Proceedings of the ICML-2004 Workshop on Statistical Relational Learning and its Connections to Other Fields (SRL-2004), Banff, Canada, July 2004

A Comparison of Inference Techniques for Semi-supervised Clustering with Hidden Markov Random Fields

Mikhail Bilenko MBILENKO@CS.UTEXAS.EDU
Sugato Basu SUGATO@CS.UTEXAS.EDU
Department of Computer Sciences, University of Texas at Austin, 1 University Station C0500, Austin, TX 78712 USA

Abstract

Recently, a number of methods have been proposed for semi-supervised clustering that employ supervision in the form of pairwise constraints. We describe a probabilistic model for semi-supervised clustering based on Hidden Markov Random Fields (HMRFs) that incorporates relational supervision. The model leads to an EM-style clustering algorithm, the E-step of which requires collective assignment of instances to cluster centroids under the constraints. We evaluate three known techniques for such collective assignment: belief propagation, linear programming relaxation, and iterated conditional modes (ICM). The first two methods attempt to globally approximate the optimal assignment, while ICM is a greedy method. Experimental results indicate that global methods outperform the greedy approach when relational supervision is limited, while their benefits diminish as more pairwise constraints are provided.

1. Introduction

There has been significant recent interest in *semi-supervised clustering*, where the goal is to improve the performance of unsupervised clustering algorithms with limited amounts of supervision in the form of labels or pairwise constraints on the data points (Wagstaff et al., 2001; Klein et al., 2002; Xing et al., 2003; Bilenko et al., 2004). In this paper, we present a probabilistic model for semi-supervised clustering with pairwise relations and compare the performance of several inference methods for cluster assignment in the context of an EM-based algorithm.

Existing methods for semi-supervised clustering fall into two general categories that we call *constraint-based* and *distance-based*. Constraint-based methods rely on user-provided labels or relational constraints to guide the algorithm towards a more appropriate data partitioning (Wagstaff et al., 2001). In distance-based approaches, an existing clustering algorithm that uses a particular clustering distortion measure is employed, but the measure is trained to satisfy the labels or constraints in the given supervised data (Klein et al., 2002; Xing et al., 2003; Cohn et al., 2003; Bar-Hillel et al., 2003). In Bilenko et al. (2004), we have proposed an integrated framework for semi-supervised clustering that combines the constraint-based and distance-based approaches in a unified probabilistic model.

We have recently shown that this semi-supervised clustering framework has an underlying probabilistic model – a Hidden Markov Random Field (HMRF) (Basu et al., 2004). Then, minimizing the integrated clustering objective function becomes equivalent to finding the maximum *a posteriori* probability (MAP) configuration of the HMRF (Zhang et al., 2001). It can be shown that the HMRF clustering model is able to incorporate any Bregman divergence (Banerjee et al., 2004) as the clustering distortion measure, which allows using the framework with such common distortion measures as KL-divergence, I-divergence, and parameterized squared Mahalanobis distance. Additionally, cosine similarity can also be used as the clustering distortion measure in the framework, which makes it useful for directional datasets.

The HMRF semi-supervised clustering model suggests an EM-based algorithm that minimizes the integrated clustering objective function to obtain a partitioning of the dataset. The E-step of the algorithm can be mapped to an inference step of a probabilistic relational model. In prior work, we used a fast greedy iterated conditional modes (ICM) inference technique in the E-step (Bilenko et al., 2004; Basu et al., 2004). Here, we compare ICM to two global approximate inference techniques for relational models: belief propagation and linear programming (LP) relaxation. Our experiments reveal that, when provided with sufficient relational supervision, ICM produces results comparable to belief propagation and LP relaxation on the constraint-based semi-supervised clustering task at a fraction of computational cost.

2. Background

2.1. The HMRF Clustering Framework

Given a set of data points $\mathcal{X}$, sets of pairwise must-link constraints $\mathcal{M}$ and cannot-link constraints $\mathcal{C}$ with sets of asso-

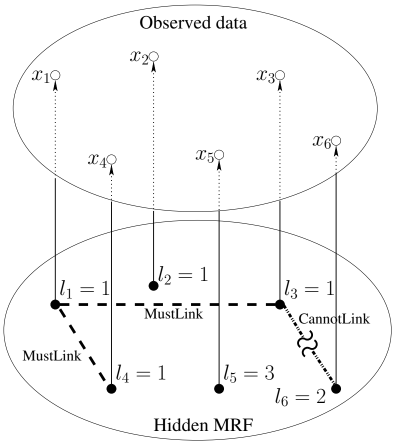

Figure 1. A Hidden Markov Random Field

ciated violation costs $\mathcal{W}$ and $\overline{\mathcal{W}}$, and a distortion measure D between the points, the semi-supervised clustering task is to optimally partition $\mathcal{X}$ into K clusters.[1] An optimal partitioning is that which minimizes the total distortion between the points and their cluster representatives according to D, while keeping constraint violations to a minimum.

This problem can be formalized as the task of label assignment in a Hidden Markov Random Field (HMRF) (Basu et al., 2004). The HMRF model consists of (1) a *hidden* field $\mathcal{L} = \{l_i\}_{i=1}^N$ of random variables that encode cluster assignments of data points; and (2) an *observable* set $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$ of random variables that correspond to observed data points. Every data point $\mathbf{x}_i$ is assumed to be generated from a conditional probability distribution $\mathbf{Pr}(\mathbf{x}_i|l_i)$ determined by the value of the corresponding hidden variable l_i. Random variables $\mathcal{X}$ are conditionally independent given the hidden variables $\mathcal{L}$, i.e., $\mathbf{Pr}(\mathcal{X}|\mathcal{L}) = \prod_{\mathbf{x}_i \in \mathcal{X}} \mathbf{Pr}(\mathbf{x}_i|l_i)$.

Note that a relational model similar to HMRFs has been proposed by Segal et al. (2003) for semi-supervised clustering with constraints, except that it only utilized must-link constraints and did not incorporate learnable distortion measures.

Fig. 1 shows a simple example of an HMRF. The observed dataset $\mathcal{X}$ consists of six points $\{\mathbf{x}_1 \ldots \mathbf{x}_6\}$ with corresponding cluster label variables $\{l_1 \ldots l_6\}$. Two must-link constraints are provided between (l_1,l_3) and (l_1,l_4), and one cannot-link constraint is provided between (l_3,l_6). The task is to partition the six points into three clusters. One clustering configuration is shown in Fig. 1, where the must-linked points $\mathbf{x}_1, \mathbf{x}_3$ and $\mathbf{x}_4$ are put in cluster 1; the point $\mathbf{x}_6$, which is cannot-linked to $\mathbf{x}_3$, is assigned to cluster 2; $\mathbf{x}_2$ and $\mathbf{x}_5$, which are not involved in any constraints, are put in clusters 1 and 3 respectively.

Every hidden random variable l_i has an associated set of neighbors $\mathcal{N}_i$. The must-link constraints $\mathcal{M}$ and cannot-link constraints $\mathcal{C}$ define the neighborhood over each hidden variable to be all the points that are must-linked or cannot-linked to the corresponding data point. A Markov Random Field is then defined over the hidden variables.

Let us consider a particular cluster label configuration $\mathcal{L}$ to be the joint event $\mathcal{L} = \{l_i\}_{i=1}^N$, which corresponds to a specific assignment of data points to clusters. By the Hammersley-Clifford theorem, the probability of a label configuration in the Markov Random Field can be expressed as a Gibbs distribution (Geman & Geman, 1984):

$$\mathbf{Pr}(\mathcal{L}) = \frac{1}{Z} e^{-V(\mathcal{L})} = \frac{1}{Z} e^{-\sum_{\mathcal{N}_i \in \mathcal{N}} V_{\mathcal{N}_i}(\mathcal{L})} \tag{1}$$

where $\mathcal{N}$ is the set of all neighborhoods, Z is a normalizing constant, and $V(\mathcal{L})$ is the overall label configuration potential function, which can be decomposed into the functions $V_{\mathcal{N}_i}(\mathcal{L})$ denoting the potential for every neighborhood $\mathcal{N}_i$.

Since we are provided with pairwise constraints over the class labels, we restrict the MRFs over the hidden variables to have pairwise potentials. The prior probability of a configuration of cluster labels $\mathcal{L}$ then becomes $\mathbf{Pr}(\mathcal{L}) = \frac{1}{Z} \exp(-\sum_i \sum_j V(i,j))$, where

$$V(i,j) = \begin{cases} f_M(\mathbf{x}_i, \mathbf{x}_j) & \text{if } (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{M} \\ f_C(\mathbf{x}_i, \mathbf{x}_j) & \text{if } (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{C} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Here, $f_M(\mathbf{x}_i, \mathbf{x}_j)$ is a non-negative function that penalizes the violation of a must-link constraint, and $f_C(\mathbf{x}_i, \mathbf{x}_j)$ is the corresponding penalty function for cannot-links. Intuitively, this form of $\mathbf{Pr}(\mathcal{L})$ gives higher probabilities to label configurations that attempt to satisfy the must-link constraints $\mathcal{M}$ and cannot-link constraints $\mathcal{C}$.

It is possible to use a learnable distortion measure D that adapts distance computations to respect the user-provided constraints. To facilitate learning of distortion measure parameters, the penalty for violating a must-link constraint between *distant* points should be higher than that between *nearby* points. This reflects the fact that if two must-linked points are far apart according to the current distortion measure, the distance measure parameters need to be modified to bring those points closer together. Inversely, the penalty for violating a cannot-link constraint between two points that are *nearby* according to the current distance measure should be higher than for two *distant* points. To reflect this reasoning, the penalty functions are chosen as follows:

$$f_M(\mathbf{x}_i, \mathbf{x}_j) = w_{ij} \varphi_D(\mathbf{x}_i, \mathbf{x}_j) \mathbb{1}[l_i \neq l_j] \tag{3}$$

$$f_C(\mathbf{x}_i, \mathbf{x}_j) = \overline{w}_{ij} \big(\varphi_{D\max} - \varphi_D(\mathbf{x}_i, \mathbf{x}_j)\big) \mathbb{1}[l_i = l_j] \tag{4}$$

where φ_D is a monotonically increasing *penalty scaling* function of the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ (typically equivalent to the distortion measure D), and $\varphi_{D\max}$ is the maximum value of φ_D for the dataset. This form of f_C ensures

[1] Must-link and cannot-link constraints specify pairs of points that should be in same and different clusters respectively.

that the penalty for violating a cannot-link constraint remains non-negative. Note that the resulting potential function V corresponds to a metric version of previously described generalized Potts potential (Kleinberg & Tardos, 1999).

The overall posterior probability of a cluster label configuration $\mathcal{L}$ is $\mathbf{Pr}(\mathcal{L}|\mathcal{X}) \propto \mathbf{Pr}(\mathcal{L})\mathbf{Pr}(\mathcal{X}|\mathcal{L})$, assuming $\mathbf{Pr}(\mathcal{X})$ to be constant. We consider $\mathbf{Pr}(\mathcal{X}|\mathcal{L})$ to have an exponential form, which encompasses most commonly used distortion measures such as squared Euclidean distance, KL divergence, and cosine similarity (Basu et al., 2004). Finding the maximum *a posteriori* (MAP) configuration of the HMRF then becomes equivalent to maximizing the posterior probability:

$$\mathbf{Pr}(\mathcal{L}|\mathcal{X}) = \frac{1}{Z} e^{-\sum_i \sum_j V(i,j)} \cdot e^{-\sum_{x_i \in \mathcal{X}} D(\mathbf{x}_i, \mu_{l_i})} \tag{5}$$

where Z is a normalizing constant. Henceforth, we will refer to the first exponential factor of $\mathbf{Pr}(\mathcal{L}|\mathcal{X})$ as the *constraint potential*, the second factor as the *distance potential*, and the negative logarithm of $\mathbf{Pr}(\mathcal{L}|\mathcal{X})$ as the *posterior energy*. Note that MAP estimation would reduce to maximum likelihood (ML) estimation of $\mathbf{Pr}(\mathcal{X}|\mathcal{L})$ if $\mathbf{Pr}(\mathcal{L})$ is constant. However, because our model accounts for dependencies between the cluster labels and $\mathbf{Pr}(\mathcal{L})$ is not constant, full MAP estimation of $\mathbf{Pr}(\mathcal{L}|\mathcal{X})$ is required.

Taking logarithms of Eqn.(5) gives the following cluster objective function, minimizing which is equivalent to maximizing the MAP probability in Eqn.(5), or, equivalently, minimizing the posterior energy of the HMRF:

$$\begin{aligned}\mathcal{J}_{\text{obj}} = &\sum_{x_i \in \mathcal{X}} D(\mathbf{x}_i, \mu_{l_i}) + \sum_{(x_i,x_j) \in \mathcal{M}} w_{ij} \varphi_D(\mathbf{x}_i, \mathbf{x}_j) \mathbb{1}[l_i \neq l_j] \\ &+ \sum_{(x_i,x_j) \in \mathcal{C}} \overline{w}_{ij} \big(\varphi_{D\max} - \varphi_D(\mathbf{x}_i, \mathbf{x}_j)\big) \mathbb{1}[l_i = l_j] + \log Z \end{aligned} \tag{6}$$

where Z is a normalizing constant. Thus, the task is to minimize $\mathcal{J}_{\text{obj}}$ over $\{\mu_h\}_{h=1}^K$, $\mathcal{L}$, and D (if the latter is parameterized). For computational efficiency, we consider $\log Z$ to be constant during the clustering iterations.

The *Expectation-Maximization* (EM) algorithm is a popular solution to such "incomplete data" problems (Dempster et al., 1977). It is well-known that K-Means is equivalent to an EM algorithm with hard clustering assignments (Kearns et al., 1997; Banerjee et al., 2004). Thus, we can use a K-Means-type iterative clustering algorithm, HMRF-KMEANS, to find a (local) maximum of the above function.

2.2. The HMRF-KMEANS Algorithm

The outline of the HMRF-KMEANS algorithm is presented in Figure 2. Initialization is performed using neighborhoods inferred from the constraints as described in (Bilenko et al., 2004). The basic idea of HMRF-KMEANS is as follows: in the E-step, given the current cluster representatives, all data points are collectively re-assigned to clusters to minimize $\mathcal{J}_{\text{obj}}$. In the M-step, the cluster representatives $\{\mu_h\}_{h=1}^K$ are re-estimated from the cluster assignments to minimize $\mathcal{J}_{\text{obj}}$ for the current assignment. The clustering distortion measure D is updated in the M-step to reduce the objective function simultaneously by transforming the space in which data lies, thus performing metric learning.

Note that this procedure is an instantiation of the generalized EM algorithm (Dempster et al., 1977; Neal & Hinton, 1998), where the objective function is reduced but not necessarily minimized in the M-step. Effectively, the E-step minimizes $\mathcal{J}_{\text{obj}}$ over cluster assignments $\mathcal{L}$, the M-step (A) minimizes $\mathcal{J}_{\text{obj}}$ over cluster representatives $\{\mu_h\}_{h=1}^K$, and the M-step (B) minimizes $\mathcal{J}_{\text{obj}}$ over the parameters of the distortion measure D. The E-step and the M-step are repeated till a specified convergence criterion is reached.

Algorithm: HMRF-KMEANS
Input: Set of data points $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$, number of clusters K,
set of *must-link* constraints $\mathcal{M} = \{(\mathbf{x}_i, \mathbf{x}_j)\}$,
set of *cannot-link* constraints $\mathcal{C} = \{(\mathbf{x}_i, \mathbf{x}_j)\}$,
distance measure D, constraint violation costs $\mathcal{W}$ and $\overline{\mathcal{W}}$.
Output: Disjoint K-partitioning $\{\mathcal{X}_h\}_{h=1}^K$ of $\mathcal{X}$ such that objective function $\mathcal{J}_{\text{obj}}$ in Eqn.(6) is (locally) minimized.
Method:
1. Initialize the K clusters centroids $\{\mu_h^{(0)}\}_{h=1}^K$, set t $\leftarrow$ 0
2. Repeat until *convergence*
2a. `E-step`: Given $\{\mu_h^{(t)}\}_{h=1}^K$, re-assign cluster labels $\{l_i^{(t+1)}\}_{i=1}^N$ on the points $\{\mathbf{x}_i\}_{i=1}^N$ to minimize $\mathcal{J}_{\text{obj}}$.
2b. `M-step(A)`: Given cluster labels $\{l_i^{(t+1)}\}_{i=1}^N$, re-calculate cluster centroids $\{\mu_h^{(t+1)}\}_{h=1}^K$ to minimize $\mathcal{J}_{\text{obj}}$.
2c. `M-step(B)`: Re-estimate distance measure D to reduce $\mathcal{J}_{\text{obj}}$.
2d. t $\leftarrow$ t+1

Figure 2. The HMRF-KMEANS algorithm

The relational nature of the supervision is significant in the E-step, where the task is to assign data points to clusters using the current estimates of the cluster representatives. In simple K-Means there is no interaction between the cluster labels, and the E-step is a simple assignment of every point to the cluster representative that is nearest to it according to the clustering distortion measure. In contrast, the HMRF model incorporates interaction between the cluster labels defined by the random field over the hidden variables. Thus, optimal assignment in the E-step of HMRF-KMEANS is a relational inference problem.

3. Inference Techniques

Below we describe three inference methods for collective assignment of data points to clusters in the E-step of HMRF-KMEANS.

3.1. The Belief Propagation Approach

A global joint assignment of the points to clusters that (locally) minimizes the objective function $\mathcal{J}_{\text{obj}}$ can be found

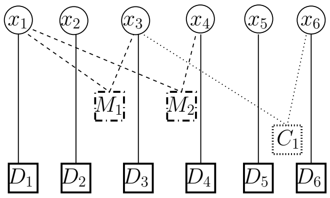

Figure 3. The Factor Graph for the HMRF in Figure 1

by performing approximate inference on the HMRF using belief propagation (Pearl, 1988). This approach is similar to the technique used by Segal et al. (2003).

To implement the message passing algorithm for approximate inference on the HMRF, we represent the HMRF as a factor graph model (Kschischang et al., 2001). The sum-product/max-product algorithm on the factor graph model has been shown to be a generalization of several well known inference algorithms on graphical models. Interpreting the HMRF model as a factor graph enables us to perform belief propagation on the HMRF using the max-product message passing algorithm on the corresponding factor graph.

The factor graph corresponding to the example HMRF in Figure 1 is shown in Figure 3. The factor graph has the following components:

(1) N variable nodes $\{\mathbf{x}_i\}_{i=1}^N$ representing the data points.

(2) N factor nodes $\{D_i\}_{i=1}^N$ that encode the distance potential components of the objective function. Each distance factor node D_i has an edge connecting it to the corresponding variable node $\mathbf{x}_i$, and a table containing different values of the distance potential function. This table has an entry for each possible cluster assignment of the variable; the j^{th} entry is $\exp(-d)$, where d is the distance from the i^{th} point to the j^{th} cluster.

(3) $|\mathcal{M}|$ factor nodes $\{M_i\}_{i=1}^{|\mathcal{M}|}$ and $|\mathcal{C}|$ factor nodes $\{C_i\}_{i=1}^{|\mathcal{C}|}$, which encode the cost of violating the must-link and cannot-link constraints, respectively. There is one factor node for each constraint, which is linked by edges to the 2 variable nodes involved in that constraint.

The constraint potential table associated with each constraint factor node maps a set of K^2 value-pairs (corresponding to possible cluster assignments to the pair of points in the constraint) to potential values. For the factor node encoding the must-link constraint $(\mathbf{x}_i, \mathbf{x}_j)$, the potential value for the entry (l_i, l_j) in the constraint potential table is 1 if $l_i = l_j$, i.e., $\mathbf{x}_i$ and $\mathbf{x}_j$ have the same cluster assignments. If $l_i \neq l_j$, the potential value is $\exp(-d(\mathbf{x}_i, \mathbf{x}_j)w_{ij})$, where $d(\mathbf{x}_i, \mathbf{x}_j)$ is the distance between the points $\mathbf{x}_i$ and $\mathbf{x}_j$ according to the current estimate of the distortion measure D, and w_{ij} is the weight of the constraint.

Similarly, for the cannot-link factor nodes, the potential tables have values of 1 for the entry (l_i, l_j) where $l_i \neq l_j$, and $\exp(-(d_{\max}(\mathbf{x}_i, \mathbf{x}_j) - d(\mathbf{x}_i, \mathbf{x}_j))\overline{w}_{ij})$ if $l_i = l_j$. The potential values of the constraint factor nodes correspond to the metric version of the Potts potential function, as explained in Section 2.1.

Finding the collective assignment of points to minimize $\mathcal{J}_{\text{obj}}$ in the E-step corresponds to running the max-product message-passing algorithm on the factor graph (Kschischang et al., 2001). Once the message-passing algorithm converges, the cluster assignment for each data point is obtained from the value in the corresponding variable node.

3.2. The Iterated Conditional Modes Approach

The Iterated Conditional Modes (ICM) inference technique (Besag, 1986) is a greedy strategy to sequentially update the cluster assignment of each point, keeping the assignments for the other points fixed. Based on a preselected random ordering, each point $\mathbf{x}_i$ is sequentially assigned to the cluster representative μ_h that minimizes the point's contribution to the objective function $\mathcal{J}_{obj}(\mathbf{x}_i, \mu_h)$:

$$\mathcal{J}_{\text{obj}}(\mathbf{x}_i, \mu_h) = D(\mathbf{x}_i, \mu_h) + \sum_{(x_i,x_j)\in\mathcal{M}} w_{ij}\varphi_D(\mathbf{x}_i, \mathbf{x}_j)\mathbb{1}[h \neq l_j] + \sum_{(x_i,x_j)\in\mathcal{C}} \overline{w}_{ij}\big(\varphi_{D\max} - \varphi_D(\mathbf{x}_i, \mathbf{x}_j)\big)\mathbb{1}[h = l_j] \quad (7)$$

Optimal assignment for every point is that which minimizes the distortion between the point and its cluster representative (first term of $\mathcal{J}_{\text{obj}}$) along with incurring a minimal penalty for constraint violations caused by this assignment (second and third terms of $\mathcal{J}_{\text{obj}}$). Once a point $\mathbf{x}$ is assigned to a cluster, the subsequent points in the sequence determined by the ordering use the current cluster assignment of $\mathbf{x}$ to calculate possible constraint violations.

After all points are assigned, the assignment process is repeated according to a new random ordering. This process proceeds until no point changes its cluster assignment between two successive iterations. ICM is guaranteed to reduce $\mathcal{J}_{\text{obj}}$ or keep it unchanged (if $\mathcal{J}_{\text{obj}}$ is already at a local minimum) in the E-step (Besag, 1986).

3.3. The LP Relaxation Approach

The task of finding an assignment of instances to clusters to minimize the objective function can be posed as an integer programming problem. Such a formulation has been proposed by Kleinberg and Tardos in the context of the general *metric labeling* problem, where they considered the cost of assigning labels to instances while attempting to satisfy a set of must-link pairwise constraints (Kleinberg & Tardos, 1999). We extend this formulation to include cannot-link constraints, which allows using it for assigning instances to clusters in the E-step of HMRF-KMEANS.

Let $\mathcal{Y} = \{y_{il}\}$, $i = 1..N$, $l = 1..K$, be a set of nonnegative binary variables that encode membership of instances in clusters: $y_{il} = 1$ signifies that the i^{th} instance belongs to the l^{th} cluster. Sets of nonnegative binary variables $\mathcal{Y}^{(\mathcal{M})} = \{y_i^{(M)}\}_{i=1}^{|\mathcal{M}|}$ and $\mathcal{Y}^{(\mathcal{C})} = \{y_i^{(C)}\}_{i=1}^{|\mathcal{C}|}$ encode violations of must-link and cannot-link pairwise constraints respectively. Each $y_k^{(M)} = 1$ signifies that the k^{th} must-link pairwise constraint $e_k = (\mathbf{x}_{k_1}, \mathbf{x}_{k_2})$ is violated, while $y_k^{(C)} = 1$ signifies that the k^{th} cannot-link pairwise constraint $e_k = (\mathbf{x}_{k_1}, \mathbf{x}_{k_2})$ is violated. The objective function to be optimized in the E-step of HMRF-KMEANS then becomes:

$$\mathcal{J}_{\text{obj}} = \sum_{\mathbf{x}_i \in \mathcal{X}} \sum_{l \in \mathcal{L}} D(\mathbf{x}_i, \mu_l)\, y_{il} + \sum_{(\mathbf{x}_{k_1}, \mathbf{x}_{k_2}) \in \mathcal{M}} w_k f_M(\mathbf{x}_{k_1}, \mathbf{x}_{k_2}) y_k^{(M)} + \sum_{(\mathbf{x}_{k_1}, \mathbf{x}_{k_2}) \in \mathcal{C}} \overline{w}_k f_C(\mathbf{x}_{k_1}, \mathbf{x}_{k_2}) y_k^{(C)} \quad (8)$$

Assigning each instance to only one cluster imposes the following linear constraint on variables in $\mathcal{Y}$:

$$\sum_{l \in \mathcal{L}} y_{il} = 1, \quad \mathbf{x}_i \in \mathcal{X} \quad (9)$$

Also, consistency of pairwise constraint violation variables in $\mathcal{Y}^{(\mathcal{M})}$ and $\mathcal{Y}^{(\mathcal{C})}$ with the assignment variables in $\mathcal{Y}$ requires satisfaction of the following linear constraints:

$$\begin{aligned} y_k^{(M)} &= \frac{1}{2} \sum_{l \in \mathcal{L}} |y_{k_1 l} - y_{k_2 l}|, \quad & e_k = (\mathbf{x}_{k_1}, \mathbf{x}_{k_2}) \in \mathcal{M}; \\ y_k^{(C)} &= 1 - \frac{1}{2} \sum_{l \in \mathcal{L}} |y_{k_1 l} - y_{k_2 l}|, \quad & e_k = (\mathbf{x}_{k_1}, \mathbf{x}_{k_2}) \in \mathcal{C} \end{aligned} \quad (10)$$

These constraints can be expressed in a linear program by replacing variables $\mathcal{Y}^{(\mathcal{M})}$ and $\mathcal{Y}^{(\mathcal{C})}$ with corresponding sets of auxiliary variables $\mathcal{Z}^{(\mathcal{M})}$ and $\mathcal{Z}^{(\mathcal{C})}$, where $z_{kl}^{(M)} = 1$ iff the k^{th} must-link pair $e_k = (\mathbf{x}_{k_1}, \mathbf{x}_{k_2})$ is violated and either $\mathbf{x}_{k_1}$ or $\mathbf{x}_{k_2}$ is assigned to l^{th} cluster. Semantics of $z_{kl}^{(C)}$ are similar: $z_{kl}^{(C)} = 1$ iff k^{th} cannot-link pair $e_k = (\mathbf{x}_{k_1}, \mathbf{x}_{k_2})$ is violated and both $\mathbf{x}_{k_1}$ and $\mathbf{x}_{k_2}$ are assigned to l^{th} cluster. Variables in $\mathcal{Y}^{(\mathcal{M})}$ and $\mathcal{Y}^{(\mathcal{C})}$ can be expressed via variables in $\mathcal{Z}^{(\mathcal{M})}$ and $\mathcal{Z}^{(\mathcal{C})}$ as follows:

$$\begin{aligned} y_k^{(M)} &= \frac{1}{2} \sum_{l \in \mathcal{L}} z_{kl}^{(M)}, \quad & e_k = (\mathbf{x}_{k_1}, \mathbf{x}_{k_2}) \in \mathcal{M} \\ y_k^{(C)} &= \sum_{l \in \mathcal{L}} z_{kl}^{(C)}, \quad & e_k = (\mathbf{x}_{k_1}, \mathbf{x}_{k_2}) \in \mathcal{C} \end{aligned} \quad (11)$$

Consistency of assignment variables in $\mathcal{Y}$ with pairwise constraint violation variables in $\mathcal{Z}^{(\mathcal{M})}$ and $\mathcal{Z}^{(\mathcal{C})}$ can then be achieved by introducing the following linear constraints:

$$z_{kl}^{(M)} \geq y_{k_1 l} - y_{k_2 l}, \quad e_k = (\mathbf{x}_{k_1}, \mathbf{x}_{k_2}) \in \mathcal{M} \quad (12)$$

$$z_{kl}^{(M)} \geq y_{k_2 l} - y_{k_1 l}, \quad e_k = (\mathbf{x}_{k_1}, \mathbf{x}_{k_2}) \in \mathcal{M} \quad (13)$$

$$z_{kl}^{(C)} \leq y_{k_1 l} + y_{k_2 l}, \quad e_k = (\mathbf{x}_{k_1}, \mathbf{x}_{k_2}) \in \mathcal{C} \quad (14)$$

$$z_{kl}^{(C)} \geq y_{k_1 l} + y_{k_2 l} - 1, \quad e_k = (\mathbf{x}_{k_1}, \mathbf{x}_{k_2}) \in \mathcal{C} \quad (15)$$

Minimization of objective function (8) under constraints (9) and (12)-(15) to solve for binary variables $\mathcal{Y}$, $\mathcal{Z}^{(\mathcal{M})}$, and $\mathcal{Z}^{(\mathcal{C})}$ is NP-hard. Kleinberg and Tardos proposed a linear programming relaxation of this integer programming problem by allowing $\mathcal{Y}$, $\mathcal{Z}^{(\mathcal{M})}$, and $\mathcal{Z}^{(\mathcal{C})}$ to be nonnegative real numbers, and provided a randomized method for rounding the real solution to the linear program to integers (Kleinberg & Tardos, 1999). We follow their approach, which allows us to perform collective assignment of all instances in $\mathcal{X}$ to cluster centroids.

4. Experiments

4.1. Methodology and Datasets

Experiments were conducted on three datasets: *Iris* from the UCI repository, the *Protein* dataset used by Xing et al. (2003) and Bar-Hillel et al. (2003), and a randomly sampled subset from the *Letters* handwritten character recognition dataset. For *Letters*, we chose three classes: $\{\mathbf{I}, \mathbf{J}, \mathbf{L}\}$, sampling 10% of the data points from the original dataset randomly. We used parameterized squared Euclidean distance as the clustering distortion measure D for these experiments.

We used pairwise F-Measure to evaluate the clustering results based on the underlying classes. F-Measure relies on the traditional information retrieval measures, adapted for evaluating clustering by considering same-cluster pairs:

$$Precision = \frac{\#PairsCorrectlyPredictedInSameCluster}{\#TotalPairsPredictedInSameCluster}$$

$$Recall = \frac{\#PairsCorrectlyPredictedInSameCluster}{\#TotalPairsInSameCluster}$$

$$F\text{–}Measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

We generated learning curves with 2-fold cross-validation for each dataset. Each point on the learning curve represents a particular number of randomly selected pairwise constraints given as input to the algorithm. Unit constraint costs were used for all constraints, since the datasets did not provide individual weights for the constraints. The clustering algorithm was run on the whole dataset, but the pairwise F-Measure was calculated only on the test set. Results were averaged over 10 runs of 2 folds. For each trial, cluster initialization was performed using neighborhoods inferred from the provided constraints (Bilenko et al., 2004), and then the HMRF-KMEANS algorithm was run with a particular inference technique in the E-step, and metric learning for Euclidean distance in the M-step.

4.2. Results and Discussion

We compared the three methods described in Section 3 for collective assignment of instances to clusters. Figures 4-6 show learning curves for the three datasets. For each

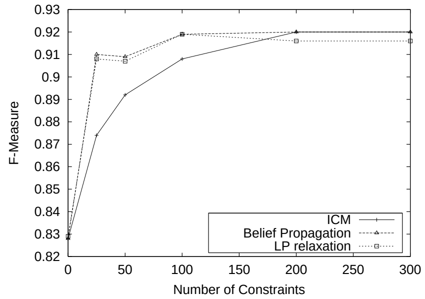

Figure 4. Iris results

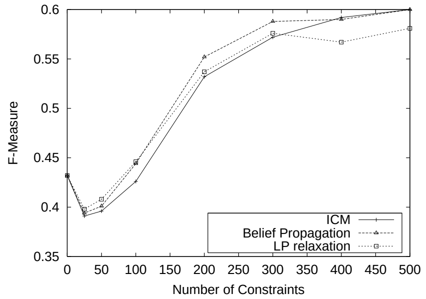

Figure 5. Protein results

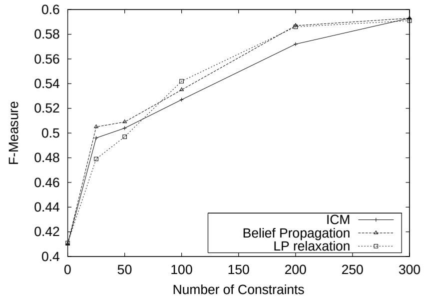

Figure 6. Letters results

dataset, ICM was faster than belief propagation and LP relaxation by at least an order of magnitude, which agrees with the relative computational complexities of these algorithms.

As the results demonstrate, global relational methods such as belief propagation and LP relaxation outperform the greedy approaches when a limited number of pairwise constraints is provided. However, as the number of provided constraints increases, returns from these computationally expensive methods diminish, and for every dataset there exists a number of constraints beyond which ICM performs no worse than the global approximate inference methods.

5. Conclusions

We have compared two methods for global approximate inference (belief propagation and LP relaxation) with a greedy approximate inference algorithm (ICM) in the context of collective assignment of data points to clusters in semi-supervised clustering with pairwise relational constraints. Our results indicate that belief propagation and LP relaxation outperform ICM when a limited number of pairwise constraints is provided. However, given a sufficiently large amount of relational supervision, the greedy algorithm for approximate inference performs on par with global methods. Thus, greedy inference techniques should be considered for scaling up semi-supervised clustering to large datasets due to their low computational cost.

6. Acknowledgments

We would like to thank Razvan Bunescu for helpful discussions. This research was supported in part by NSF grants IIS-0325116 and IIS-0117308.

References

Banerjee, A., Merugu, S., Dhillon, I. S., & Ghosh, J. (2004). Clustering with Bregman divergences. *Proceedings of SDM-2004*.

Bar-Hillel, A., Hertz, T., Shental, N., & Weinshall, D. (2003). Learning distance functions using equivalence relations. *Proceedings of ICML-2003*, (pp. 11–18).

Basu, S., Bilenko, M., & Mooney, R. J. (2004). A probabilistic framework for semi-supervised clustering. In submission, available at `http://www.cs.utexas.edu/~ml/publication`.

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society, Series B (Methodological)*, *48*, 259–302.

Bilenko, M., Basu, S., & Mooney, R. J. (2004). Integrating constraints and metric learning in semi-supervised clustering. To appear in *Proceedings of ICML-2004*.

Cohn, D., Caruana, R., & McCallum, A. (2003). *Semi-supervised clustering with user feedback* (Technical Report TR2003-1892). Cornell University.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, *39*, 1–38.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE PAMI*, *6*, 721–742.

Kearns, M., Mansour, Y., & Ng, A. Y. (1997). An information-theoretic analysis of hard and soft assignment methods for clustering. *Proceedings of UAI-1997* (pp. 282–293).

Klein, D., Kamvar, S. D., & Manning, C. (2002). From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. *Proceedings of ICML-2002* (pp. 307–314).

Kleinberg, J., & Tardos, E. (1999). Approximation algorithms for classification problems with pairwise relationships: Metric labeling and Markov random fields. *Proceedings of FOCS-1999* (pp. 14–23).

Kschischang, F. R., Frey, B., & Loeliger, H.-A. (2001). Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, *47*, 498–519.

Neal, R. M., & Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. *Learning in Graphical Models* (pp. 355–368). MIT Press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.

Segal, E., Wang, H., & Koller, D. (2003). Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, *19*, i264–i272.

Wagstaff, K., Cardie, C., Rogers, S., & Schroedl, S. (2001). Constrained K-Means clustering with background knowledge. *Proceedings of ICML-2001* (pp. 577–584).

Xing, E. P., Ng, A. Y., Jordan, M. I., & Russell, S. (2003). Distance metric learning, with application to clustering with side-information. *NIPS 15* (pp. 505–512).

Zhang, Y., Brady, M., & Smith, S. (2001). Hidden Markov random field model and segmentation of brain MR images. *IEEE Transactions on Medical Imaging*, *20*, 45–57.