\mathbf{x},y,\mathbf{w}) = \frac{1}{r}\sum_{i=1}^{n} \log(1+\exp(-r(y_i\mathbf{w}^T\mathbf{x}_i - 1))) + \lambda \mathbf{w}^T\mathbf{w}$$

converges to the loss of SVM when r increases. Thus, the Modified Logistics Clustering $\min_{\mathbf{y},\mathbf{w}} g_r(\mathbf{x},y,\mathbf{w})$ converges to the primal form of MMC:

$$\begin{aligned} \min_{\mathbf{y},\mathbf{w},b} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi_i \\ s.t. \quad & y_i(\mathbf{w}^T\varphi(\mathbf{x}_i)+b) \geq 1-\xi_i, \xi_i \geq 0, y_i \in \{-1,1\} \end{aligned}$$

Therefore, we have the first conclusion. The proof of second conclusion is similar to the proof of Corollary 5. Due to the space limitation, we neglect the details. ■

4. Accelerated Maximum Relative Margin Clustering

In this section, we accelerate the optimization process of MRMC based on the Proposition 4. Of course that **Algorithm 1** can handle MRMC, however, this method is too general and does not aim at the specific problem. The algorithm spends too much time on estimating posterior probabilities and these cost seem unworthy because the terms in (9) contains posterior probability, e.g., $\sum_{y_i=\pm 1} P(y_i|\mathbf{x}_i)[1-y_i f(\mathbf{x}_i,\mathbf{w},b))]_+$, are just the minimum function's approximation as we proved. On the other hand, only the penalty $\mathbf{w}^T\mathbf{w}$ is not enough for large margin clustering; the class-balance constraint, $\frac{1}{n}\sum_{i=1}^{n}\max(y_i,0) = r$, plays a very important role in avoiding trivial solution (Xu et al. (2004); Zhang et al. (2007)). But it cannot be added into deterministic annealing EM algorithm straightforwardly. Thus, we need to modify the EM algorithm for MRMC. Based on the approximation (12) of MRMC, we can utilize the property to reduce these costs and satisfy the extra constraint.

In E-step, by approximating MRMC via (12), it is unnecessary to calculate the posterior probability, $\delta(sgn(f(\mathbf{x},\mathbf{w})))$ is enough. Meanwhile, to guarantee that the objective function strictly decreases after adding the class-balance constraint in E-step, just labeling the data positive and negative with fixed ratio may be not suitable. We employ the pair-switch strategy as (Joachims (1999)): if $\exists(y_k^t, y_l^t)$ where $y_k^t = 1, y_l^t = -1$ satisfies the condition:

$$\mathcal{L}(1,f(\mathbf{x}_k,\mathbf{w},b)) + \mathcal{L}(-1,f(\mathbf{x}_l,\mathbf{w},b)) > \mathcal{L}(1,f(\mathbf{x}_l,\mathbf{w},b)) + \mathcal{L}(-1,f(\mathbf{x}_k,\mathbf{w},b)) \qquad (13)$$

then switch the labels of $\mathbf{x}_i$ and $\mathbf{x}_j$. Moreover, to accelerate the learning procedure further, we can fix the labels of some points when we are assured about them in iterations. In M-step, we formulate the optimization to SVM style quadratic programming which can be solved by sequential minimal optimization. Besides keeping affine invariance and controlling the spread of projections (Shivaswamy and Jebara (2010)), the additional relative margin constraint can also prevent premature convergence and increase the possibility to get the global optimal or better local optimal (Zhang et al. (2007)).

Theorem 7 *The objective function (12) with the class-balance constraint is decreasing at each iteration and* ***Algorithm 2*** *converges in a finite number of steps.*

Proof For saving space, we denote the loss for i-th instance in t-th iteration $\mathcal{L}(1, f(\mathbf{x}_i,\mathbf{w}^t, b^t))$ as ξ_{1i}^t and $\mathcal{L}(-1, f(\mathbf{x}_i, \mathbf{w}^t, b^t))$ as ξ_{-1i}^t, where $(\mathbf{w}^t, b^t)$ is the optimal in t-th iteration. Assume that the instances pair $y_k^t = 1, y_l^t = -1$ satisfies the switch condition, we have

$$\begin{aligned}
& \lambda \mathbf{w}^{tT}\mathbf{w}^t + \sum_{i=1}^{n} \left(\delta(y_i^t)\xi_{1i}^t + (1-\delta(y_i^t))\xi_{-1i}^t\right) + \mu B^t \\
= \; & \lambda \mathbf{w}^{tT}\mathbf{w}^t + \sum_{i=1, i\neq k,l}^{n} \left(\delta(y_i^t)\xi_{1i}^t + (1-\delta(y_i^t))\xi_{-1i}^t\right) + \xi_{1k}^t + \xi_{-1l}^t + \mu B^t \\
> \; & \lambda \mathbf{w}^{tT}\mathbf{w}^t + \sum_{i=1, i\neq k,l}^{n} \left(\delta(y_i^t)\xi_{1i}^t + (1-\delta(y_i^t))\xi_{-1i}^t\right) + \xi_{-1k}^t + \xi_{1l}^t + \mu B^t \\
\geq \; & \min_{\mathbf{w},b} \lambda \mathbf{w}^T\mathbf{w} + \sum_{i=1}^{n} \left(\delta(y_i^{t+1})\xi_{1i} + (1-\delta(y_i^{t+1}))\xi_{-1i}\right) + \mu B
\end{aligned}$$

The first inequality holds due to the pair-switch criterion (13). It is easy to verify that the new labelling still satisfies the class-balance constraints. This means that the optimal objective decreases after E-step in **Algorithm 2**. Since there is only a finite number of permutation of instances, the algorithm will converge to a stable labelling. ■

Although the proposed algorithm is solving a sequence of Quadratic Programming(QP) as iterSVR (Zhang et al. (2007)), there are still some important differences between them. In iterSVR, the QP formulates as a Support Vector Regression which loses the sparse property while we solve SVM with relative margin constraints in each iteration. Instead of predicting labels directly in iterSVR, we switch the labels guaranteeing the convergence and the class-balance constraint.

Time Complexity Analysis Compared with SDP based MMC and GMMC that are $O(n^{6.5})$ and $O(n^{4.5})$ (Zhang et al. (2007)), our algorithm is efficient. In M-step, we solve an SVM style optimization whose empirical complexity usually scales between $O(n)$ and $O(n^{2.3})$. In E-step, the pair-switch takes $O(n^2)$ at most. Empirically, the number of iterations is usually much smaller than the number of instances. Thus, ARMC is much faster. This will be proved by the experiments in Section 5.

Multi-Class Clustering Extension Although the $P(\mathbf{y}|\mathbf{x})$ we employed in ARMC is designed for two-class clustering problems, we can extend the ARMC to multi-class clustering problems easily. Following the divisive hierarchical clustering methods (Shi and Malik (2000); Zhang et al. (2007)), we can execute the two-class clustering method recursively.

Algorithm 2 Accelerated Maximum Relative Margin Clustering (ARMC)

Input: data $\mathcal{D}$, kernel matrix K
Output: $\mathbf{y}$
Initialize $\mathbf{w}$ and assign data with the initial $\mathbf{w}$ under the class-balance constraint
repeat
M-step: Solve

$$min_{\mathbf{w},b} \quad \sum_{i=1}^{n}\big((1-\delta(y_i^t))[1+f(\mathbf{x}_i,\mathbf{w},b)]_+ +\delta(y_i^t)[1-f(\mathbf{x}_i,\mathbf{w},b)]_+\big)+\mu B+\lambda\mathbf{w}^T\mathbf{w}$$
$$s.t. \qquad |\mathbf{w}^T\varphi(\mathbf{x}_i)+b|<B \quad i=1,\ldots,n$$

E-step: **if** $\exists(y_k^t,y_l^t)$ satisfies the condition (13)
then switch the labels of $\mathbf{x}_k$ and $\mathbf{x}_l$.
until labeling stably

Another multi-class clustering extension approach is transforming the hinge loss in $P(\mathbf{y}|\mathbf{x})$ as (Xu et al. (2006)), then plugging it into our framework. In this paper, we focus on two-class clustering problems because of the limitation of space and the details of multi-class clustering extension of ARMC will be discussed in our future work.

5. Experiments

We implement the ARMC algorithm, the generalized maximum margin clustering(GMMC) (Valizadegan and Jin (2006)), normalized spectral clustering(NC) (Ng et al. (2001)), and the k-means in **Matlab**. We use **CVX**[4] to solve the SDP problem in GMMC. **SVM**^{light}[5] is employed in ARMC. Experiments are run on a 2.40GHz Intel(R) Core(TM)2 Duo PC running Windows XP.

We first evaluate on some synthetic data sets to validate our algorithm. The results are illustrated in Figure. 2. The first two scatter-plot are the results on *2moons* and *2circles*. The last two pictures are the clustering results on mixture Gaussian distributions with different distances between two central points. It can be seen that our algorithm can provide a convincing division even in the problem that the clusters overlap heavily.

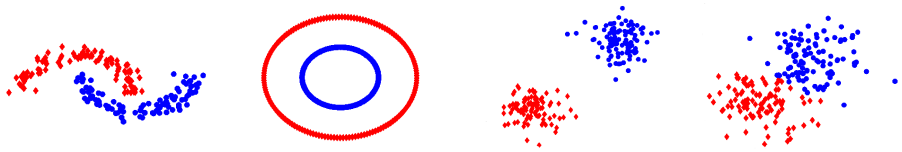


Figure 2: Illustration on some synthetic data sets

Because the clustering ability of divisive hierarchical multi-class clustering extension depends on the discriminative ability of the proposed algorithm on two-class clustering

4. The CVX matlab code be obtained from http://www.stanford.edu/∼boyd/cvx/
5. The program of **SVM**^{light} is available at http://svmlight.joachims.org/

problems, we focus on comparing the performances in the two-class clustering setting on real-world datasets in this paper following (Xu et al. (2004); Valizadegan and Jin (2006); Zhang et al. (2007); Li et al. (2009)).

Benchmark data sets of Handwritten Digits, `MNIST` and `USPS`, are taken for performance evaluation. It has been demonstrated that the instances with different labels form different clusters in handwritten digits datasets (Hinton and Salakhutdinov (2006)); thus evaluating clustering on such datasets is reasonable (Guyon et al. (2009)). The `MNIST` datasets contains 60,000 images of handwritten Arabic number 0 to 9. Each of the instances has been size-normalized in a 28×28 pixels image. The `USPS` contains 11,000 images of handwritten number size-normalized 16×16. We take the most difficult discriminative pairs such as 1 vs. 7, 3 vs. 8, 5 vs. 8 and 8 vs. 9 from `MNIST` and `USPS`. The experimental instances are sampled randomly from the dataset with size increasing from 100 to 1000. To tune the parameters conveniently, we scale the grayscale of each pixel from an integer in [0, 255] to a real number in [0, 1]. For normalized spectral clustering, k, the number of nearest neighbors, is tuned from 3 to 12. RBF kernel is used throughout the experiments to evaluate the similarity between instances and the σ^2 is chosen from $\{10^i | i = -2, \ldots, 3\}$. The λ of ARMC is selected from the set of $\{0.02, 0.2, 2, 20, 200\}$ and we simplify the model by fixing B and μ beforehand. For GMMC, we set $C_e = 10^4$ as (Valizadegan and Jin (2006)) claimed. All couples (C_δ, σ^2) in $\{10^0, \ldots, 10^6\} \times \{10^{-2}, \ldots, 10^3\}$ are considered.

We follow strategy in (Xu et al. (2004); Valizadegan and Jin (2006); Zhang et al. (2007); Li et al. (2009)) to evaluate the accuracy of algorithms. We first remove the label of instances and perform clustering algorithms to divide the data, and then utilize misclassification error rate according to true labels. The best clustering results of 10-trials are reported in Figure. 3 and Figure. 4. Because of the high complexity of SDP, GMMC costs too much to handle more than 700 data. Our results on `MNIST` and `USPS` are the best or achieve the state-of-the-art in most situations except the digit pair 3 vs. 8 on `USPS`. Another phenomenon in experiments is that although GMMC achieves better accuracy sometimes, the variance of GMMC is larger than the other algorithms. In contrast, our algorithm is stable. Maybe the reason is that GMMC is sensitive to parameters and any changes of dataset will affect the performance heavily.

Speed of algorithm is another important issue besides accuracy. SDP based MMC solving methods are computationally expensive and not practical to handle even medium sized problems. We demonstrate the speed advantage of the proposed algorithm. `Heart` and `Ionosphere` are adopted, which contain 270 instances and 351 instances, respectively. We also evaluate the running time on subsets of `MNIST` and `TDT2`. We select 400 instances from the pair 3 vs. 8, 1 vs. 7 randomly and 2 categories from `TDT2` randomly, each of which contains 300 instances. The empirical running time is showed in Table 1 and we can conclude that ARMC is much more practical and faster compared with SDP based MMC.

6. Conclusion

We introduce the low-density separation assumption into clustering making it possible to extend the learning ability of discriminative models to unsupervised setting. As an implementation of this assumption, we present an information-theoretic framework which is based on *Minimum Conditional Entropy Principle*. Most of the popular clustering algo-

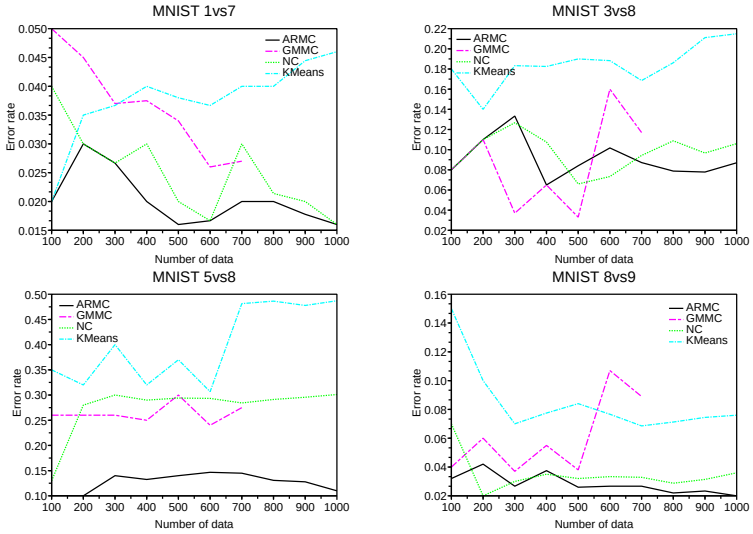


Figure 3: Experimental results on MNIST

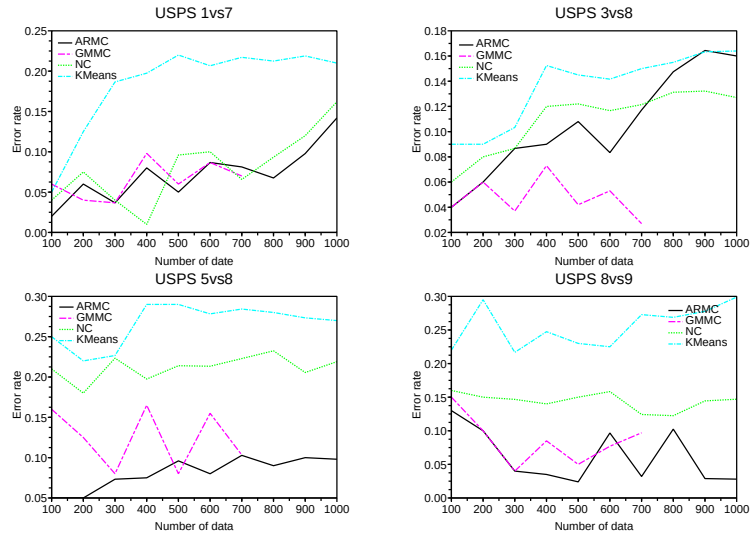


Figure 4: Experimental results on USPS

rithms, such as spectral clustering and maximum margin clustering, connect closely to this framework. Moreover, we derive a novel algorithm, *Accelerated Maximum Relative Margin Clustering*, which maximizes the margin while considering spread of projections and affine invariance, for clustering. The efficiency and effectiveness of the proposed algorithm have been demonstrated empirically.

For future work, the stability and universal consistency of the proposed algorithm need to be examined from theoretical perspective. Deriving more efficient discriminative clustering algorithms based on the framework is also an interesting and attractive direction.

Acknowledgments

This work was supported by NSFC #60275025. We thank the reviewers for the suggestions.

Table 1: List of Performance Comparison

	Dataset	k-means	NC	GMMC	ARMC
Time	Heart(Stalog)	0.0037	0.4139	13.03	1.302
(sec.)	Ionosphere	0.0044	0.7991	26.84	12.42
	MNIST 3vs8	0.2392	1.4381	23.84	2.637
	MNIST 1vs7	0.2896	1.4132	29.99	2.869
	TDT2	0.5066	4.9626	79.65	9.771
Error Rate	Heart(Stalog)	**28.8**	33.3	30.7	30.3
(%)	Ionosphere	28.6	14.8	17.1	**11.7**
	MNIST 3vs8	18.3	10.8	**6.5**	**6.5**
	MNIST 1vs7	4.0	3.0	3.8	**2.0**
	TDT2	25.0	**2.2**	8.0	4.0

References

M. Belkin, P. Niyogi, and V. Sindhwani. On manifold regularization. In *AISTATS'05*, 2005.

S. Ben-David and U. von Luxburg. Relating clustering stability to properties of cluster boundaries. In *COLT'08*, 2008.

S. Ben-David, T. Lu, D. Pál, and M. Sotáková. Learning low-density separators. In *AISTATS'09*, 2009.

O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

B. Dai and G. Niu. Compact margin machine. In *PAKDD'10*, 2010.

I. S. Dhillon, Y. Guan, and B. Kulis. A unified view of kernel k-means, spectral clustering and graph cuts. Technical Report TR-04-25, U.Texas Dept. of Computer Science, 2005.

Y. Grandvalet and Y. Bengio. Semi-supervised learning by entropy minimization. In *NIPS'04*, 2004.

I. Guyon, U. von Luxburg, and R. Williamson. Clustering: Science or art? In *NIPS 2009 Workshop on Clustering Theory*, 2009.

G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, July 2006.

F. Jiao, S. Wang, C. Lee, R. Greiner, and D. Schuurmans. Semi-supervised conditional random fields for improved sequence segmentation and labeling. In *ACL'06*, 2006.

T. Joachims. Transductive inference for text classification using support vector machines. In *ICML '99*, 1999.

J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML'01*, 2001.

Y.F. Li, I. W. Tsang, J. T. Kwok, and Z.H. Zhou. Tighter and convex maximum margin clustering. In *AISTATS'09*, 2009.

A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS'01*, 2001.

T. J. O'Neill. Normal discrimination with unclassified observations. *Journal of the American Statistical Association*, 73(364):821–826, December 1978.

M. Seeger. A taxonomy for semi-supervised learning methods. In Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors, *Semi-Supervised Learning*. The MIT Press, 2006.

J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.

P. K. Shivaswamy and T. Jebara. Maximum relative margin and data-dependent regularization. *Journal of Machine Learning Research*, 11, 2010.

P. Sollich. Probabilistic methods for support vector machines. In *NIPS'00*, 2000.

H. Valizadegan and R. Jin. Generalized maximum margin clustering and unsupervised kernel learning. In *NIPS'06*, 2006.

L. Xu, J. Neufeld, B. Larson, and D. Schuurmans. Maximum margin clustering. In *NIPS'04*, 2004.

L. Xu, D. Wilkinson, F. Southey, and D. Schuurmans. Discriminative unsupervised learning of structured predictors. In *ICML'06*, 2006.

L. Xu, M. White, and D. Schuurmans. Optimal reverse prediction: a unified perspective on supervised, unsupervised and semi-supervised learning. In *ICML '09*, 2009.

J. Ye, Z. Zhao, and M. Wu. Discriminative k-means for clustering. In *NIPS'08*, 2008.

X. T. Yuan and B. G. Hu. Robust feature extraction via information theoretic learning. In *ICML'09*, 2009.

A. L. Yuille, P. Stolorz, and J. Utans. Statistical physics, mixtures of distributions, and the em algorithm. *Neural Computation.*, 6(2):334–340, 1994. ISSN 0899-7667.

J. Zhang, R. Jin, Y. Yang, and A. Hauptmann. Modified logistic regression: An approximation to svm and its applications in large-scale text categorization. In *ICML'03*, 2003.

K. Zhang, I. W. Tsang, and J. T. K. Maximum margin clustering made practical. In *ICML'07*, 2007.

B. Zhao, F. Wang, and C. S. Zhang. Efficient maximum margin clustering via cutting plane algorithm. In *SDM'08*, 2008.