Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), pp. 59-68, Seattle, WA, August 2004.

A Probabilistic Framework for Semi-Supervised Clustering

Sugato Basu
Dept. of Computer Sciences
University of Texas at Austin
Austin, TX 78712
sugato@cs.utexas.edu

Mikhail Bilenko
Dept. of Computer Sciences
University of Texas at Austin
Austin, TX 78712
mbilenko@cs.utexas.edu

Raymond J. Mooney
Dept. of Computer Sciences
University of Texas at Austin
Austin, TX 78712
mooney@cs.utexas.edu

ABSTRACT

Unsupervised clustering can be significantly improved using supervision in the form of pairwise constraints, i.e., pairs of instances labeled as belonging to same or different clusters. In recent years, a number of algorithms have been proposed for enhancing clustering quality by employing such supervision. Such methods use the constraints to either modify the objective function, or to learn the distance measure. We propose a probabilistic model for semi-supervised clustering based on Hidden Markov Random Fields (HMRFs) that provides a principled framework for incorporating supervision into prototype-based clustering. The model generalizes a previous approach that combines constraints and Euclidean distance learning, and allows the use of a broad range of clustering distortion measures, including Bregman divergences (e.g., Euclidean distance and I-divergence) and directional similarity measures (e.g., cosine similarity). We present an algorithm that performs partitional semi-supervised clustering of data by minimizing an objective function derived from the posterior energy of the HMRF model. Experimental results on several text data sets demonstrate the advantages of the proposed framework.

1. INTRODUCTION

Large amounts of unlabeled data are available in many real-life data-mining tasks, e.g., uncategorized messages in an automatic email classification system, genes of unknown functions for doing gene function prediction, etc. Labeled data is often limited and expensive to generate, since labeling typically requires human expertise. Consequently, *semi-supervised learning*, which uses both labeled and unlabeled data, has become a topic of significant recent interest [11, 24, 33]. In this paper, we focus on *semi-supervised clustering*, where the performance of unsupervised clustering algorithms is improved with limited amounts of supervision in the form of labels on the data or constraints [38, 6, 27, 39, 7].

Existing methods for semi-supervised clustering fall into two general categories which we call *constraint-based* and *distance-based*. Constraint-based methods rely on user-provided labels or constraints to guide the algorithm towards a more appropriate data partitioning. This is done by modifying the objective function for evaluating clusterings so that it includes satisfying constraints [15], enforcing constraints during the clustering process [38], or initializing and constraining the clustering based on labeled examples [6]. In distance-based approaches, an existing clustering algorithm that uses a particular clustering distortion measure is employed; however, it is trained to satisfy the labels or constraints in the supervised data. Several adaptive distance measures have been used for semi-supervised clustering, including string-edit distance trained using Expectation Maximization (EM) [10], KL divergence trained using gradient descent [13], Euclidean distance modified by a shortest-path algorithm [27], or Mahalanobis distances trained using convex optimization [39].

We propose a principled probabilistic framework based on Hidden Markov Random Fields (HMRFs) for semi-supervised clustering that combines the constraint-based and distance-based approaches in a unified model. We motivate an objective function for semi-supervised clustering derived from the posterior energy of the HMRF framework, and propose a EM-based partitional clustering algorithm HMRF-KMEANS to find a (local) minimum of this objective function. Previously, we proposed a unified approach to semi-supervised clustering that was experimentally shown to produce more accurate clusters than other methods on several data sets [8]. However, this approach is restricted to using Euclidean distance as the clustering distortion measure. In this paper, we show how to generalize that model to handle non-Euclidean measures. Our generalization can utilize any *Bregman divergence* [3], which includes a wide variety of useful distances, e.g., KL divergence. In a number of applications, such as text-clustering using a vector-space model, a directional similarity measure based on the angle between vectors is more appropriate [1]. Consequently, clustering algorithms that utilize distortion measures appropriate for directional data have recently been developed [18, 2]. Our unified semi-supervised clustering framework based on HMRFs is also applicable to such directional similarity measures.

To summarize, the proposed approach aids unsupervised clustering by incorporating labeled data in the following three ways:

- Improved initialization, where initial cluster centroids are estimated from the neighborhoods induced from constraints;
- Constraint-sensitive assignment of instances to clusters, where points are assigned to clusters so that the overall distortion of the points from the cluster centroids is minimized, while a minimum number of must-link and cannot-link constraints are violated;
- Iterative distance learning, where the distortion measure is re-estimated during clustering to warp the space to respect user-specified constraints as well as to incorporate data variance.

We present experimental results on clustering text documents that demonstrate the advantages of our approach.

2. BACKGROUND

2.1 Motivation of Framework

In this work, we will focus on partitional prototype-based clustering as our underlying unsupervised clustering model, where a set of data points is partitioned into a pre-specified number of clusters (each cluster having a representative or prototype) so that a well-defined cost function, involving a distortion measure between the points and the cluster representatives, is minimized. A popular clustering algorithm in this category is K-Means [29].

Earlier research on semi-supervised clustering has considered supervision in the form of labeled points [6] or constraints [38, 39, 5]. In this paper, we will be considering the model where supervision is provided in the form of *must-link* and *cannot-link* constraints, indicating respectively that a pair of points should be or should not be put in the same cluster. For each pairwise constraint, the model assigns an associated cost of violating that constraint. Considering supervision in the form of constraints is more realistic than requiring class labels in many unsupervised-learning applications, e.g. clustering for speaker identification in a conversation [5], or clustering GPS data for lane-finding [38]: while class labels may be unknown, a user can still specify whether pairs of points belong to same or different clusters. Constraint-based supervision is also more general than class labels: a set of classified points implies an equivalent set of pairwise constraints, but not vice versa.

Our semi-supervised clustering model considers a set of data points $\mathcal{X}$ with a specified distortion measure D between the points. Supervision is provided as a set $\mathcal{M}$ of must-link constraints (with a set of associated violation costs $\mathcal{W}$) and a set $\mathcal{C}$ of cannot-link constraints (with associated violation costs $\overline{\mathcal{W}}$). The task is to partition the data into K clusters so that the total distortion between the points and the corresponding cluster representatives according to the given measure D is minimized while a minimum number of constraints are violated. Since we restrict our attention to hard clustering, every point is assigned to a single cluster in our model.

A word on the notation and terminology used in this paper: boldface variables, e.g., $\mathbf{x}$, represent vectors; calligraphic upper-case alphabets, e.g., $\mathcal{X}$, refer to sets, whose representatives are enumerated as $\{\mathbf{x}_i\}_{i=1}^N$ (except $\mathcal{J}$, which always denotes an objective function); x_{im} represents the m^{th} component of the d-dimensional vector $\mathbf{x}_i$. The term "distance measure" is used synonymously with "distortion measure" throughout the paper.

2.2 Hidden Markov Random Field

To incorporate pairwise constraints along with an underlying distortion measure between points into a unified probabilistic model, we consider Hidden Markov Random Fields (HMRFs). An HMRF has the following components:

- A *hidden* field $\mathcal{L} = \{l_i\}_{i=1}^N$ of random variables, whose values are unobservable. In the clustering framework, the set of hidden variables are the unobserved cluster labels on the points, indicating cluster assignments. Every hidden variable l_i takes values from the set $\{1,\ldots,K\}$, which are the indices of the clusters.

- An *observable* set $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$ of random variables, where every random variable $\mathbf{x}_i$ is generated from a conditional probability distribution $\mathbf{Pr}(\mathbf{x}_i|l_i)$ determined by the corresponding hidden variable l_i. The random variables $\mathcal{X}$ are conditionally independent given the hidden variables $\mathcal{L}$, i.e., $\mathbf{Pr}(\mathcal{X}|\mathcal{L}) = \prod_{\mathbf{x}_i\in\mathcal{X}} \mathbf{Pr}(\mathbf{x}_i|l_i)$. In our framework, the set of observable variables for the HMRF corresponds to the given data points.

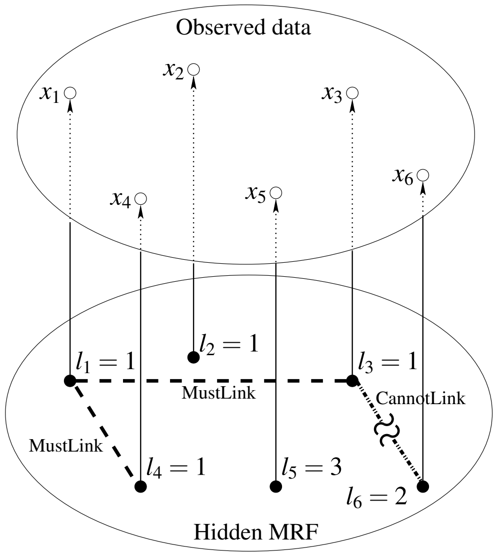

Figure 1: A Hidden Markov Random Field

Fig. 1 shows a simple example of an HMRF. The observed dataset $\mathcal{X}$ consists of six points $\{\mathbf{x}_1 \ldots \mathbf{x}_6\}$, which have corresponding cluster labels $\{l_1 \ldots l_6\}$. Two must-link constraints are provided between (l_1,l_3) and (l_1,l_4), while 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. 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.

Each 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 the hidden labels, such that the neighbors of a point $\mathbf{x}_i$ are all points that are must-linked or cannot-linked to it. The random field defined over the hidden variables is a Markov Random Field, where the probability distribution of the hidden variables obeys the following Markov property:

$$\forall i, \mathbf{Pr}(l_i|\mathcal{L} - \{l_i\}) = \mathbf{Pr}(l_i|\{l_j : l_j \in \mathcal{N}_i\}) \tag{1}$$

So, the probability distribution of the value of l_i for the data point $\mathbf{x}_i$ depends only on the cluster labels of the points that are must-linked or cannot-linked to $\mathbf{x}_i$.

Let us consider a particular cluster label configuration $\mathcal{L}$ to be the joint event $\mathcal{L} = \{l_i\}_{i=1}^N$. By the Hammersley-Clifford theorem [22], the probability of a label configuration can be expressed as a Gibbs distribution [21], so that

$$\mathbf{Pr}(\mathcal{L}) = \frac{1}{Z_1}\exp\big(-V(\mathcal{L})\big) = \frac{1}{Z_1}\exp\big(-\sum_{\mathcal{N}_i\in\mathcal{N}} V_{\mathcal{N}_i}(\mathcal{L})\big) \tag{2}$$

where $\mathcal{N}$ is the set of all neighborhoods, Z_1 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 $\mathcal{N}_i$ in the label configuration $\mathcal{L}$.

Since we are provided with pairwise constraints over the class labels, we restrict the MRFs over the hidden variable to have pairwise potentials. The prior probability of a configuration of cluster labels $\mathcal{L}$ then becomes $\mathbf{Pr}(\mathcal{L}) = \frac{1}{Z_1}\exp(-\sum_i \sum_j V(i,j))$, where

$$V(i,j) = \begin{cases} f_M(\mathbf{x}_i,\mathbf{x}_j) & \text{if } (\mathbf{x}_i,\mathbf{x}_j) \in \mathcal{M} \\ f_C(\mathbf{x}_i,\mathbf{x}_j) & \text{if } (\mathbf{x}_i,\mathbf{x}_j) \in \mathcal{C} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

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. Note that the third condition in the definition of $V(i,j)$ is necessary since not all points are involved in the constraints. Intuitively, this form of $\mathbf{Pr}(\mathcal{L})$ gives higher probabilities to label configurations that satisfy most of the must-link constraints $\mathcal{M}$ and cannot-link constraints $\mathcal{C}$, thereby discouraging the violation of the user-specified constraints.

2.3 MAP Estimation in HMRFs

Given a particular configuration of the hidden variables (unknown cluster labels), the variables in the observable field of the HMRF (the data points) are generated using specified conditional probability distributions. The conditional probability of the observation set $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$ for a given configuration $\mathcal{L} = \{l_i\}_{i=1}^N$ is given by $\mathbf{Pr}(\mathcal{X}|\mathcal{L})$, which in the clustering framework is of the form:

$$\mathbf{Pr}(\mathcal{X}|\mathcal{L}) = p(\mathcal{X},\{\boldsymbol{\mu}_h\}_{h=1}^K) \tag{4}$$

where $p(\mathcal{X},\{\boldsymbol{\mu}_h\}_{h=1}^K)$ is a probability density function parameterized by the cluster representatives $\{\boldsymbol{\mu}_h\}_{h=1}^K$. This function is related to the clustering distortion measure D, as we will show in Section 2.4.

The overall posterior probability of a cluster label configuration $\mathcal{L}$ is $\mathbf{Pr}(\mathcal{L}|\mathcal{X}) \propto \mathbf{Pr}(\mathcal{L})\mathbf{Pr}(\mathcal{X}|\mathcal{L})$, considering $\mathbf{Pr}(\mathcal{X})$ to be a constant C. Hence, finding the maximum a-posteriori (MAP) configuration of the HMRF becomes equivalent to maximizing the posterior probability:

$$\mathbf{Pr}(\mathcal{L}|\mathcal{X}) = \left(\frac{1}{Z_2}\exp\big(-\sum_i\sum_j V(i,j)\big)\right)\cdot p(\mathcal{X},\{\boldsymbol{\mu}_h\}_{h=1}^K) \tag{5}$$

where $Z_2 = CZ_1$. The negative logarithm of $\mathbf{Pr}(\mathcal{L}|\mathcal{X})$ is known as *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.

Since the cluster representatives as well as the cluster labels for the points are unknown in a clustering setting, maximizing Eqn.(5) is an "incomplete-data problem", for which a popular solution method is *Expectation Maximization* (EM) [16]. It is well-known that K-Means is equivalent to an EM algorithm with hard clustering assignments [26, 6, 3]. Section 3.2 describes a K-Means-type hard partitional clustering algorithm, HMRF-KMEANS, that finds a (local) maximum of the above function.

The posterior probability $\mathbf{Pr}(\mathcal{L}|\mathcal{X})$ in Eqn.(5) has 2 components: the first factor evaluates each label configuration, corresponding to cluster assignments of every point, and gives a higher probability to a configuration that satisfies more of the given must-link and cannot-link constraints. A particular label configuration determines the cluster assignments and hence the cluster representatives. The second factor estimates the probability of generating the observed data points using the conditional distributions, which are parameterized by the cluster representatives and depend on the distortion measure. The overall posterior probability of the cluster label configuration of all the points therefore takes into account both the cluster distortion measure and the constraints in a principled unified framework.

2.4 Clustering Objective Function

Eqn.(5) suggests a general framework for incorporating constraints into clustering. Particular choices of the constraint penalty functions f_M and f_C, and the conditional probabilities $p(\mathcal{X},\{\boldsymbol{\mu}_h\}_{i=1}^K)$ would be motivated by the distortion measure appropriate for the clustering task.

When considering the second term in Eqn.(5), we restrict our attention to probability densities of the exponential form:

$$p(\mathcal{X},\{\boldsymbol{\mu}_h\}_{h=1}^K) = \frac{1}{Z_3}\exp\big(-\sum_{x_i\in\mathcal{X}} D(\mathbf{x}_i,\boldsymbol{\mu}_{l_i})\big) \tag{6}$$

where $D(\mathbf{x}_i,\boldsymbol{\mu}_{l_i})$ is the distortion between $\mathbf{x}_i$ and $\boldsymbol{\mu}_{l_i}$, and Z_3 is a normalization constant. Different clustering models fall into this exponential form:

- $\mathbf{x}_i$ and $\boldsymbol{\mu}_{l_i}$ are vectors and D is the square of the L_2 norm: the cluster conditional probability is a unit variance Gaussian [26];
- $\mathbf{x}_i$ and $\boldsymbol{\mu}_{l_i}$ are probability distributions and D is the KL-divergence: the cluster conditional probability is a multinomial distribution [17];
- $\mathbf{x}_i$ and $\boldsymbol{\mu}_{l_i}$ are vectors of unit length (according to the L_2 norm) and D is one minus the dot-product: the cluster conditional probability is a von-Mises Fisher (vMF) distribution with unit concentration parameter [2], which is essentially the spherical analog of a unit variance Gaussian.

We will discuss the connection between specific distortion measures that we will study in this paper and their corresponding cluster conditional probabilities in more detail in Section 3.1.

Let us now examine the potential function V in the first term of Eqn.(5). In previous work, only must-linked points were considered in the neighborhood of a Markov Random Field with the *generalized Potts* potential function [12, 28]. In this potential function, the must-link penalty is $f_M(\mathbf{x}_i,\mathbf{x}_j) = w_{ij}\mathbb{1}[l_i \neq l_j]$, where w_{ij} is the cost for violating the must-link constraint (i,j), and $\mathbb{1}$ is the indicator function ($\mathbb{1}$[true] = 1, $\mathbb{1}$[false] = 0). This function specifies that the cost of violating a must-link constraint $(\mathbf{x}_i,\mathbf{x}_j)$ is w_{ij} irrespective of the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$.

In a semi-supervised clustering framework where we want to use the constraint violations to learn the underlying distance measure, the penalty for violating a must-link constraint between *distant* points should be higher than that between *nearby* points. This would reflect the fact that if two must-linked points are far apart according to the current distortion measure and are hence put in different clusters, the measure is inadequate and needs to be modified to bring those points closer together. So, the must-link penalty function is chosen to be

$$f_M(\mathbf{x}_i,\mathbf{x}_j) = w_{ij}\varphi_D(\mathbf{x}_i,\mathbf{x}_j)\mathbb{1}[l_i \neq l_j] \tag{7}$$

where φ_D is the *penalty scaling* function, which we choose to be a monotonically increasing function of the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ according to the current distortion measure. Specific penalty functions φ_D for different distortion measures D are described in Section 3.1.

Analogously, 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. This would encourage the distance learning step to put cannot-linked points farther apart. The cannot-link penalty function can be accordingly chosen to be

$$f_C(\mathbf{x}_i,\mathbf{x}_j) = \overline{w}_{ij}\big(\varphi_{D\max} - \varphi_D(\mathbf{x}_i,\mathbf{x}_j)\big)\mathbb{1}[l_i = l_j] \tag{8}$$

where $\varphi_{D\max}$ is the maximum value of the scaling function φ_D for the dataset. This form of f_C ensures that the penalty for violating a cannot-link constraint remains non-negative, since the second term is never greater than the first. Note that these f_M and f_C penalty functions make the MRF over the hidden variables non-isotropic (i.e., the values of the potential between pairs of random variables in the field are non-uniform), but the overall model is still a valid HMRF.

Putting this into Eqn.(5) and taking logarithms 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:

$$\mathcal{J}_{\text{obj}} = \sum_{x_i \in \mathcal{X}} D(\mathbf{x}_i, \boldsymbol{\mu}_{l_i}) + \sum_{(x_i,x_j)\in\mathcal{M}} w_{ij}\varphi_D(\mathbf{x}_i,\mathbf{x}_j)\mathbb{1}[l_i \neq l_j] + \sum_{(x_i,x_j)\in\mathcal{C}} \overline{w}_{ij}\big(\varphi_{D\max} - \varphi_D(\mathbf{x}_i,\mathbf{x}_j)\big)\mathbb{1}[l_i = l_j] + \log Z \quad (9)$$

where $Z = Z_2 Z_3$. Thus, the task is to minimize $\mathcal{J}_{\text{obj}}$ over $\{\boldsymbol{\mu}_h\}_{h=1}^K$, $\mathcal{L}$, and D (if the latter is parameterized).

3. ALGORITHM

3.1 Adaptive Distortion Measures

The choice of a distortion measure D for a particular clustering problem depends on the properties of the domain under consideration. A number of popular distortion measures, including Euclidean distance and Kullback-Leibler divergence, belong to a general family of functions known as *Bregman divergences* [3]. Another popular class of distortion measures includes *directional* similarity functions such as normalized dot product (cosine similarity) and Pearson's correlation [31]. Selection of the most appropriate distortion measure for a clustering task should take into account intrinsic properties of the dataset. For example, Euclidean distance is most appropriate for low-dimensional data with distribution close to the normal distribution, while normalized dot product best captures similarity of directional data where differences in angles between vectors are important, while vector lengths are not. For Bregman divergences and directional similarity measures like cosine similarity, it has been shown that there exist efficient K-Means-type iterative relocation algorithms that minimize the corresponding clustering cost functions [2, 3].

For many realistic datasets, off-the-shelf distortion measures may fail to capture the correct notion of similarity in a clustering setting. Unsupervised measures like Mahalanobis distance and Pearson correlation attempt to correct similarity estimates using the global mean and variance of the dataset. However, these measures may still fail to estimate distances accurately if the attributes' true contribution to similarity is not correlated with their variance. Recently, several semi-supervised clustering approaches have been proposed that incorporate adaptive similarity functions, including parameterization of Jensen-Shannon divergence [13] and Euclidean distance [5, 39]. In initial work [8], we have shown how Euclidean distance can be parameterized and learned in a principled manner in a semi-supervised clustering setting. We now turn to two other popular distortion measures, cosine similarity and Kullback-Leibler divergence, and describe how their adaptive versions can be used as distortion measures in our HMRF-based framework.

3.1.1 *Parameterized Cosine Similarity*

Cosine similarity can be parameterized using a symmetric positive-definite matrix $\mathbf{A}$, which leads to the following distortion measure:

$$D_{\cos_{\mathbf{A}}}(\mathbf{x}_i,\mathbf{x}_j) = 1 - \frac{\mathbf{x}_i^T \mathbf{A}\mathbf{x}_j}{\|\mathbf{x}_i\|_{\mathbf{A}}\|\mathbf{x}_j\|_{\mathbf{A}}} \quad (10)$$

where $\|\mathbf{x}\|_{\mathbf{A}}$ is the weighted L_2 norm: $\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\mathbf{x}^T\mathbf{A}\mathbf{x}}$. Such parameterization is equivalent to projecting every instance $\mathbf{x}$ onto a space spanned by $\mathbf{A}^{1/2}$: $\mathbf{x} \to \mathbf{A}^{1/2}\mathbf{x}$. Since unparameterized cosine similarity is a natural measure for prototype-based clustering under the assumption that the data is generated by a mixture of von Mises-Fisher (vMF) distributions [2], $D_{\cos_{\mathbf{A}}}(\mathbf{x}_i,\mathbf{x}_j)$ can be thought of as a distortion measure for data generated by a mixture of vMF distributions in the projected space. Because for realistic high-dimensional domains computing the full matrix $\mathbf{A}$ would be extremely expensive computationally, we focus our attention on diagonal $\mathbf{A}$, which is equivalent to using a vector of weights $\mathbf{a} = \text{diag}(\mathbf{A})$. Therefore, from now on we will be referring to the cosine measure in Eqn.(10) as $D_{\cos_{\mathbf{a}}}(\mathbf{x}_i,\mathbf{x}_j)$.

To use $D_{\cos_{\mathbf{a}}}(\mathbf{x}_i,\mathbf{x}_j)$ as the distortion measure in the clustering framework described in Section 2.4, we also use it as the penalty scaling function $\varphi_D(\mathbf{x}_i,\mathbf{x}_j) = D_{\cos_{\mathbf{a}}}(\mathbf{x}_i,\mathbf{x}_j)$, which leads to the following objective function:

$$\begin{aligned}\mathcal{J}_{\cos_{\mathbf{a}}} = &\sum_{x_i\in\mathcal{X}} D_{\cos_{\mathbf{a}}}(\mathbf{x}_i,\boldsymbol{\mu}_{l_i}) \\ &+ \sum_{(x_i,x_j)\in\mathcal{M}} w_{ij} D_{\cos_{\mathbf{a}}}(\mathbf{x}_i,\mathbf{x}_j)\mathbb{1}[l_i\neq l_j] \\ &+ \sum_{(x_i,x_j)\in\mathcal{C}} \overline{w}_{ij}\big(D_{\cos_{\mathbf{a}}\max} - D_{\cos_{\mathbf{a}}}(\mathbf{x}_i,\mathbf{x}_j)\big)\mathbb{1}[l_i = l_j] + \log Z\end{aligned} \quad (11)$$

where $D_{\cos_{\mathbf{a}}\max} = 1$.

3.1.2 *Parameterized I-Divergence*

In certain domains, data is described by probability distributions, e.g. text documents can be represented as probability distributions over words generated by a multinomial model [35]. KL-divergence is a widely used distance measure for such data: $D_{KL}(\mathbf{x}_i,\mathbf{x}_j) = \sum_{m=1}^d x_{im}\log\frac{x_{im}}{x_{jm}}$, where $\mathbf{x}_i$ and $\mathbf{x}_j$ are probability distributions over d events: $\sum_{m=1}^d x_{im} = \sum_{m=1}^d x_{jm} = 1$. In previous work, Cohn et al. parameterized KL-divergence multiplying m-th component by a weight γ_m: $D'_{KL}(\mathbf{x}_i,\mathbf{x}_j) = \sum_{m=1}^d \gamma_m x_{im}\log\frac{x_{im}}{x_{jm}}$. It can be shown that after such parameterization D'_{KL} is no longer a Bregman divergence over probability distributions, which is undesirable since convergence is no longer guaranteed for the algorithm described in [13].

Instead of KL-divergence, we employ a related measure, I-divergence, which also belongs to the class of Bregman divergences [3]. I-divergence has the following form: $D_I(\mathbf{x}_i,\mathbf{x}_j) = \sum_{m=1}^d x_{im}\log\frac{x_{im}}{x_{jm}} - \sum_{m=1}^d (x_{im}-x_{jm})$; $\mathbf{x}_i$ and $\mathbf{x}_j$ no longer need to be probability distributions but can be any non-negative vectors. For probability distributions, I-divergence and KL-divergence are equivalent. We parameterize I-divergence by a vector of non-negative weights $\mathbf{a}$:

$$D_{I_{\mathbf{a}}}(\mathbf{x}_i,\mathbf{x}_j) = \sum_{m=1}^d a_m x_{im}\log\frac{x_{im}}{x_{jm}} - \sum_{m=1}^d a_m(x_{im}-x_{jm}) \quad (12)$$

Such parameterization can be thought of as scaling every attribute in the original space by a weight contained in the corresponding component of $\mathbf{a}$, and then taking I-divergence in the transformed space. This implies that $D_{I_{\mathbf{a}}}$ is a Bregman divergence with respect to the transformed space.

The clustering framework described in Section 2.4 requires us to define an appropriate penalty scaling function $\varphi_D(\mathbf{x}_i,\mathbf{x}_j)$ to be used in the HMRF potential functions as described in Eqns.(3) and (7-8). Since we consider unordered constraint pairs, $\varphi_D(\mathbf{x}_i,\mathbf{x}_j)$ must be symmetric to penalize constraints appropriately. To meet this requirement, we will use a sum of weighted I-divergences from $\mathbf{x}_i$ and $\mathbf{x}_j$ to the mean vector $\frac{\mathbf{x}_i+\mathbf{x}_j}{2}$. This "I-divergence to the mean", $D_{IM_{\mathbf{a}}}$, is analogous to Jensen-Shannon divergence, which is the symmetric "KL-divergence to the mean" [14], and is defined as follows:

$$\varphi_D(\mathbf{x}_i,\mathbf{x}_j) = D_{IM_\mathbf{a}}(\mathbf{x}_i,\mathbf{x}_j) = \sum_{m=1}^{d} a_m \big(x_{im}\log\frac{2x_{im}}{x_{im}+x_{jm}} + x_{jm}\log\frac{2x_{jm}}{x_{im}+x_{jm}}\big) \quad (13)$$

This formulation leads to the following objective function:

$$\mathcal{J}_{I_\mathbf{a}} = \sum_{x_i\in\mathcal{X}} D_{I_\mathbf{a}}(\mathbf{x}_i,\boldsymbol{\mu}_{l_i}) + \sum_{(x_i,x_j)\in\mathcal{M}} w_{ij} D_{IM_\mathbf{a}}(\mathbf{x}_i,\mathbf{x}_j)\mathbb{1}[l_i \neq l_j] + \sum_{(x_i,x_j)\in\mathcal{C}} \overline{w}_{ij}\big(D_{IM_\mathbf{a}\max} - D_{IM_\mathbf{a}}(\mathbf{x}_i,\mathbf{x}_j)\big)\mathbb{1}[l_i = l_j] + \log Z \quad (14)$$

The two parameterized distortion measures $D_{\cos_\mathbf{a}}$ and $D_{IM_\mathbf{a}}$ have underlying generative models: weighted cosine corresponds to a von-Mises Fisher (vMF) distribution in the projected space, while I-divergence corresponds to multinomial distributions with rescaled probabilities. Thus, $\Pr(\mathcal{X}|\mathcal{L})$ in Eqn.(4) is well-defined for the underlying HMRF model in both these cases, and minimizing objective functions $\mathcal{J}_{\cos_\mathbf{a}}$ and $\mathcal{J}_{I_\mathbf{a}}$ leads to maximizing $\Pr(\mathcal{L}|\mathcal{X})$ for the corresponding underlying models.

3.2 EM Framework

As discussed in Section 2.2, $\mathcal{J}_{\text{obj}}$ can be minimized by a K-Means-type iterative algorithm HMRF-KMEANS. The outline of the algorithm is presented in Fig. 2. The basic idea of HMRF-KMEANS is as follows: in the E-step, given the current cluster representatives, every data point is re-assigned to the cluster which minimizes its contribution to $\mathcal{J}_{\text{obj}}$. In the M-step, the cluster representatives $\{\boldsymbol{\mu}_h\}_{h=1}^K$ 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. Note that this corresponds to the generalized EM algorithm [32, 16], 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 $\{\boldsymbol{\mu}_h\}_{h=1}^K$, 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. The specific details of the E-step and M-step are discussed in the following sections.

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.(9) is (locally) minimized.
Method:
1. Initialize the K clusters centroids $\{\boldsymbol{\mu}_h^{(0)}\}_{h=1}^K$, set t ← 0
2. Repeat until *convergence*
2a. `E-step`: Given $\{\boldsymbol{\mu}_h^{(t)}\}_{h=1}^K$, 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 $\{\boldsymbol{\mu}_h^{(t+1)}\}_{h=1}^K$ to minimize $\mathcal{J}_{\text{obj}}$.
2c. `M-step(B)`: Re-estimate distance measure D to reduce $\mathcal{J}_{\text{obj}}$.
2d. t ← t+1

Figure 2: HMRF-KMEANS **algorithm**

Note that calculating the normalizing constant Z in Eqn.(9) is computationally intensive for most distortion measures, e.g. for cosine similarity, this corresponds to computing a Bessel function [2]. So, we make an approximation by considering $\log Z$ to be constant throughout the clustering iterations, and hence drop that term from Eqn.(9).

3.3 Initialization

Good initial centroids are essential for the success of partitional clustering algorithms such as K-Means. In previous work, it was shown that using limited supervision in the form of labeled points results in good initial centroids for partitional clustering [6]. In our case, supervision is provided as pairwise constraints instead of labeled points. However, we follow the same motivation of inferring good initial centroids from the constraints.

We try to utilize both the constraints and unlabeled data during initialization. For this, we follow a two stage initialization process.

Neighborhood inference: We begin by taking the transitive closure of the must-link constraints to get connected components consisting of points connected by must-links. Let there be λ connected components, which are used to create λ neighborhoods $\{\mathcal{N}_p\}_{p=1}^{\lambda}$. These define the must-link neighborhoods in the MRF over the hidden cluster variables.

Assuming consistency of the constraints, we then infer additional constraints from the neighborhoods. We augment the set $\mathcal{M}$ with the must-link constraints inferred from the transitive closure that were not in the initial set. For each pair of neighborhoods $\mathcal{N}_p$ and $\mathcal{N}_{p'}$ that have at least one cannot-link between them, we add cannot-link constraints between every pair of points in $\mathcal{N}_p$ and $\mathcal{N}_{p'}$ and augment the cannot-link set $\mathcal{C}$ with these entailed constraints. This step corresponds to inferring as much information as possible about the neighborhood structure of the hidden MRF, under the assumption of consistency of the constraints.

From this point onwards in the paper, we will overload notation and refer to the augmented must-link and cannot-link sets as $\mathcal{M}$ and $\mathcal{C}$ respectively. Note that if we know that the given set of constraints are noisy, implying that the constraints are not consistent, we will not add these additional inferred constraints to $\mathcal{M}$ and $\mathcal{C}$ and only work with the constraints provided initially.

Cluster selection: The first stage produces λ neighborhood sets $\{\mathcal{N}_p\}_{p=1}^{\lambda}$. These neighborhoods are used as initial clusters for the HMRF-MEANS algorithm. If $\lambda = K$, λ cluster centers are initialized with the centroids of all the λ neighborhood sets. If $\lambda < K$, λ clusters are initialized from the neighborhoods, and the remaining $K-\lambda$ clusters are initialized with points obtained by random perturbations of the global centroid of $\mathcal{X}$.

If $\lambda > K$, K neighborhoods are selected as initial clusters using the clustering distortion measure. Farthest-first traversal is a good heuristic for initialization in prototype-based partitional clustering algorithms [23]. The goal in farthest-first traversal is to find K points that are maximally separated from each other in terms of a given distance function. In our case, we apply a weighted variant of farthest-first traversal to the centroids of the λ neighborhoods, where the weight of each centroid is proportional to the size of the corresponding neighborhood. We consider the weighted distance between two centroids to be the distance between them according to the distortion measure multiplied by the weights of the two centroids. Thus, weighted farthest-first is biased to select centroids that are relatively far apart as well as large in size.

During weighted farthest first selection, the algorithm maintains a set of centroids that have been visited so far. The centroid of the largest neighborhood is selected as the starting point and added to the visited set. At every point in the algorithm, the unvisited centroid with the farthest weighted distance (smallest weighted similarity) from the visited set is chosen. If there is a tie, it is resolved by selecting the centroid farthest from the global centroid of the data. This point is added to the visited set, and the process is continued till K centroids are visited. Finally, the K neighborhood centroids chosen by weighted farthest-first traversal are set as the K initial cluster centroids for HMRF-KMEANS.

Overall, this two-stage initialization procedure is able to take into account both unlabeled and labeled data to obtain cluster representatives that provide a good initial partitioning of the dataset.

3.4 E-step

In the E-step, assignments of data points to clusters are updated 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. As a result, computing the assignment of data points to cluster representatives to minimize the objective function is computationally intractable in any non-trivial HMRF model [36].

There exist several techniques for computing cluster assignments that approximate the optimal solution in this framework, e.g., iterated conditional modes (ICM) [9, 40], belief propagation [34, 36], and linear programming relaxation [28]. We follow the ICM approach, which is a greedy strategy to sequentially update the cluster assignment of each point, keeping the assignments for the other points fixed.

The algorithm performs cluster assignments in random order for all points. Each point $\mathbf{x}_i$ is assigned to the cluster representative $\boldsymbol{\mu}_h$ that minimizes the point's contribution to the objective function $\mathcal{J}_{obj}(\mathbf{x}_i,\boldsymbol{\mu}_h)$:

$$\mathcal{J}_{\text{obj}}(\mathbf{x}_i,\boldsymbol{\mu}_h) = D(\mathbf{x}_i,\boldsymbol{\mu}_h) + \sum_{(x_i,x_j)\in\mathcal{M}} w_{ij}\varphi_D(\mathbf{x}_i,\mathbf{x}_j)\mathbb{1}[h \neq l_j] + \sum_{(x_i,x_j)\in\mathcal{C}} \overline{w}_{ij}\big(\varphi_{D\max} - \varphi_D(\mathbf{x}_i,\mathbf{x}_j)\big)\mathbb{1}[h = l_j] \tag{15}$$

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 $\mathcal{J}_{\text{obj}}$). After all points are assigned, they are randomly re-ordered, and the assignment process is repeated. 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 [9].

Overall, the assignment of points to clusters incorporates pairwise supervision by discouraging constraint violations proportionally to their severity, which guides the algorithm towards a desirable partitioning of the data.

3.5 M-step

The M-step of the algorithm consists of two parts. First, cluster representatives $\{\boldsymbol{\mu}_h\}_{h=1}^K$ are re-estimated from points currently assigned to them to decrease the objective function $\mathcal{J}_{\text{obj}}$ in Eqn.(9). It has recently been shown that for Bregman divergences each cluster representative calculated in the M-step of the EM algorithm is equivalent to the expectation value over the points in that cluster, which is essentially their arithmetic mean [3]. Additionally, it has been experimentally demonstrated that for distribution-based clustering, smoothing cluster representatives by a prior using a deterministic annealing schedule leads to considerable improvements [17]. With smoothing controlled by a parameter α, each cluster representative $\boldsymbol{\mu}_h$ is estimated as follows when $D_{I_\mathbf{a}}$ is the distortion measure:

$$\boldsymbol{\mu}_h^{(I_\mathbf{a})} = \frac{1}{1+\alpha}\left(\frac{\sum_{\mathbf{x}_i\in\mathcal{X}_h}\mathbf{x}_i}{|\mathcal{X}_h|} + \alpha\frac{1}{n}\right) \tag{16}$$

For directional measures, each cluster representative is the arithmetic mean projected onto unit sphere [2]. Taking the weighting into account, centroids are estimated as follows when $D_{\cos_\mathbf{a}}$ is the distortion measure:

$$\boldsymbol{\mu}_h^{(\cos_\mathbf{a})} = \frac{\sum_{\mathbf{x}_i\in\mathcal{X}_h}\mathbf{x}_i}{\|\sum_{\mathbf{x}_i\in\mathcal{X}_h}\mathbf{x}_i\|_\mathbf{A}} \tag{17}$$

Since constraints do not take part in cluster representative re-estimation, this step remains the same as in K-Means for Bregman divergences, and the same as in SPKMEANS for weighted cosine similarity [18].

Second, if a parameterized variant of a distortion measure is used, e.g. $D_{\cos_\mathbf{a}}$ or $D_{I_\mathbf{a}}$ shown above, the distortion measure parameters must be updated to decrease the objective function. For certain distance measure parameterizations, minimization via taking partial derivatives and solving for the parameter values may be feasible, e.g. for Euclidean distance [8]. In general, however, a closed-form solution may be unattainable. In such cases, gradient descent provides an alternative avenue for learning distortion measure weights. For the two distortion measures described above, $D_{\cos_\mathbf{a}}$ and $D_{I_\mathbf{a}}$, every weight a_m would be updated using the update rule $a_m = a_m + \eta\frac{\partial\mathcal{J}_{\text{obj}}}{\partial a_m}$, where:

$$\begin{aligned}\frac{\partial\mathcal{J}_{\text{obj}}}{\partial a_m} &= \sum_{x_i\in\mathcal{X}}\frac{\partial D(\mathbf{x}_i,\boldsymbol{\mu}_{l_i})}{\partial a_m} \\ &+ \sum_{(x_i,x_j)\in\mathcal{M}} w_{ij}\frac{\partial D(\mathbf{x}_i,\mathbf{x}_j)}{\partial a_m}\mathbb{1}[l_i\neq l_j] \\ &+ \sum_{(x_i,x_j)\in\mathcal{C}}\overline{w}_{ij}\left[\frac{\partial D_{\max}}{\partial a_m} - \frac{\partial D(\mathbf{x}_i,\mathbf{x}_j)}{\partial a_m}\right]\mathbb{1}[l_i = l_j]\end{aligned} \tag{18}$$

For the two particular distortion measures that we are considering, $D_{\cos_\mathbf{a}}$ and $D_{I_\mathbf{a}}$, gradients $\frac{\partial D(\mathbf{x}_i,\mathbf{x}_j)}{\partial a_m}$ are the following:

$$\frac{\partial D_{\cos_\mathbf{a}}(\mathbf{x}_i,\mathbf{x}_j)}{\partial a_m} = \frac{x_{im}x_{jm}\|\mathbf{x}_i\|_\mathbf{A}\|\mathbf{x}_j\|_\mathbf{A} - \mathbf{x}_i^T\mathbf{A}\mathbf{x}_j\frac{x_{im}^2\|\mathbf{x}_j\|_\mathbf{A}^2 + x_{jm}^2\|\mathbf{x}_i\|_\mathbf{A}^2}{2\|\mathbf{x}_i\|_\mathbf{A}\|\mathbf{x}_j\|_\mathbf{A}}}{\|\mathbf{x}_i\|_\mathbf{A}^2\|\mathbf{x}_j\|_\mathbf{A}^2} \tag{19}$$

$$\frac{\partial D_{I_\mathbf{a}}(\mathbf{x}_i,\mathbf{x}_j)}{\partial a_m} = x_{im}\log\frac{x_{im}}{x_{jm}} - (x_{im} - x_{jm}) \tag{20}$$

Intuitively, the distance learning step results in modifying the distortion measure so that similar data points are brought closer together, while dissimilar points are pulled apart. This process leads to a transformed data space, which facilitates partitioning of the unlabeled data that respects supervised constraints provided by the user and reflects natural variance in the data.

4. EXPERIMENTS

4.1 Datasets

When clustering sparse high-dimensional data, e.g. text documents represented using the vector space model, it is particularly

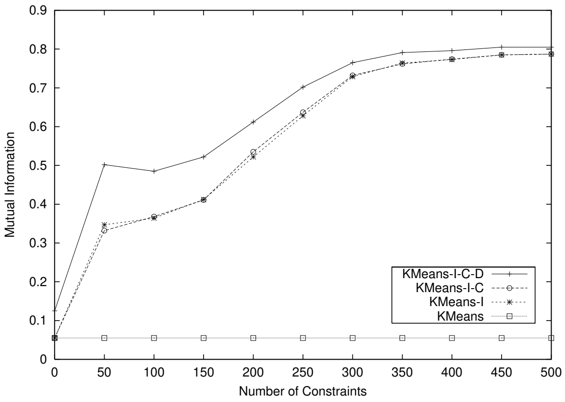

Figure 3: Clustering results for $D_{\cos_a}$ on *News-Different-3* dataset

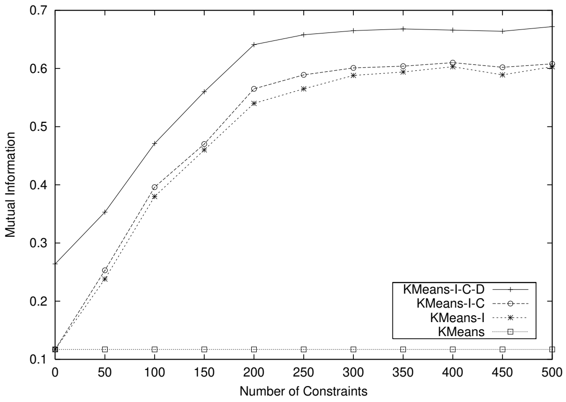

Figure 4: Clustering results for D_{I_a} on *News-Different-3* dataset

difficult to cluster small datasets. This is due to the fact that clustering algorithms can easily get stuck in local optima on such datasets, which leads to poor clustering quality. In previous studies with SP-KMEANS algorithm applied to document collections whose size is small compared to the dimensionality of the word space, it has been observed that there is little relocation of documents between clusters for most initializations, which leads to poor clustering quality after convergence of the algorithm [17].

This scenario is likely in many realistic applications. For example, when clustering the search results in a web-search engine like Vivísimo[1], typically the number of webpages that are being clustered is in the order of hundreds. However the dimensionality of the feature space, corresponding to the number of unique words in all the webpages, is in the order of thousands. Moreover, each webpage is sparse, since it contains only a small number of all the possible words. Supervision in the form of pairwise constraints can be beneficial in such cases and may significantly improve clustering quality. To demonstrate the effectiveness of our semi-supervised clustering framework, we consider 3 data sets that have the characteristics of being sparse, high-dimensional, and having a small number of points compared to the dimensionality of the space.

We derived 3 datasets from the *20-Newsgroups* collection.[2] This collection has messages harvested from 20 different Usenet newsgroups, 1000 messages from each newsgroup. From the original dataset, a reduced dataset was created by taking a random subsample of 100 documents from each of the 20 newsgroups. Three datasets were created by selecting 3 categories from the reduced collection. *News-Similar-3* consists of 3 newsgroups on similar topics (`comp.graphics`, `comp.os.ms-windows`, `comp.windows.x`) with significant overlap between clusters due to cross-posting. *News-Related-3* consists of 3 newsgroups on related topics (`talk.politics.misc`, `talk.politics.guns`, and `talk.politics.mideast`). *News-Different-3* consists of articles posted in 3 newsgroups that cover different topics (`alt.atheism`, `rec.sport.baseball`, `sci.space`) with well-separated clusters. The vector-space model of *News-Similar-3* has 300 points in 1864 dimensions, *News-Related-3* has 300 points in 3225 dimensions, and *News-Different-3* had 300 points in 3251 dimensions. Since the overlap between topics in *News-Similar-3* and *News-Related-3* is significant, they are more challenging datasets than *News-Different-3*.

All the datasets were pre-processed by stop-word removal, TF-IDF weighting, removal of very high-frequency and low-frequency words, etc., following the methodology of Dhillon et al. [18].

4.2 Clustering Evaluation

We used *normalized mutual information* (NMI) as our clustering evaluation measure. NMI is an external clustering validation metric that estimates the quality of the clustering with respect to a given underlying class labeling of the data: it measures how closely the clustering algorithm could reconstruct the underlying label distribution in the data [37, 19]. If C is the random variable denoting the cluster assignments of the points and K is the random variable denoting the underlying class labels on the points [2], then the NMI measure is defined as:

$$NMI = \frac{I(C;K)}{(H(C)+H(K))/2} \tag{21}$$

where $I(X;Y) = H(X) - H(X|Y)$ is the mutual information between the random variables X and Y, $H(X)$ is the Shannon entropy of X, and $H(X|Y)$ is the conditional entropy of X given Y [14]. NMI effectively measures the amount of statistical information shared by the random variables representing the cluster assignments and the user-labeled class assignments of the data points.

4.3 Methodology

We generated learning curves using 20 runs of 2-fold cross-validation for each dataset. For studying the effect of constraints in clustering, 50% of the dataset is set aside as the test set at any particular fold. The different points along the learning curve correspond to constraints that are given as input to the semi-supervised clustering algorithm. These constraints are obtained from the training set corresponding to the remaining 50% of the data by randomly selecting pairs of points from the training set, and creating must-link or cannot-link constraints depending on whether the underlying classes of the two points are same or different. Unit constraint costs $\mathcal{W}$ and $\overline{\mathcal{W}}$ were used for all constraints, original and inferred, since the datasets did not provide individual weights for the constraints. Based on a few pilot studies, gradient step size η was chosen to have values $\eta = 1.75$ for clustering with $D_{\cos_a}$ and $\eta = 1.0^{-8}$ for clustering with D_{I_a}; weights were restricted to be non-negative. In a realistic setting, these parameters could be tuned using cross-validation with a hold-out set. The clustering algorithm was run on the whole dataset, but NMI was calculated only on the test set. The learning curve results were averaged over the 20 runs.

[1] http://www.vivisimo.com

[2] http://www.ai.mit.edu/people/jrennie/20Newsgroups

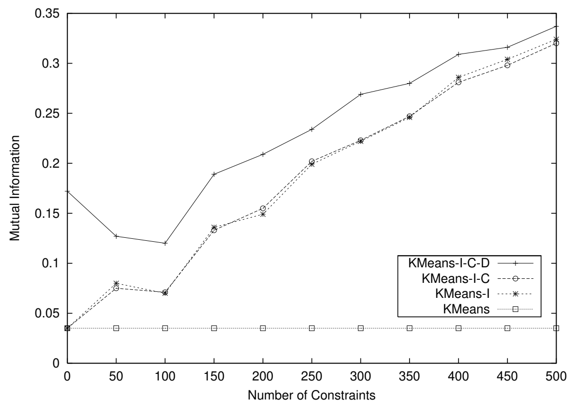

Figure 5: Clustering results for $D_{\cos_a}$ on *News-Related-3* dataset

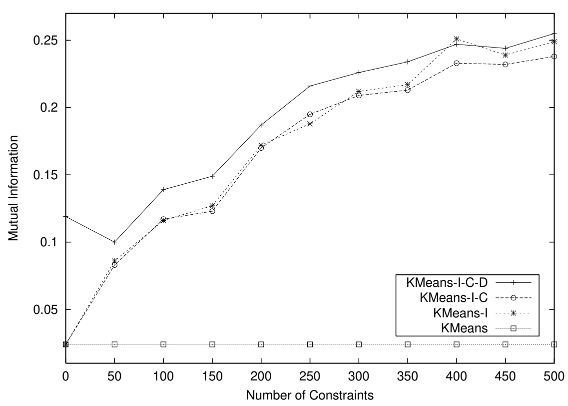

Figure 6: Clustering results for D_{I_a} on *News-Related-3* dataset

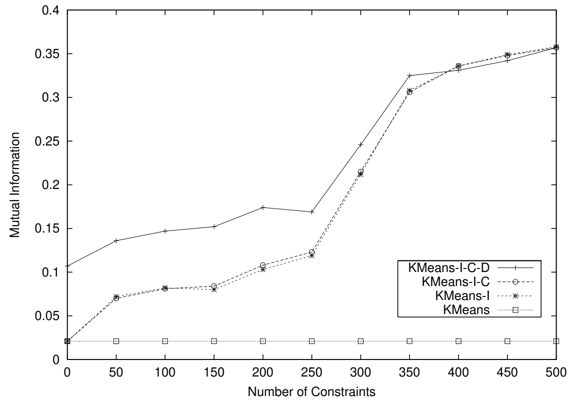

Figure 7: Clustering results for $D_{\cos_a}$ on *News-Similar-3* dataset

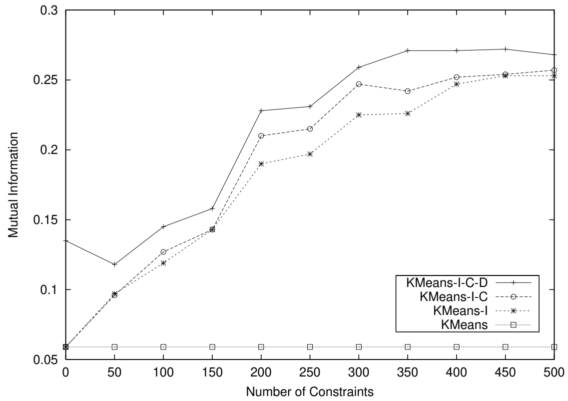

Figure 8: Clustering results for D_{I_a} on *News-Similar-3* dataset

4.4 Results and Discussion

We compared the proposed HMRF-KMEANS algorithm with two ablations as well as unsupervised K-Means clustering. The following variants were compared for distortion measures $D_{\cos_a}$ and D_{I_a} as representatives for Bregman divergences and directional measures respectively:

- KMEANS-I-C-D is the complete HMRF-KMEANS algorithm that includes use of supervised data in initialization (I) as described in Section 3.3, incorporates constraints in cluster assignments (C) as described in Section 3.4, and performs distance learning (D) as described in Section 3.5;
- KMEANS-I-C is an ablation of HMRF-KMEANS that uses pairwise supervision for initialization and cluster assignments, but does not perform distance learning;
- KMEANS-I is a further ablation that only uses the constraints to initialize cluster representatives;
- KMEANS is the unsupervised K-Means algorithm.

Figs. 3, 5, and 7 demonstrate the results for experiments where weighted cosine similarity $D_{\cos_a}$ was used as the distortion measure, while Figs. 4, 6, and 8 summarize experiments where weighted I-divergence D_{I_a} was used.

As the results demonstrate, the full HMRF-KMEANS algorithm outperforms the unsupervised K-Means baseline as well as the ablated versions of HMRF-KMEANS for both $D_{\cos_a}$ and D_{I_a}. Relative performance of KMEANS-I-C and KMEANS-I indicates that using supervision for initializing cluster representatives is highly beneficial, while the constraint-sensitive cluster assignment step does not lead to significant additional improvements for $D_{\cos_a}$. For D_{I_a}, KMEANS-I-C outperforms KMEANS-I on *News-Different-3* (Fig. 4) and *News-Similar-3* (Fig. 8) which indicates that incorporating constraints in the cluster assignment process is useful for these datasets. This result is reversed for *News-Related-3* (Fig. 6), implying that in some cases using constraints in the E-step may be unnecessary, which agrees with previous results on other domains [6]. However, incorporating supervised data in all the 3 stages of the algorithm in KMEANS-I-C-D, namely initialization, cluster assignment, and distance update, always leads to substantial performance improvement.

As can be seen from results for 0 pairwise constraints in Figs. 3-8, distance learning is beneficial even in the absence of any pairwise constraints, since it is able to capture the relative importance of the different attributes in the unsupervised data. In the absence of supervised data or when no constraints are violated, distance learning attempts to minimize the objective function by adjusting the weights given the distortion between the unsupervised datapoints and their corresponding cluster representatives.

In realistic application domains, supervision in the form of constraints would be in most cases provided by human experts, in which case it is important that any semi-supervised clustering algorithm performs well with a small number of constrains. KMEANS-I-C-D starts outperforming its variants and the unsupervised clustering baseline early on in the learning curve, and is therefore a very appropriate algorithm to use in actual semi-supervised data clustering systems.

Overall, our results show that the HMRF-KMEANS algorithm effectively incorporates labeled and unlabeled data in three stages, each of which improves the clustering quality.

5. RELATED WORK

A related unified model for semi-supervised clustering with constraints was recently proposed by Segal et al. [36]. Their model is a unified *Markov network* that combines a binary Markov network derived from pairwise protein interaction data and a Naive Bayes Markov network modeling gene expression data. Our proposed HMRF framework is more general than this formulation, since it works with a broad class of clustering distortion measures, including Bregman divergences and directional similarity measures. In contrast, the formulation of Segal et al. considers only a Gaussian cluster conditional probability distribution, which corresponds to having Mahalanobis distance as the underlying clustering distance measure. Additionally, the HMRF-KMEANS algorithm performs distance learning in the unified framework, which is not done in the Markov Network model.

The HMRF-KMEANS algorithm proposed in this paper is related to the EM algorithm for HMRF model-fitting proposed by Zhang et al. [40]. However, HMRF-KMEANS performs an additional step of distance learning in the M-step, which is not considered in the HMRF-EM algorithm. The discussion of the HMRF-EM algorithm was also restricted only to Gaussian conditional distributions, which has been generalized in our formulation.

There has been other research in semi-supervised clustering focusing individually on either constraint-based or distance-based semi-supervised clustering. COP-KMEANS is a constraint-based clustering algorithm that has a heuristically motivated objective function [38]. Our method, on the other hand, has an underlying probabilistic model based on Hidden Markov Random Fields. Bansal et al. [4] also proposed a framework for pairwise constrained clustering, but their model performs clustering using only the constraints, whereas our formulation uses both constraints and an underlying distortion measure between the points.

In recent work on distance-based semi-supervised clustering with pairwise constraints, Cohn et al. [13] used gradient descent for weighted Jensen-Shannon divergence in the context of EM clustering. Xing et al. [39] utilized a combination of gradient descent and iterative projections to learn a Mahalanobis distance for K-Means clustering. The Redundant Component Analysis (RCA) algorithm used only must-link constraints to learn a Mahalanobis distance using convex optimization [5]. Spectral learning is another recent method that utilizes supervision to transform the clustering distance measure using spectral methods [25]. All these distance learning techniques for clustering train the distance measure first using only supervised data, and then perform clustering on the unsupervised data. In contrast, our method integrates distance learning with the clustering process and utilizes both supervised and unsupervised data to learn the distortion measure.

6. FUTURE WORK

We have presented the general probabilistic framework for incorporating pairwise supervision into a prototype-based clustering algorithm, as well as two instantiations of that framework for particular distortion measures. There are several open issues that would be interesting to explore in future work.

Investigating alternative approaches to training distortion measures in the M-step of our algorithm may lead to improved performance of the algorithm. Our initial results as well as other recent work on distance learning for clustering [27, 8, 5, 39] suggest that transforming the data space can be highly beneficial for clustering quality. Therefore, we conjecture that developing alternative feature selection or feature extraction approaches, which perform other types of data space transformation using supervised data, is a promising direction for future work.

The weighted farthest-first algorithm for cluster initialization that we have described in Section 3.3 has proven itself very useful. We intend to explore theoretical implications of this initialization algorithm in the HMRF model, as well as develop alternative techniques that utilize both labeled and unlabeled data for initializing cluster representatives.

While we have used the ICM algorithm for constraint-sensitive cluster assignment in the HMRF model, other methods have also been proposed for this task, e.g. loopy belief propagation [36]. Extensive experimental comparison of these strategies would be informative for future work on iterative reassignment algorithms like HMRF-KMEANS in the HMRF framework. We also want to run experiments to study the sensitivity of the HMRF-KMEANS algorithm to the constraint violation parameters $\mathcal{W}$ and $\overline{\mathcal{W}}$, as done in Segal et al. [36].

Finally, we want to apply our algorithm to other application domains. One interesting problem in bioinformatics is to improve the quality of clustering genes with unknown functions by utilizing constraints between the genes derived from domain knowledge. Segal et al. [36] used constraints derived from protein-protein interactions while clustering gene expression data using Mahalanobis distance as the underlying distortion measure. We want to apply our HMRF-KMEANS algorithm to different kinds of gene representations, for which different clustering distance measures would be appropriate, e.g., Pearson's correlation would be an appropriate distortion measure for gene microarray data [20], I-divergence would be useful for the phylogenetic profile representation of genes [30], etc. We plan to run experiments for clustering these datasets using the HMRF-KMEANS algorithm, where the constraints will be inferred from protein interaction databases as well as from function pathway labels that are known for a subset of the genes.

7. CONCLUSIONS

We have introduced a theoretically motivated framework for semi-supervised clustering that employs Hidden Random Markov Fields (HMRFs) to utilize both labeled and unlabeled data in the clustering process. The framework can be used with a number of distortion measures, including Bregman divergences and directional measures, and it accommodates trainable measures that can be adapted to specific datasets. We introduced the HMRF-KMEANS algorithm that performs clustering in this framework and incorporates supervision in the form of pairwise constraints in all stages of the clustering algorithm: initialization, cluster assignment, and parameter estimation. We presented two instantiations of the algorithm based on two particular distortion measures that are popular for high-dimensional data: KL divergence and cosine similarity. Experimental evaluation has shown that the algorithm derived from the HMRF framework leads to improved cluster quality on realistic textual datasets over unsupervised clustering and ablations of the proposed approach.

8. ACKNOWLEDGMENTS

We would like to thank Srujana Merugu for insightful comments. This research was supported by the National Science Foundation under grants IIS-0117308 and ITR: IIS-0325116, and by a Faculty Fellowship from IBM Corporation.

9. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, New York, 1999.

[2] A. Banerjee, I. Dhillon, J. Ghosh, and S. Sra. Generative model-based clustering of directional data. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003)*, pages 19–28, 2003.

[3] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. In *Proceedings of the 2004 SIAM International Conference on Data Mining (SDM-04)*, 2004.

[4] N. Bansal, A. Blum, and S. Chawla. Correlation clustering. In *Proceedings of the 43rd IEEE Symposium on Foundations of Computer Science (FOCS-02)*, pages 238–247, 2002.

[5] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning distance functions using equivalence relations. In *Proceedings of 20th International Conference on Machine Learning (ICML-2003)*, pages 11–18, 2003.

[6] S. Basu, A. Banerjee, and R. J. Mooney. Semi-supervised clustering by seeding. In *Proceedings of 19th International Conference on Machine Learning (ICML-2002)*, pages 19–26, 2002.

[7] S. Basu, A. Banerjee, and R. J. Mooney. Active semi-supervision for pairwise constrained clustering. In *Proceedings of the 2004 SIAM International Conference on Data Mining (SDM-04)*, 2004.

[8] S. Basu, M. Bilenko, and R. J. Mooney. Comparing and unifying search-based and similarity-based approaches to semi-supervised clustering. In *Proceedings of the ICML-2003 Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, pages 42–49, 2003.

[9] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society, Series B (Methodological)*, 48(3):259–302, 1986.

[10] M. Bilenko and R. J. Mooney. Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003)*, pages 39–48, 2003.

[11] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100, 1998.

[12] Y. Boykov, O. Veksler, and R. Zabih. Markov random fields with efficient approximations. In *Proceedings of IEEE Computer Vision and Pattern Recognition Conference (CVPR-98)*, pages 648–655, 1998.

[13] D. Cohn, R. Caruana, and A. McCallum. Semi-supervised clustering with user feedback. Technical Report TR2003-1892, Cornell University, 2003.

[14] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

[15] A. Demiriz, K. P. Bennett, and M. J. Embrechts. Semi-supervised clustering using genetic algorithms. In *Artificial Neural Networks in Engineering (ANNIE-99)*, pages 809–814, 1999.

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977.

[17] I. S. Dhillon and Y. Guan. Information theoretic clustering of sparse co-occurrence data. In *Proceedings of the Third IEEE International Conference on Data Mining (ICDM-03)*, pages 517–521, 2003.

[18] I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42:143–175, 2001.

[19] B. E. Dom. An information-theoretic external cluster-validity measure. Research Report RJ 10219, IBM, 2001.

[20] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences, USA*, 95:14863–14848, 1998.

[21] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–742, 1984.

[22] J. M. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. Unpublished manuscript, 1971.

[23] D. Hochbaum and D. Shmoys. A best possible heuristic for the k-center problem. *Mathematics of Operations Research*, 10(2):180–184, 1985.

[24] T. Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML-99)*, pages 200–209, 1999.

[25] S. D. Kamvar, D. Klein, and C. D. Manning. Spectral learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-2003)*, pages 561–566, 2003.

[26] M. Kearns, Y. Mansour, and A. Y. Ng. An information-theoretic analysis of hard and soft assignment methods for clustering. In *Proceedings of 13th Conference on Uncertainty in Artificial Intelligence (UAI-97)*, pages 282–293, 1997.

[27] D. Klein, S. D. Kamvar, and C. Manning. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In *Proceedings of the The Nineteenth International Conference on Machine Learning (ICML-2002)*, pages 307–314, 2002.

[28] J. Kleinberg and E. Tardos. Approximation algorithms for classification problems with pairwise relationships: Metric labeling and Markov random fields. In *Proceedings of the 40th IEEE Symposium on Foundations of Computer Science (FOCS-99)*, pages 14–23, 1999.

[29] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[30] E. M. Marcotte, I. Xenarios, A. van der Bliek, and D. Eisenberg. Localizing proteins in the cell from their phylogenetic profiles. *Proceedings of the National Academy of Science*, 97:12115–20, 2000.

[31] K. V. Mardia and P. Jupp. *Directional Statistics*. John Wiley and Sons Ltd., 2nd edition, 2000.

[32] R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 355–368. MIT Press, 1998.

[33] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39:103–134, 2000.

[34] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo,CA, 1988.

[35] F. C. N. Pereira, N. Tishby, and L. Lee. Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL-93)*, pages 183–190, 1993.

[36] E. Segal, H. Wang, and D. Koller. Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, 19:i264–i272, July 2003.

[37] A. Strehl, J. Ghosh, and R. Mooney. Impact of similarity measures on web-page clustering. In *Workshop on Artificial Intelligence for Web Search (AAAI 2000)*, pages 58–64, July 2000.

[38] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl. Constrained K-Means clustering with background knowledge. In *Proceedings of 18th International Conference on Machine Learning (ICML-2001)*, pages 577–584, 2001.

[39] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems 15*, pages 505–512, Cambridge, MA, 2003. MIT Press.

[40] Y. Zhang, M. Brady, and S. Smith. Hidden Markov random field model and segmentation of brain MR images. *IEEE Transactions on Medical Imaging*, 20(1):45–57, 2001.