To appear in IEEE conference on computer vision and pattern recognition, 2005

Learning With Constrained and Unlabelled Data

Tilman Lange[1] langet@inf.ethz.ch
Martin H.C. Law[2] lawhiu@cse.msu.edu
Anil K. Jain[2] jain@cse.msu.edu
Joachim M. Buhmann[1] jbuhmann@inf.ethz.ch
1. Institute of Computational Science
ETH Zurich
CH-8050 Zurich, Switzerland
2. Dept. of Computer Science and Engineering
Michigan State University
East Lansing, MI 48823, USA

Abstract

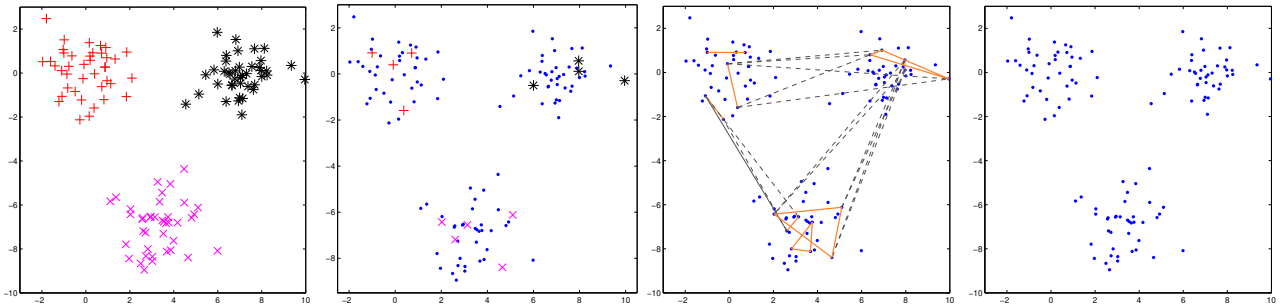
Classification problems abundantly arise in many computer vision tasks – being of supervised, semi-supervised or unsupervised nature. Even when class labels are not available, a user still might favor certain grouping solutions over others. This bias can be expressed either by providing a clustering criterion or cost function and, in addition to that, by specifying pairwise constraints on the assignment of objects to classes. In this work, we discuss a unifying formulation for labelled and unlabelled data that can incorporate constrained data for model fitting. Our approach models the constraint information by the maximum entropy principle. This modeling strategy allows us (i) to handle constraint violations and soft constraints, and, at the same time, (ii) to speed up the optimization process. Experimental results on face classification and image segmentation indicates that the proposed algorithm is computationally efficient and generates superior groupings when compared with alternative techniques.

1. Introduction

Many problems in computer vision can be cast as classification and grouping problems. Examples include low level image segmentation and object recognition/classification. Often, a clear distinction is made between problems that are (i) supervised or (ii) unsupervised, the first involving only labelled data while the latter involving only unlabelled data in the process of learning. Recently, there has been a growing interest in a hybrid setting, called *semi-supervised*, where the labels of only a portion of the data set are available for training. The unlabelled data, instead of being discarded, are used in the training process to provide information about the data density $p(\mathbf{x})$, so that the joint data and label density $p(\mathbf{x}, y)$ can be more appropriately inferred. Partially labelled data are typical in applications where data collection is easy but data labelling is expensive. Remote sensing serves as a good example: taking a high resolution SAR image is relatively easy compared to the labor-intensive process of correctly labelling pixels in the scene. In molecular biology, the functional classification of proteins based on sequence or secondary structure information represents an example of similar nature: the data acquisition process is relatively cheap while the cost of identifying the correct functional category is high.

Instead of specifying the class labels, a "weaker" way of specifying a priori knowledge about the desired model is via *constraints*. A pairwise *must-link* constraint corresponds to the requirement that two objects should be assigned the same label, whereas the labels of two objects participating in a *must-not-link* constraint should be different. Must-link constraints are generally easier to model because they usually represent an equivalence relation. Constraints can be particularly beneficial in data clustering [6], where precise definitions of classes are absent. In the search for good models, one would like to include all (trustworthy) information that is available, no matter whether it is about unlabelled, data with constraints, or labelled data. Figure 1 illustrates this spectrum of different types of prior knowledge that can be included in the process of classifying data.

Table 1 summarizes different approaches in the literature to clustering with constraints – for which we provide a novel approach. In the second and the third type of approaches, the subjects of inference are the labels of the objects. These bear a close similarity to the transductive learning setting as introduced by Vapnik in [20]. From a probabilistic point of view, one specifies a prior on the class labels for points participating in constraints or for labelled points. Combinations of labels that violate the constraints / prior label information are either forbidden by having zero prior probability, or they are penalized by having small prior probability values. Labels of originally unlabelled or constrained points are affected by the prior knowledge only indirectly by the parameter estimates. This, however, can lead to "discontinuous" label assignment: two points at exactly the same location, one with constraint and one without, can be assigned different labels! In [23], smoothness of the cluster



(a) Supervised (b) Partially labelled (c) Partially constrained (d) Unsupervised

Figure 1. Spectrum between supervised and unsupervised learning: dots correspond to points without any labels. Points with labels are denoted by circles, asterisks and crosses. In (c), the must-link and must-not-link constraints are denoted by solid and dashed lines, respectively.

Summary	Key ideas	Examples
Distance editing	Modify the distance/proximity matrix due to the constraints	[10, 9]
Constraints on labels	The cluster labels are inferred under the restriction that the constraints are always satisfied	[21, 22, 23, 19]
Penalize violation	Penalty for constraint violations. ICM used for greedy optimization.	[3, 4, 2]
Modify generation model	Generation process of data participating in constraints is modified leading to parameter estimates consistent with constraints	[12, 11]

Table 1. Different algorithms for clustering with constraints

labels is enforced to avoid this. In [10], the distance metric is modified in view of the constraints to propagate the effect of a constraint to the neighboring space. However, this approach is not very robust: a single constraint can provide a shortcut and change the distance matrix dramatically.

In this work, we tackle the problem of constraint-based classification in a model-based framework. Unlabelled and constrained data are integrated in a manner analogous to the integration of labelled and unlabelled data. The parameters are estimated by a combination of the unlabelled and constrained data. We start with a natural formulation of semi-supervised learning that enables a smooth transition between supervised and completely unsupervised problem settings. This formulation is extended to incorporate prior information in the form of pairwise constraints: for the constraint data, we introduce a maximum entropy prior on the hidden class labels and thereby switching the sampling paradigm. In order to maintain computational feasibility, a mean field approximation is employed that leads to accurate posterior estimates for the constraint data. In contrast to the approach in [3], the approximation allows us to employ more sophisticated optimization compared to [3] and [21, 22]. Furthermore, the method by Shental *et al.* [19] does not allow the specification of the importance of constraint examples.

We want to emphasize that the approach presented here is of rather general applicability: it can be easily applied to mixture-based approaches as well as to all clustering principles that rely on minimizing the distance to cluster centroids with respect to a Bregman divergence [1] (e.g. k-means with squared Euclidean error or histogram clustering with Kullback-Leibler divergence [5]). Furthermore, thanks to the formal equivalence proved in [18], which states that pairwise clustering problems can be recast as k-means problems in a suitable embedding space, our approach is also applicable to pairwise data clustering problems.

2. Integrating Partial Label Information

In unsupervised model-based classification, the data density $p(\mathbf{x}|\Theta)$ is usually a function of the parameters Θ. In this setting, learning consists of identifying a model that is suitable for the data by parameter optimization. Following the Maximum Likelihood (ML) approach, parameter estimates are obtained by minimizing the negative data log-likelihood $\mathcal{L}(\mathcal{X}^u;\Theta) = -\sum_{\mathbf{x}\in\mathcal{X}^u} \log p(\mathbf{x}|\Theta)$ of the (unlabelled) data $\mathcal{X}^u$, where $p(\mathbf{x}|\Theta) = \sum_\nu \pi_\nu p(\mathbf{x}|\theta_\nu)$ in model-based clustering, with ν as cluster label and θ_ν as the parameter for the ν-th class-conditional density. Algorithmically, the often intractable global optimization over the parameter space is replaced by an iterative, local optimization procedure, the *Expectation-Maximization (EM)* algorithm [16]. The latter is also an intrinsic part of the *Deterministic Annealing (DA)* optimization procedure [17].

In supervised model-based classification, one has a data set $\mathcal{X}^l$ and the corresponding labels, denoted by y_i for $\mathbf{x}_i \in \mathcal{X}^l$. For simplicity, we assume $y_i \in \{1, \ldots, k\}$, where k is the number of classes. The application of the maximum likelihood principle to this data tells us to choose the parameters Θ that minimize $\mathcal{L}(\mathcal{X}^l, \mathcal{Y}; \Theta) = -\sum_{\mathbf{x}_i \in \mathcal{X}^l} \log p(\mathbf{x}_i, y_i|\Theta)$, where again $p(\mathbf{x}_i, y_i|\Theta) = \pi_{y_i} p(\mathbf{x}_i|\theta_{y_i})$.

In semi-supervised learning, we have both labelled and unlabelled data where the latter should be employed in order to get an improved model (i.e. parameter) estimate.[1] It is natural to require that the estimated parameters yield high likelihood for *both*, labelled and unlabelled data, which can be achieved by considering

$$\min_{\Theta} \left(\alpha \mathcal{L}(\mathcal{X}^u; \Theta) + (1-\alpha) \mathcal{L}(\mathcal{X}^l, \mathcal{Y}; \Theta) \right) \qquad (2.1)$$

as the objective function. Here, $\alpha \in [0, 1]$ controls the influence of the labelled/unlabelled data on the parameter estimation. For very small α, the unlabelled data is almost ignored while for α close to 1, the labelled data hardly enters the model estimate. Clearly, the choice of α is critical in this context since it might significantly determine the resulting model, in particular in the case of a model mismatch. By choosing $\alpha = |\mathcal{X}^u|/|\mathcal{X}^u \cup \mathcal{X}^l|$, every data point – independent of it being labelled or unlabelled – is considered equally important while by setting $\alpha = 1/2$ both *data sources*, labelled and unlabelled, will have the same influence. A different strategy is to choose the largest α with *minimal empirical test error* on the labelled data $\mathcal{X}^l$. By means of this strategy, one integrates the log-likelihood and the empirical classification risk into the objective function. Our formulation ensures that the model parameters are actually modified and affected by the prior information, not just the posterior assignment probabilities of some objects. This formulation has the consequence that the final parameter estimate essentially becomes a convex combination of the parameter estimates due to purely unlabelled data and purely labelled data, with α as a trade-off parameter between parameter estimates. For Gaussian class-conditional densities, for example, the class-specific means μ_ν are estimated in each EM iteration by

$$\mu_\nu = \frac{\alpha \sum_{\mathbf{x}_i \in \mathcal{X}^u} \rho_i(\nu) \mathbf{x} + (1-\alpha) \sum_{\mathbf{x}_i \in \mathcal{X}^l} \mathbf{1}\{y_i = \nu\} \mathbf{x}}{\alpha \sum_{\mathbf{x}_i \in \mathcal{X}^u} \rho_i(\nu) + (1-\alpha) \sum_{\mathbf{x}_i \in \mathcal{X}^l} \mathbf{1}\{y_i = \nu\}}. \qquad (2.2)$$

Here, $\rho_i(\nu)$ represents the posterior probability estimated from the *unlabelled data*. For central clustering with Bregman divergences, one also obtains an estimate analogous to the one in eq. (2.2). A similar update equation for covariance matrices is straightforwardly obtained by taking the derivative of convex combination of the two log-likelihoods.

[1] We assume that there is at least one labelled sample for each class.

3. Integrating Pairwise Constraints

The focus of the present work is the integration of pairwise must-link and must-not-link constraints into the process of model fitting. We want to achieve this in a way similar to the integration of partially labelled data as described in section 2. Our perspective is that specifying constraints amounts to specifying an *object-specific* prior model for the assignment of constraint data to different classes. This contrasts the sampling paradigm underlying a standard mixture model, which is given by the following two-stage process: (i) a class is picked with probability π_ν, and (ii) the datum $\mathbf{x}$ is generated according to $p(\mathbf{x}|\theta_\nu)$. For constrained data, the first step of the sampling process is no longer object independent. We provide a *Maximum Entropy (ME)* [7, 8] prior model defined on the hidden variables that captures the dependencies, and propose an efficient implementation of the model by means of a Mean-Field approximation. At first, however, we discuss constraint specification.

3.1 Constraint Specification

Suppose a user provides the information about objects i and j that they should be linked together, i.e., be assigned to the same group. We introduce a binary indicator variable $a_{i,j}$ such that it is 1 if i and j should be in the same group, and 0 otherwise. If the must-link constraints contain no errors, it is natural to assume that the must-link constraints represent an *equivalence relation*, i.e., they should be symmetric, reflexive and transitive. Therefore, the transitive closure of the user-provided, must-link constraints represents a useful augmentation of the constraint set. For an equivalence relation, there exists a partitioning of the set in relation. By considering the graph with $(a_{i,j})$ as adjacency matrix, the connected components (*cliques* in the augmented graph) of the graph correspond to the equivalence classes. While performing this augmentation is a must for certain approaches, e.g. for the one in [19], it is optional in our approach. Note that augmentation of constraints can increase the number of erroneous constraints if there exist mis-specified constraints.

Similar to must-link constraints, *must-not-link* constraints can be expressed by employing an additional indicator variable $b_{i,j}$ with $b_{i,j} = 1$ if i and j should not be linked, and 0 otherwise. Negative or must-not-link constraints, despite their symmetry, do not represent an equivalence relation. However, given the transitive closure of positive constraints, there is some structure that can be exploited. Suppose there is a negative constraint between i and j, i.e. $b_{i,j} = 1$. The negative constraints can be augmented by adding negative constraints between i' and j' where $a_{i',i} = 1$ and $a_{j',j} = 1$. In other words, negative

constraints can be considered as constraints *between components*. Again, we want to emphasize that performing this augmentation is also optional for must-not-link constraints; e.g. , augmenting conflicting constraints is not reasonable.

3.2 Including Constraints in the Inference

Consider the data set $\mathcal{X}^c$ which consists of all data that participate in at least one constraint. Since only constraints and no labels are prescribed for the data in $\mathcal{X}^c$, we consider the label y_i for $\mathbf{x}_i \in \mathcal{X}^c$ as a hidden or latent variable.

We want to penalize a constraint violation whenever the latent variables in a constraint are different (the same) while they are supposed to be the same (different). Hence, the penalty for violation of positive and negative constraints becomes $a_{i,j}\mathbf{1}\{y_i \neq y_j\}$, and $b_{i,j}\mathbf{1}\{y_j = y_j\}$, respectively, where $\mathbf{1}$ denotes the indicator function.

As stated above, the user specifies a preference for or against certain labellings of the constrained data. We turn this information into a prior on the label assignment for the data in $\mathcal{X}^c$ by applying the maximum entropy principle: find the prior distribution $p(\mathbf{y}) = p(y_1, \ldots, y_n)$ for the cluster labels of the data points $\mathbf{x}_i \in \mathcal{X}^c$ such that the entropy $H(p)$ is maximized while the expected number of constraint violations,

$$\sum_{y_1=1}^{k} \cdots \sum_{y_n=1}^{k} p(\mathbf{y}) \sum_{i,j} \left(a_{i,j}\mathbf{1}\{y_i \neq y_j\} + b_{i,j}\mathbf{1}\{y_i = y_j\}\right), \tag{3.1}$$

is bounded by κ^+ for positive and κ^- for negative constraints. Note, that we can rewrite the problem of finding the maximum entropy distribution as a Lagrangian functional with Lagrange parameters λ^+ and λ^-; the latter control the amount of penalty for a constraint violation. The solution to this inference problem is the so-called Gibbs distribution, and, in our case, it is

$$\frac{1}{Z}\prod_{i,j} \exp\left(-\lambda^+ a_{i,j}\mathbf{1}\{y_i \neq y_j\} - \lambda^- b_{i,j}\mathbf{1}\{y_i = y_j\}\right), \tag{3.2}$$

where Z is the normalization constant. A similar prior has been proposed independently in [14] using a heuristic argument. The prior can be considered as the distribution in a pairwise Markov random field defined on the label variables where the graph structure is given by the constraint parameters $a_{i,j}$ and $b_{i,j}$. The result is a prior factorial over the edges in the Markov Random Field; it can be regarded as maximally non-committal to fluctuations in the data since the ME principle assumes the least about the label information apart from the information derived from the constraints. Furthermore, depending on the choice of λ^+ and λ^-, constraint violations are possible. For $\lambda^+ \to \infty, \lambda^- \to \infty$, the prior strictly enforces the constraints for the data in $\mathcal{X}^c$. We note that, procedures like EM or Deterministic Annealing require the computation of posterior assignment distributions for each single datum, i.e. the posterior over the assignment variables needs to be marginalized. Clearly, direct marginalization is only feasible for a small number of constraints, or when the constraints are highly decoupled.

In [3], the authors avoided the need to perform marginalization by resorting to a different, more greedy hill climbing heuristic, the *Iterative Conditional Mode (ICM)*. As the results in the experimental section indicate, the drawback of such a procedure is that it gets stuck very easily in poor local minima which is particularly dangerous in the context of constraint clustering. In order to use more sophisticated optimization techniques such as EM or DA, the problem of estimating marginalized posteriors can no longer be circumvented. In order to keep the optimization tractable, we approximate the posterior in the E-step by the mean field approximation.

Mean-Field Approximation for Posterior Inference Assume that the data given in $\mathcal{X}^c$ are independent, i.e. the data densities are factorial. By Bayes rule, we have

$$p(\mathbf{y}|\mathcal{X}^c) = \frac{1}{Z}\prod_i \exp\left(-h_i(y_i)\right) p(\mathbf{y}), \tag{3.3}$$

where, e.g., $h_i(y_i) = -\log p(\mathbf{x}_i|y_i)$ for Gaussian class-conditional densities or $h_i(y_i) = \|\mathbf{x}_i - \mu_{y_i}\|^2$ for the DA version of k-means.

In the mean field approximation, one tries to find a *factorial approximation*, the mean field approximation, $q(\mathbf{y}) = \prod_i q_i(y_i)$ of the posterior $p(\mathbf{y}|\mathcal{X}^c)$ such that the Kullback-Leibler divergence between the approximate and true posterior distributions is minimized, i.e.

$$\min_q \sum_{\mathbf{y}} q(\mathbf{y}) \log\left(\frac{q(\mathbf{y})}{p(\mathbf{y}|\mathcal{X}^c)}\right), \tag{3.4}$$

such that $\sum_\nu q_i(\nu) = 1$, for all i. Because the approximation is factorial, the computation of the marginalized posterior probabilities becomes feasible, a prerequisite to optimize the model efficiently. Note that the above KL divergence can be decomposed as

$$-H(q) - \mathrm{E}_q\left[\log p(\mathbf{y}|\mathcal{X}^c)\right] \tag{3.5}$$

where $H(q)$ denotes the entropy of the mean field approximation and E_q denotes the expectation w.r.t. q. We seek to minimize the expression in eq. (3.4) by looking for stationary points for the $q_i(\nu)$. Set $\gamma_{ij} = \lambda^+ a_{ij} - \lambda^- b_{ij}$ and $\Delta_{\nu,\mu} = 1 - \delta_{\nu,\mu}$, where $\delta_{\nu,\mu}$ is the Kronecker delta function. Using this convention, we can summarize the exponents in eq. (3.2) by $\gamma_{i,j}\Delta_{\nu,\mu}$ if $y_i = \nu$ and $y_j = \mu$. We want to emphasize that this approximation is only used for constrained data.

Taking the derivative of eq. (3.4) w.r.t $q_i(\nu)$ and setting it to zero leads to

$$q_i(\nu) = \frac{1}{Z_i} \exp\left(-h_i(\nu) - \sum_{j \neq i} \sum_{\mu} q_j(\mu)\gamma_{i,j}\Delta_{\nu,\mu}\right), \tag{3.6}$$

where

$$Z_i = \sum_{\nu} \exp\left(-h_i(\nu) - \sum_{j \neq i} \sum_{\mu} q_j(\mu)\gamma_{i,j}\Delta_{\nu,\mu}\right). \tag{3.7}$$

Since $\Delta_{\nu,\mu} = 1$ only if $\mu \neq \nu$, we can further simplify the expression for $q_i(\nu)$ to

$$q_i(\nu) = \frac{1}{Z_i} \exp\left(-h_i(\nu) - \sum_{j \neq i}(1 - q_j(\nu))\gamma_{i,j}\right). \tag{3.8}$$

Eventually, we have arrived at a factorial approximation of the marginal posterior probabilities. For the constrained data, these update equations can be used in the E-step for posterior probability estimation.

Model Fitting with Constraints So far, we have assumed that every data point in $\mathcal{X}^c$ participates in a constraint and we minimize the negative log-likelihood $-\log p(\mathcal{X}^c; \Theta, \mathcal{C})$ where $\mathcal{C}$ is used to denote the set of constraints. The constrained data and the unlabelled data can be integrated in a manner similar to eq. (2.1): suppose the given data $\mathcal{X}$ can be decomposed into unlabelled data $\mathcal{X}^u$ and data $\mathcal{X}^c$ that participate in pairwise assignment constraints. Furthermore, let $\alpha \in [0, 1]$. The same convex combination can be used

$$\min_{\Theta} \big(\alpha \mathcal{L}(\mathcal{X}^u; \Theta^u_{\text{prior}}, \Theta_{\text{model}}) + (1-\alpha)\mathcal{L}(\mathcal{X}^c; \Theta^c_{\text{prior}}, \Theta_{\text{model}}, \mathcal{C})\big), \tag{3.9}$$

which shifts the focus from pure posterior inference to improved parameter estimation. Similarly, (labelled + constrained) as well as (labelled + constrained + unlabelled) data can be combined into a single objective function. In particular, the optimal Θ can still be found by EM or DA, while allowing the inclusion of partially labelled as well as constrained data. The result of the minimization is a parameter estimate that takes all the available prior information into account. For the class-conditional densities, we arrive at a similar formula as we did in the semi-supervised case, e.g. for the means, we have

$$\mu_\nu = \frac{\alpha \sum_{\mathbf{x}_i \in \mathcal{X}^u} \rho_i(\nu)\mathbf{x} + (1-\alpha)\sum_{\mathbf{x}_i \in \mathcal{X}^c} q_i(\nu)\mathbf{x}}{\alpha \sum_{\mathbf{x}_i \in \mathcal{X}^u} \rho_i(\nu) + (1-\alpha)\sum_{\mathbf{x}_\in \mathcal{X}^c} q_i(\nu)}, \tag{3.10}$$

which amounts again to a convex combination of parameter estimates due to labelled and constrained data. Note, however, that the priors are different for unlabelled and constrained data.

The appropriate choice of α largely depends on what the user wants to achieve. If we set $\alpha = |\mathcal{X}^u|/|\mathcal{X}^c \cup \mathcal{X}^u|$ again, we assign equal importance to all the data points while for $\alpha = 1/2$ labelled and constrained data have the same importance in the inference. We can also use the search strategy mentioned above which controls α such that the number of constraint violations is minimized – in analogy to the minimization of the empirical risk (see section 2). Note that the coupling parameter α is different from the Lagrange parameters λ^+ and λ^-: α controls the importance of constrained data set $\mathcal{X}^c$ as opposed to unlabelled data set, while the Lagrange parameters λ^+ and λ^- only affect the data in $\mathcal{X}^c$.

4. Experimental Results

The approach described in section 3 is applied to deterministic annealing (DA) [17] for Gaussian mixture models and squared error clustering, leading to a DA clustering algorithm with constraints. This algorithm is tested on different synthetic and real world data sets. The clustering with constraints algorithms[2] by Shental *et al.* [19] and Basu *et al.* [3] are also run on all the data sets for comparison. For the algorithm in [3], both PCKMEANS and MPCKMEANS have been tried and they give nearly identical results for all data sets. Thirteen different constraint penalty values ranging from 1 to 4000 are used for the algorithm in [3]; only the best result of their algorithms is reported. In order to evaluate the results of the different methods, we use *F-scores*, i.e. the harmonic mean of precision and recall, to compare two classifications. Note, that an F-score of one amounts to perfect agreement of two solutions.

Figure 2 shows a 2D synthetic data set with 200 points, together with an example set of constraints. Since the horizontal separation between the point clouds is smaller than the vertical separation, the two-cluster unsupervised solution is to group the data into "upper" and "lower" clusters. With the presence of the constraints, however, the data points should be grouped into "left" and "right" clusters. The actual constraints are generated by first sampling point pairs randomly and then converting each pair to either a must-link or must-not-link constraint according to its location. Different levels of constraint information are considered: 1%, 5%, 10%, 15%, 30%, or 50% of constraints are considered relative to the total number of samples in the data set in order to account for the construction of the transitive closure on constraint graphs. We run the proposed

[2]We would like to thank the authors for putting the implementation of their algorithms online: `http://www.cs.huji.ac.il/~tomboy/code/ConstrainedEM_plusBNT.zip` for [19] and `http://www.cs.utexas.edu/users/ml/risc/code/` for [3].

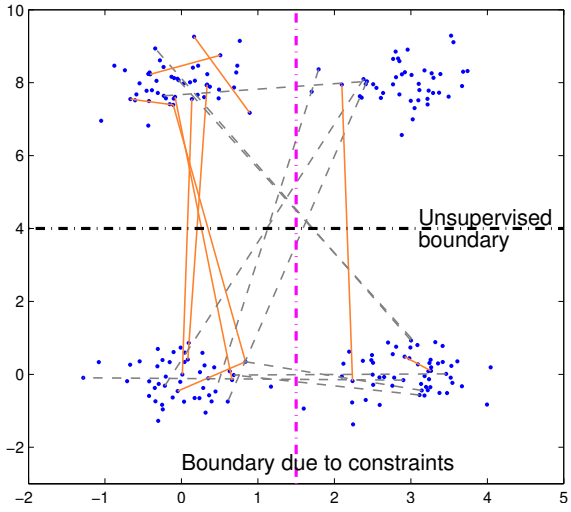


Figure 2. Synthetic data set. Solid lines: must-link constraints. Dashed lines: must-not-link constraints.

	1%	5%	10%	15%	30%	50%
Shental	0.540	0.560	0.560	0.580	0.535	1.0
Basu	0.540	0.545	0.535	0.590	0.535	1.0
Proposed	1.0	1.0	0.995	0.995	0.995	0.995

Table 2. F-scores of the toy data set.

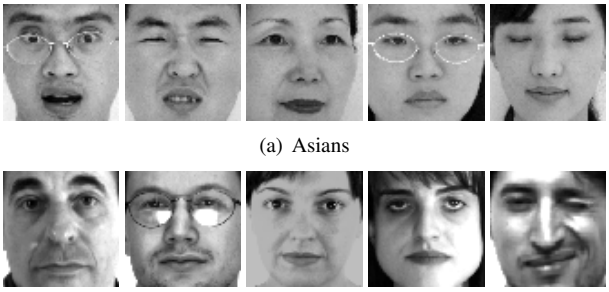
algorithm with $\lambda^+ = \lambda^- = 1000$ and recover the desired boundary almost exactly (with at most one erroneous point) for *all* levels of constraint information. On the contrary, the desired boundary is recovered by the algorithm in [19] and [3] only when 50% of constraints are present. The F-scores are shown in table 2. Note that a random grouping would have a F-score of 0.5 in this case. In order to demonstrate the effect of mis-specified constraints, we have randomly flipped 20% of the constraints for the 50% data set. The best result for the method in [19] is an F-score of 0.835. In contrast, our proposed method behaves favorably: the mis-specified constraints have hardly any effect of the decision boundary learnt and, hence, we obtain again an F-score of 0.995. We conclude that our approach is more robust towards erroneous constraints in this case.

Our second experiment is about an ethnicity classification problem [13], where the goal is to classify if a face image belongs to an Asian or not. The data set consists of 2630 images with size 64×64 from multiple databases, including the PF01 database[3], the Yale database[4], the AR database [15] and the non-public NLPR database[5]. Some example images are shown in Figure 3. A face image is represented by the first 30 eigenface coefficients. Again, different levels

[3] http://nova.postech.ac.kr/archives/imdb.html.

[4] http://cvc.yale.edu/projects/yalefaces/yalefaces.html.

[5] Provided by Dr. Yunhong Wang, National Laboratory for Pattern Recognition, Beijing.



(a) Asians

(b) Non-Asians

Figure 3. Example face images in the ethnicity classification problem.

	1%	10%	15%	30%
Shental	0.925	0.946	0.891	0.973
Basu	0.568	0.565	0.570	0.809
Proposed	0.923	0.915	0.922	0.963

Table 3. F-scores of the ethnicity classification problem by different clustering with constraints algorithms.

(1%, 10%, 15% and 30%) of constraint information (which has been derived from the known ground-truth labelling) are considered. The F-scores of different algorithms are shown in Table 3. We can see that the proposed algorithm significantly outperforms the algorithm by Basu *et al.* and is competitive with the algorithm by Shental *et al*.

Our third experiment is about the newsgroup data sets[6] used in [3]. It consists of three data sets, each of which contains roughly 300 documents from three different topics. The topics are regarded as the classes to be recovered. Latent semantic indexing is used to transform the term frequency and inverse document frequency normalized document vector to a 20D feature vector. Again, we have access to a ground-truth labelling of the data which we used to derive a varying number of constraints – as in the last two experiments. The F-scores are shown in table 4. We can see that the proposed algorithm is also very competitive: the method in [3] is outperformed on most problem instances again. We observe similar behavior on two of the three data sets in comparison with the approach in [19].

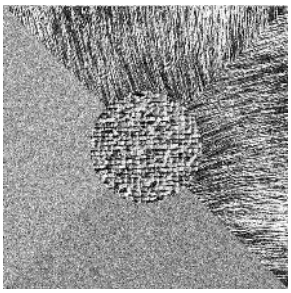
Our final experiment is on an image segmentation task. We use a Mondrian image (Figure 4(a)) consisting of five regions: three regions with strong texture, and two regions of very noisy gray-level segments, are to be identified. This 512 by 512 image is divided into a 101-by-101 grid. A 24-dimensional feature vector is extracted for each site: 12 features originate from a 12-bin histogram of gray-level values, while the remaining 12 correspond to the averages of Gabor filter responses for four orientations at three different scales at each site. The segment labels of different sites are

[6] http://www.cs.utexas.edu/users/ml/risc/.

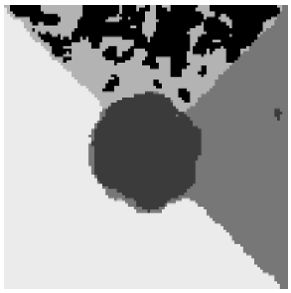
Data set		1%	10%	15%	30%
same-300	Shental	0.412	0.429	0.516	0.487
	Basu	0.515	0.459	0.472	0.552
	Proposed	0.491	0.588	0.527	0.507
similar-300	Shental	0.560	0.553	0.531	0.532
	Basu	0.515	0.492	0.549	0.530
	Proposed	0.54	0.54	0.53	0.514
diff-300	Shental	0.877	0.554	0.907	0.871
	Basu	0.677	0.582	0.558	0.608
	Proposed	0.533	0.658	0.571	0.594

Table 4. F-scores of the newsgroup data sets with different numbers of constraints.

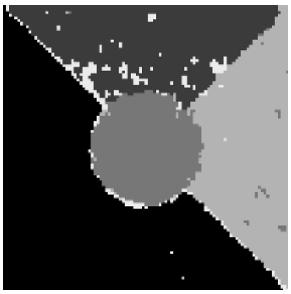
generated from a ground-truth image. Since, the texture information dominates the gray value information, clustering with unlabelled data fails to recover the ground-truth information. This also holds true for the data set with 1% and 5% of the data in constraints (see figure 4(c)). The segmented image with 10% of sites in constraints is shown in figure 4(d). Here, we almost perfectly identify the ground-truth information, since the algorithm is able to distinguish between the gray-level segments. The F-scores by various algorithms are listed in Table 5. The proposed method holds an edge when at least 10% of data are in constraints, and it can discover the desired segmentation (approximately) with the least amount of constraint information. The quality gap is particularly large in this case. Furthermore, the approach in [19] had a very high running time, in particular for examples with a large number of constraints.



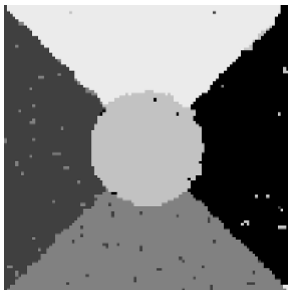
(a) Original Image



(b) Segmentation, no constraints



(c) Segmentation, 1% and 5% pixels in constraints



(d) Segmentation, 10% pixels in constraints

Figure 4. Results of image segmentation. (a): source image. (b) to (d): segmentation results with different numbers of constraints.

	1%	5%	10%	15%
Shental	0.830	0.831	0.840	0.829
Basu	0.761	0.801	0.821	0.776
Proposed	0.772	0.829	0.972	0.98

Table 5. F-scores of the Image segmentation task.

5. Conclusion

The traditional boundary between supervised and unsupervised learning has been blurred by the recent advances of learning with partially labelled and constrained data. In this paper, we have proposed a general framework for incorporating different sources of information in learning classifications of a data set, with supervised learning and unsupervised learning arising as special cases. By inferring on the parameters instead of the posterior assignment probabilities, we avoid the pitfall of inconsistent labelling rules. This combination approach also decouples the learning task; instead of worrying about how different types of information interact with each other, we can focus on building the most appropriate model for a single source of information (constraints in our case). To this end, we adopt the maximum entropy principle to derive a prior distribution for the assignment labels. The maximum entropy principle assumes the least about the label information apart from the information derived from the constraints. The mean field approximation technique is adopted to keep the computation tractable: the computation requirement in each iteration is similar to that of a standard EM iteration. This can be much more efficient than the algorithm in [19] in the case where the constraints lead to a large clique in the corresponding graphical model. The factorial distribution due to mean field approximation is a stationary point of the variational free energy and, thereby, aims at finding the best factorial distribution in terms of the Kullback-Leibler divergence to the true distribution. The use of deterministic annealing in our approach has avoided getting trapped in poor local minima, which can be the case for the ICM technique used in [3]. This is particularly valuable in clustering with constraints, where the energy landscape can be more "rugged" than in standard clustering tasks.

There are several avenues for future work. The proposed framework can be applied to clustering with other types of Bregman divergence, such as histogram clustering. We can also consider information other than partial labels and constraints. Finally, we would like to investigate more about the interplay between unlabelled, labelled and constraint information in both theoretical and practical sense.

Acknowledgement

Anil Jain and Martin Law are supported by ONR contract # N00014-01-0266. The authors would like to thank Dr. Yunhong Wang for providing the NLPR face database, and Xiaoguang Lu for introducing us to this database. Tilman Lange would like to thank the CSE Department of the Michigan State University for its hospitality.

References

[1] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh. Clustering with bregman divergences. In *Proc. SIAM International Conference on Data Mining*, pages 234–245, Apr. 2004.

[2] S. Basu, A. Banerjee, and R. J. Mooney. Active semi-supervision for pairwise constrained clustering. In *Proceedings of the SIAM International Conference on Data Mining*, pages 333–344, 2004.

[3] S. Basu, M. Bilenko, and R. Mooney. A probabilistic framework for semi-supervised clustering. In *Proceedings of the 10th ACM SIGKDD, International Conference on Knowledge Discovery and Data Mining*, pages 59–68, 2004.

[4] M. Bilenko, S. Basu, and R. J. Mooney. Integrating constraints and metric learning in semi-supervised clustering. In *Proceedings of the 21st International Conference on Machine Learning*, pages 81–88, 2004.

[5] T. Hofmann and J. Puzicha. Statistical models for co-occurrence data. Technical Report AIM-1625, MIT, 1998.

[6] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, Sept. 1999.

[7] E. T. Jaynes. Information theory and statistical mechanics i. *Phys. Rev.*, 106:620–630, 1957.

[8] E. T. Jaynes. Information theory and statistical mechanics ii. *Phys. Rev.*, 108:171–190, 1957.

[9] S. Kamvar, D. Klein, and C. D. Manning. Spectral learning. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 561–566, 2003.

[10] D. Klein, S. D. Kamvar, and C. D. Manning. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In *Proc. the 19th International Conference on Machine Learning*, pages 307–314, 2002.

[11] M. Law, A. Topchy, and A. Jain. Model-based clustering with probabilistic constraints. In *Proc. SIAM International Conference on Data Mining*, 2005.

[12] M. H. Law, A. Topchy, and A. K. Jain. Clustering with soft and group constraints. In *Proc. the Joint IAPR International Workshops on Structural, Syntactic, And Statistical Pattern Recognition*, pages 662–670, Lisbon, Portugal, Aug. 2004.

[13] X. Lu and A. K. Jain. Ethnicity identification from face images. In *Proceedings of SPIE*, volume 5404, pages 114–123, 2004.

[14] Z. Lu and T. Leen. Semi-supervised learning with penalized probabilistic clustering. In *Advances in Neural Information Processing Systems 17*, pages 849–856. MIT Press, 2005.

[15] A. Martinez and R. Benavente. The AR face database. Technical Report 24, CVC, 1998. `http://rvl1.ecn.purdue.edu/~aleix/aleix_face_DB.html`.

[16] G. McLachlan and K. Basford. *Mixture Models: Inference and Application to Clustering*. Marcel Dekker, New York, 1988.

[17] K. Rose, E. Gurewitz, and G. Fox. Vector quantization and deterministic annealing. *IEEE Trans. Information Theory*, 38(4):1249–1257, 1992.

[18] V. Roth, J. Laub, M. Kawanabe, and J. M. Buhmann. Optimal cluster preserving embedding of nonmetric proximity data. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(12):1540–1551, December 2003.

[19] N. Shental, A. Bar-Hillel, T. Hertz, and D. Weinshall. Computing gaussian mixture models with EM using equivalence constraints. In *Advances in Neural Information Processing Systems 16*. MIT Press, 2004.

[20] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 2nd edition, 1999.

[21] K. Wagstaff and C. Cardie. Clustering with instance-level constraints. In *Proc. the 17th International Conference on Machine Learning*, pages 1103–1110, 2000.

[22] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl. Constrained k-means clustering with background knowledge. In *Proc. the 18th International Conference on Machine Learning*, pages 577–584, 2001.

[23] S. X. Yu and J. Shi. Segmentation given partial grouping constraints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):173–183, 2004.