Combinatorial Markov Random Fields

Ron Bekkerman[1], Mehran Sahami[2], and Erik Learned-Miller[1]

[1] Department of Computer Science, University of Massachusetts, Amherst MA 01002,
{ronb|elm}@cs.umass.edu,
[2] Google Inc., 1600 Amphitheatre Parkway, Mountain View CA 94043,
sahami@google.com

Abstract. A *combinatorial random variable* is a discrete random variable defined over a combinatorial set (e.g., a power set of a given set). In this paper we introduce *combinatorial Markov random fields (Comrafs)*, which are Markov random fields where some of the nodes are combinatorial random variables. We argue that Comrafs are powerful models for unsupervised and semi-supervised learning. We put Comrafs in perspective by showing their relationship with several existing models. Since it can be problematic to apply existing inference techniques for graphical models to Comrafs, we design two simple and efficient inference algorithms specific for Comrafs, which are based on combinatorial optimization. We show that even such simple algorithms consistently and significantly outperform Latent Dirichlet Allocation (LDA) on a document clustering task. We then present Comraf models for semi-supervised clustering and transfer learning that demonstrate superior results in comparison to an existing semi-supervised scheme (constrained optimization).

1 Introduction

Three decades have passed since McGurk and MacDonald published their work [1] revealing the multi-modal nature of speech perception: sound and moving lips compose one system, so to better process audio signals, an audio/video interaction should be modeled. Since then, machine learning researchers have widely exploited data multi-modality, using many approaches, such as multi-modal neural networks [2], multivariate information bottleneck [3], and more recently multi-view expectation maximization [4] and multi-way distributional clustering [5].

Multi-modality plays an important role in unsupervised learning; given no class labels, learning results mostly depend on data representation. For example, one cannot expect a system to cluster documents by topic if only their *lengths* are given. However, when documents are represented as bags of *words*, meaningful clustering can be built. Moreover, if in addition to bags of words, another representation based on documents' *authorship* is obtained, the two modalities show different angles of documents' topicality and thus provide useful structure to documents' representation that can be leveraged during learning.

In many real-world situations multiple modalities of data can be easily observed. Indeed, consider an email inbox, where in addition to message bodies, one can observe subject lines, names of senders and recipients, markup items, attachments etc. Nevertheless, early multi-modal systems rarely went beyond two modalities (documents/words, audio/video, genes/samples, etc.). Currently, with the availability of massive computational power, using more than two modalities is a feasible and attractive research opportunity.

In many cases, each modality interacts differently with the others, with some interactions being negligibly weak. Hence, when many modalities are available (each of which having its own interaction pattern with the others), we can construct a graph representation of the modalities and their interactions. In previous work, Friedman et al. [3] use a Bayesian network to define *input* and *output* spaces in the multivariate Information Bottleneck; Bekkerman et al. [5] use a *pairwise interaction graph* to describe dependencies between the modalities. In both those studies, the graph is an auxiliary, descriptive component of the model.

Our approach uses the *Markov random field (MRF)* formalism (see, e.g., [6]). In Section 2, we propose a *combinatorial Markov random field (Comraf)*, which allows us to model each modality of the data as a single *combinatorial random variable* in the MRF graph, with edges representing probabilistic interactions between the modalities. Comraf models are (a) compact – the number of nodes in a Comraf is the order of the number of modalities, which allows for easier *model learning*; and (b) data-driven – no generative assumptions are made, which minimizes the model's bias. The main contribution of this work is to present a general framework for multi-modal learning, which is based on the *most probable explanation (MPE) inference* in a Comraf. For unsupervised learning, we show that Comrafs are a general framework that subsumes a number of existing models as special cases (Section 3) and allows us to also explore new modeling possibilities for other learning tasks, such as semi-supervised clustering and transfer learning (Section 4). We show that Comrafs lend themselves to naturally modeling multi-model data, obtaining strong empirical results (Section 5).

2 Combinatorial MRFs

Definition 1. *A* combinatorial random variable *(or* combinatorial r.v.*) X^c is a discrete random variable defined over a combinatorial set.*

A *combinatorial set* in mathematical parlance means a set of all subsets, partitionings, permutations etc. of a given finite set. To capture this intuition, we define a finite set A as *combinatorial* if its size is exponential with respect to another finite set B, i.e. $\log |A| = O(|B|)$. As an example, a combinatorial r.v. X^c can be defined over all the outcomes of *lotto 6 of 49*, in which 6 balls are selected from 49 enumerated balls to produce an outcome of the lottery. In this case, set B consists of 49 balls, while set A consists of $\binom{49}{6}$ possible choices of 6 balls from B. In a *fair* lottery, the distribution of X^c is uniform: each outcome is drawn with probability $1/\binom{49}{6}$. However, in an unfair lottery, some outcomes are more probable than others.

From the theoretical perspective, a combinatorial r.v. behaves exactly as an ordinary discrete random variable with a finite domain. However, from the practical point of view, a combinatorial r.v. is different: in most real-world cases, the event space of X^c is so large that the distribution $P(X^c)$ cannot be explicitly specified. Moreover, the MPE task for combinatorial r.v.'s can be computationally hard. Considering an unfair lottery example, in which the distribution of X^c is flat (close to uniform), say, the probability of value $\{7, 23, 29, 35, 48, 49\}$ is 0 and the probability of value $\{4, 18, 28, 37, 39, 43\}$ is $2/\binom{49}{6}$, while the rest of the values still have the probability $1/\binom{49}{6}$. An exponentially long sampling process is required to detect the most probable value.

It is easy to come up with other examples of combinatorial r.v.'s: all the possible translations of a sentence, orderings in a ranked list of retrieved documents, etc. In this paper, we consider combinatorial r.v.'s over all *partitionings* of a given set. In most complex systems random variables interact with each other. Such interactions are usually represented in a directed or undirected graphical model. In multi-modal systems, which are the focus of our paper, interactions between modalities are symmetric, so the undirected case is more appealing.

A *Markov random field (MRF)* is a model (G, P), where G is an undirected graph whose nodes $\mathbf{X} = \{X_1, \ldots, X_m\}$ represent random variables and whose edges $\mathbf{E}$ denote interactions between these variables. P is a joint probability distribution defined over $\mathbf{X}$. The *Markov property* holds in this model.

Definition 2. *A* combinatorial Markov random field (Comraf) *is an MRF, at least one node of which is a combinatorial random variable.*

2.1 MPE inference in Comrafs

An *inference* procedure in MRFs answers questions about the model, such as what is the most likely assignment $\mathbf{x}^* = \{x_1^*, \ldots, x_m^*\}$ to variables $\{X_1, \ldots, X_m\}$ (i.e. MPE). Naturally, answering most of such questions is NP-hard since it potentially requires considering every possible assignment. Thus, most inference techniques fall into the category of approximation methods.

The Hammersley-Clifford theorem [7] states that the joint distribution over nodes of an MRF is a Gibbs distribution: $P(\mathbf{x}) = \frac{1}{Z_\mathbf{f}} \exp \sum_i f_i(\mathbf{x})$, where $f_i(\mathbf{x})$ are arbitrary potential functions defined over cliques in G, and $Z_\mathbf{f}$ is a normalization factor called a partition function. Unsupervised learning problems are usually solved using the *maximum likelihood (ML)* framework (see, e.g. [8]), where model parameters that best explain the data are estimated. Most ML methods deal with approximating $Z_\mathbf{f}$, which is generally a hard task, because $Z_\mathbf{f}$ depends on the particular choice of f_i's and is a sum over all the possible configurations. However, in our setting the potentials f_i are fixed for each clique, the partition function $Z_\mathbf{f}$ becomes a constant, so $\log P(\mathbf{x}) \propto \sum_i f_i(\mathbf{x})$. Thus, for MPE, it is sufficient to directly optimize:

$$\mathbf{x}^* = \arg\max_{\mathbf{x}} P(\mathbf{x}) = \arg\max_{\mathbf{x}} \sum_i f_i(\mathbf{x}). \tag{1}$$

This relatively simple formulation is still quite powerful, as it allows us to use a wide variety of potential functions that might be too complicated to use in the general setting where the partition function still needs to be approximated.

3 Unsupervised learning with Comrafs

To illustrate the power of the Comraf framework, we initially focus on unsupervised learning (e.g., data clustering) and show how several existing clustering schemes are specific instances on Comrafs. Let $s_1, s_2, ..., s_N$ be a dataset of N i.i.d. samples drawn from some discrete distribution. Let $\mathcal{X} = \{x_1, x_2, ..., x_n\}$ be the set of n unique values comprising the event space from which samples s_i are drawn. We now define a random variable X such that $P(X = x_i)$ is given by the empirical frequencies of samples with value x_i in the dataset (i.e., X has a multinomial distribution estimated using maximum likelihood).

Define a *hard clustering* $\tilde{x}^c$ to be a partitioning of $\mathcal{X}$. Let $\mathcal{X}^c = \{\tilde{x}^c_1, \tilde{x}^c_2, ..., \tilde{x}^c_K\}$ be the combinatorial set of all K partitionings of $\mathcal{X}$, where K is exponential in the size of $\mathcal{X}$. We will refer to the subsets of the j-th partitioning $\tilde{x}^c_j$ as $\{\tilde{x}_{j,1}, \tilde{x}_{j,2}, ..., \tilde{x}_{j,k_j}\}$. That is, the first subscript is the index of the particular partitioning, and the second subscript is the subset within that partitioning.

Define $\tilde{X}_j$ to be a random variable over the subsets (*clusters*) in a partitioning $\tilde{x}^c_j$, with the probability of a selected cluster being the probability of choosing any one of its elements, that is, $P(\tilde{X} = \tilde{x}_{j,i}) = \sum_{x \in \tilde{x}_{j,i}} P(x)$. Finally, define $\tilde{X}^c$ to be a combinatorial r.v. with the event space $\mathcal{X}^c$. In this work, we shall use parallel notation for different modalities of data, replacing the "x's" in the above notation with variables appropriate for the data source. For example, w_i would represent a specific word in a dataset, $\tilde{w}^c$ a partitioning of words, and so on.

Interactions between combinatorial r.v.'s (possibly, with ordinary r.v.'s) are represented by edges in a Comraf graph. To use the objective from Equation (1), we should choose relevant cliques in the Comraf graph and define potential functions over these cliques. To make the inference feasible, we consider only the smallest cliques, i.e. adjacent pairs. Since our inference objective allows using complicated potential functions (see Section 2.1), we use the *mutual information (MI)* between r.v.'s defined over values of adjacent nodes. Let $\tilde{x}^c_i$ and $\tilde{y}^c_j$ be such values (particular partitionings of two modalities). A potential is then defined:

$$f(\tilde{x}^c_i, \tilde{y}^c_j) = I(\tilde{X}_i; \tilde{Y}_j) = \sum_{i',j'} P(\tilde{x}_{i,i'}, \tilde{y}_{j,j'}) \log \frac{P(\tilde{x}_{i,i'}, \tilde{y}_{j,j'})}{P(\tilde{x}_{i,i'}) P(\tilde{y}_{j,j'})}.$$

Our motivation for choosing MI as a potential function is as follows: a linear combination of MI terms has traditionally been used as a clustering criterion, both in 1-way clustering methods, such as Information Bottleneck (IB) [9], and in 2-way methods [10]. Friedman et al. [3] generalize the IB clustering criterion to a multivariate case: in place of MI, they use Multi-Information, which naturally factors over a directed graphical model. With little effort, we can show that Multi-Information also factors over a tree-structured undirected graphical model, reducing to a sum of pairwise MI terms defined over edges of the tree. However, in the case of an arbitrary Comraf graph, the Multi-Information can only be *approximated* by a sum of pairwise MI terms. Estimating the quality of such an approximation remains an open question that we will address in future work. Presently, we show how existing models can be cast as Comrafs:

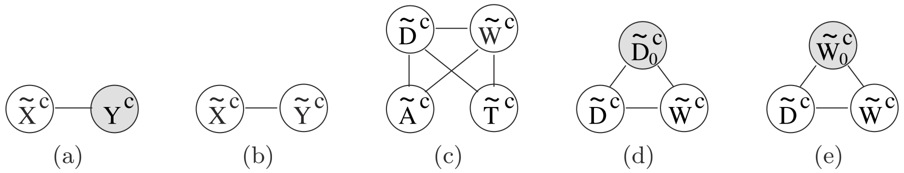


Fig. 1. Comraf graphs for: (a) hard version of Information Bottleneck; (b) information-theoretic co-clustering; (c) 4-way MDC; (d) semi-supervised clustering; (e) clustering with transfer learning.

A hard version of Information Bottleneck [9] is a special case of a Comraf. In IB, a clustering $\tilde{x}^{c*}$ is constructed that maximizes information about a variable Y (and minimizes information about X), i.e. $\tilde{x}^{c*} = \arg\max_{\tilde{x}^c_j}(I(\tilde{X}_j;Y) - \beta I(\tilde{X}_j;X))$, where β is a Lagrange multiplier. The compression constraint $I(\tilde{X}_j;X)$ can be omitted if the number of clusters is fixed: $|\tilde{x}^c_j| = k$. Consider graph G in Figure 1(a), where a shaded Y^c represents an *observed* variable.[3] On the only clique in G we define one potential which is the mutual information $I(\tilde{X}_j;Y)$. The MPE objective is then: $\tilde{x}^{c*} = \arg\max_{\tilde{x}^c_j} P(\tilde{x}^c_j, y^c) = \arg\max_{\tilde{x}^c_j} I(\tilde{X}_j;Y)$.

Information-theoretic co-clustering [10] is a task of simultaneously clustering documents $\mathcal{X}$ and their words $\mathcal{Y}$, while minimizing the information loss $I(X;Y) - I(\tilde{X}_j, \tilde{Y}_j)$ under the constraint $|\tilde{x}^c_j| = k_1$ and $|\tilde{y}^c_j| = k_2$. Note that $I(X;Y)$ is a constant for a given dataset. This scheme is a special case of a Comraf as well: given graph G in Figure 1(b), in analogy to the Comraf model of IB, we define the only potential $I(\tilde{X}_j;\tilde{Y}_j)$. Then the information-theoretic co-clustering can be represented as an MPE procedure in the Comraf: $(\tilde{x}^{c*}, \tilde{y}^{c*}) = \arg\max_{\tilde{x}^c_j,\tilde{y}^c_j} P(\tilde{x}^c_j, \tilde{y}^c_j) = \arg\max_{\tilde{x}^c_j,\tilde{y}^c_j} I(\tilde{X}_j;\tilde{Y}_j)$.

Multi-way distributional clustering (MDC) [5] is a generalization of [10], where the data has a number of interdependent modalities (such as documents, words, authors, titles, etc.). Interactions between the modalities are represented using a *pairwise interaction graph* that has no probabilistic interpretation. Actually, these interactions can be represented in a Comraf, where the modalities are combinatorial r.v.'s $\tilde{\mathbf{X}}^{\mathbf{c}} = \{\tilde{X}^c_1, \ldots, \tilde{X}^c_m\}$ that are nodes in a graph G with edges $\mathbf{E}$. The MPE scheme is then:

$$\tilde{\mathbf{x}}^{\mathbf{c}*} = \arg\max_{\tilde{\mathbf{x}}^{\mathbf{c}}_j} P(\tilde{\mathbf{x}}^{\mathbf{c}}_j) = \arg\max_{\tilde{\mathbf{x}}^{\mathbf{c}}_j} \sum_{(\tilde{X}^c_i,\tilde{X}^c_{i'})\in\mathbf{E}} I(\tilde{X}_{i,j};\tilde{X}_{i',j}). \tag{2}$$

Here the first subscript is the index of a combinatorial r.v., while the second subscript is the index of this r.v.'s particular value (a partitioning). Equation (2) is equivalent to the MDC objective proposed in [5]. An example Comraf graph for a 4-way MDC (that corresponds to simultaneously clustering documents, words, authors and titles) is shown in Figure 1(c).

[3] For discussion on observed variables see Section 4.

We note that by casting IB and Information-theoretic co-clustering as Comrafs, we not only show the generality of the framework, but also demonstrate that the generalization of these methods to additional modalities of data is naturally accomplished via Comrafs. In the case of MDC, viewing this model as a Comraf allows us to consider generalizations to other tasks, such as semi-supervised learning via the introduction of observed variables in the model.

3.1 Clustering as inference in a Comraf

Due to unique characteristics of combinatorial r.v.'s, it is problematic to apply existing inference algorithms to Comrafs. Here we propose an inference method specific for Comrafs, which is based on combinatorial optimization. We then craft two simple and efficient inference algorithms based on the proposed method.

Given that a variable X has n values that are clustered into k clusters, the combinatorial r.v. $\tilde{X}^c$ has k^n values, all of which can be represented as points in an n-dimensional lattice L: a point $\tilde{x}^c = (i_1, i_2, \ldots, i_n)$ corresponds to the fact that value x_1 of X belongs to cluster i_1, value x_2 belongs to cluster i_2, $\ldots$, value x_n belongs to cluster i_n.[4] In the lattice L there is a (possibly non-unique) point $\tilde{x}^{c*} = (i_1^*, i_2^*, \ldots, i_n^*)$ which is most likely. Since the lattice consists of an exponential number of points, the task of finding the most likely point can be computationally hard. We will attempt to approximate the solution using a quasi-random walk in the lattice. Let us start with two definitions.

Definition 3. *A* transaction $(\ldots, i_j, \ldots) \rightarrow (\ldots, i'_j, \ldots)$*, where* $i_j \neq i'_j$*, is an elementary operation in traversing the lattice* L *of possible clusterings, in which* x_j *is moved from cluster* i_j *to cluster* i'_j*.*

Definition 4. *A* path *in* L *is a sequence of transactions. A path is called* advantageous *if it leads to a more likely clustering, otherwise it is* disadvantageous*.*

Note that we can view both splits and mergers of clusters as transactions. A split of a cluster $i_{j'}$ is a transaction $(\ldots, i_{j'}, \ldots) \rightarrow (\ldots, i'_{j'}, \ldots)$, where $\exists j \neq j' : i_{j'} = i_j$ and $\forall j \neq j' : i'_{j'} \neq i_j$. That is, cluster $i_{j'}$ contained at least two elements (x_j and $x_{j'}$), one of which ($x_{j'}$) has been transferred into a newly created cluster $i'_{j'}$. A merger of clusters $i_{j'}$ and $i'_{j'}$ is a transaction $(\ldots, i_{j'}, \ldots) \rightarrow (\ldots, i'_{j'}, \ldots)$, where $\exists j \neq j' : i'_{j'} = i_j$ and $\forall j \neq j' : i_{j'} \neq i_j$, i.e. cluster $i_{j'}$ contained only one element that has been added to the existing cluster $i'_{j'}$ so that the cluster $i_{j'}$ does not exist anymore. These operations will help us to represent both agglomerative (bottom-up) and divisive (top-down) clustering schema as inference in Comrafs.

By applying splits, mergers and other transactions, we construct paths in the lattice of possible clusterings. Thus, to approximate the MPE of a combinatorial r.v. $\tilde{X}^c$, we apply the simplest, greedy combinatorial optimization algorithm—*hill climbing*:[5] we attempt to construct a path in L which is as advantageous as possible on each step, given the available computational resources.

[4] For now, we consider only *hard* clustering, where $P(i_j|x_j) = 1$ for a value x_j assigned to cluster i_j. Generalization of the Comraf model to *soft* clustering is our future task.

[5] More complex algorithms, such as Branch and Bound, while applicable, may be infeasible to use because of their high computational complexity.

Algorithm 1 A template of an MPE procedure in Comrafs.

```
Input:
 $G$ – Comraf graph of nodes $\{\tilde{X}_1^c, \ldots, \tilde{X}_m^c\}$ and edges $\mathbf{E}$
 $P(X_i, \ldots, X_m)$ – joint probability distribution of data, factorized over $G$
 $l$ – number of optimization iterations
Output:
 Most likely $\tilde{x}_{1,l}^c, \ldots, \tilde{x}_{m,l}^c$

 Initialization:
 for $i = 1, \ldots, m$ do
   Select a point in $L_i$ to be an initial value $\tilde{x}_{i,0}^c$ of $\tilde{X}_i^c$
 Compute the initial joint $P(\tilde{X}_{1,0}, \ldots, \tilde{X}_{m,0})$, factorized over $G$
 Main loop:
 for $j = 1, \ldots, l$ do
   Select variable $\tilde{X}_{i'}^c$ for optimization
   Construct advantageous path $\left(\tilde{x}_{i',j-1}^c \rightarrow \tilde{x}_{i',j}^c\right)$ in $L_{i'}$
   For all $i \neq i'$ do $\tilde{x}_{i,j}^c = \tilde{x}_{i,j-1}^c$
```

In a Comraf that has more than one combinatorial r.v., the Comraf inference algorithm becomes a variation of the *Iterative Conditional Mode (ICM)* method [11]. ICM optimizes each node of an MRF iteratively (in a round-robin fashion), given its Markov blanket. At an ICM iteration applied to a node $\tilde{X}_i^c$, the MPE objective from Equation (2) with $O(|\mathbf{X}|^2)$ terms is reduced to:

$$\tilde{x}_i^{c*} = \arg\max_{\tilde{x}_{i,j}^c} \sum_{i':\ (\tilde{X}_i^c, \tilde{X}_{i'}^c) \in \mathbf{E}} I(\tilde{X}_{i,j}; \tilde{X}_{i',j}) \tag{3}$$

that sums over only $O(|\mathbf{X}|)$ neighbors of $\tilde{X}_i^c$.

A template pseudo-code for the MPE approximation in a Comraf is given in Algorithm 1. For each combinatorial r.v. $\tilde{X}_i^c$ in the Comraf, we first select and fix its initial value as a point in the lattice L_i. We then round-robin over each $\tilde{X}_i^c$, for which we search for an advantageous path in L_i. When this path is constructed, we fix its destination point to be a new value of $\tilde{X}_i^c$ and move to another node. We repeat this procedure l times. To transform this template into an actual algorithm, we need to make the following choices:

- Selecting initial values for each combinatorial r.v. in the Comraf. Either random assignment of data points into k clusters or an assignment of all data points into one cluster are two simple choices, while other methods (such as those incorporating prior knowledge) are possible.
- Determining an ordering for variables in the optimization procedure. One obvious approach is a plain or weighted round-robin, but more sophisticated choices can also be made.
- Constructing an advantageous path in L. A greedy method would increase the likelihood with each transaction, leading to a local maximum of the objective. However, we could also consider a stochastic approach in which

some disadvantageous transactions are tolerated assuming that they may lead closer to the global maximum.

The latter point is of especial importance. We propose two algorithms for constructing advantageous paths. In both, we first split or merge clusters in order to meet the traditional requirement on the number of clusters. Then, in the *sequential* algorithm, we iterate over each data point in some ordering, and assign it into its best cluster (the one for which the objective is maximized). In the *randomized* algorithm, we repeat the following step a predefined number of times:[6] we uniformly at random select a data point x_i and a cluster $\tilde{x}_j$, and assign x_i into $\tilde{x}_j$ if this transaction improves the objective.

4 Semi-supervised and transfer learning with Comrafs

The Comraf model is a convenient framework for performing semi-supervised and transfer learning. Prior to presenting details of particular Comrafs, let us define the concepts of hidden and observed states in the Comraf model. A combinatorial r.v. is *hidden* if it can take any value from its event space. A combinatorial r.v. is *observed* if its value is preset and fixed.

4.1 Semi-supervised clustering with Comrafs

Semi-supervised clustering is a clustering task that takes advantage of labeled examples. Usually, semi-supervised clustering is performed when the number of available labeled examples is not sufficient to construct a good classifier (e.g., the constructed classifier would overfit), or when the the labeled data is noisy or skewed to a few classes. Assuming that *most* of the labeled data is accurate, our goal is to incorporate it into the (unsupervised) Comraf model.

In this paper, we consider a uni-labeled case: each labeled data point $x_i|_{i=1}^n$ belongs to one ground truth category $t_j|_{j=1}^k$. We propose an *intrinsic* Comraf approach for incorporating labeled data into clustering (by introducing observed nodes to a Comraf graph), and compare it with an existing *constrained optimization* scheme.

Intrinsic approach. Comrafs offer an elegant method for incorporating labeled data, which does not require any significant changes in the model. First, note that labels define a natural partitioning of the labeled data: for each label t_j let $\tilde{x}_{0j}$ be a subset of $\mathcal{X}$ labeled with t_j, i.e. $\tilde{x}_{0j} = \{x_i | t_i = t_j\}$. We now define a r.v. $\tilde{X}_0$ over the partitioning $\tilde{x}_0^c = \{\tilde{x}_{0j} | j = 1, \ldots, k\}$, and we also define a combinatorial r.v. $\tilde{X}_0^c$ over all the possible partitionings of the set $\mathcal{X}$. Since the partitioning $\tilde{x}_0^c$ is *given* to us, the variable $\tilde{X}_0^c$ is *observed*, with $\tilde{x}_0^c$ being its fixed value. Observed combinatorial random variables appear shaded on a Comraf graph – see, e.g., Figure 1(c). The objective function from Equation (3) and the MPE inference procedure remain unchanged (with the only difference being

[6] Equal (for fair comparison) to the number of iterations in the sequential algorithm.

that there is no need in optimizing the observed nodes): at each ICM iteration the current node is optimized with respect to the *fixed* values of its neighbors, whereas the values of the observed nodes are fixed by definition.

Constrained optimization. Wagstaff and Cardie [12] perform semi-supervised clustering with two types of boolean constraints. The *must-link* constraint *ml* equals 1 if two equally labeled data points are assigned into different clusters; the *cannot-link* constraint *cl* equals 1 if two differently labeled data points are assigned into the same cluster. A clustering objective function incorporates the constraints, e.g. in Comrafs (Equation (3)) for each combinatorial r.v. $\tilde{X}_i^c$ it is:

$$\tilde{x}_i^{c*} = \arg\max_{\tilde{x}_{i,j}^c} \sum_{i':\ (\tilde{X}_i^c, \tilde{X}_{i'}^c) \in \mathbf{E}} I(\tilde{X}_{i,j}; \tilde{X}_{i',j}) - \sum_{i'} w_{i,i'}\ ml_{i,i'} - \sum_{i'} w_{i,i'}\ cl_{i,i'},$$

where $w_{i,i'}$ are weights that we set at $+\infty$, which means that all constraints must be satisfied. Note that in a general case we are free to choose any non-negative weights. In order to fairly compare two semi-supervised methods, for both of them we must use the same underlying clustering algorithm. We use the sequential MPE inference algorithm (see Section 3.1) in both cases.

4.2 Transfer learning with Comrafs

Transfer learning is the problem of applying the knowledge learned in one task to effectively solve another learning task. In this paper, we represent the acquired knowledge as a partitioning $\tilde{y}_0^c$ pre-built for data $\mathcal{Y}$ that can be used for constructing a partitioning $\tilde{x}^c$ of data $\mathcal{X}$. We note that the intrinsic scheme for semi-supervised clustering presented above allows us to directly use labeled data not from $\mathcal{X}$ but rather from *another* collection $\mathcal{Y}$. Thus, in analogy to the semi-supervised case, we introduce an observed combinatorial r.v. $\tilde{Y}_0^c$ with a fixed value $\tilde{y}_0^c$. During the inference process, we construct $\tilde{x}^{c*}$ that maximizes information about $\tilde{y}_0^c$, while applying the same objective function as in Equation (3).

5 Experimentation

Following [10], we use *micro-averaged accuracy* for evaluation of our clustering methods. Let $\tilde{x}^c$ be a clustering of the data $\mathcal{X}$. Let T be the set of ground truth categories. We fix the number of clusters to match the number of categories $|\tilde{x}^c| = |T| = k$. For each cluster $\tilde{x}_j$, let $\gamma_T(\tilde{x}_j)$ be the maximal number of $\tilde{x}_j$'s elements that belong to one category. Then, accuracy $Acc(\tilde{x}_j, T)$ of a cluster $\tilde{x}_j$ with respect to T is defined as $Acc(\tilde{x}_j, T) = \gamma_T(\tilde{x}_j)/|\tilde{x}_j|$. The micro-averaged accuracy of a clustering $\tilde{x}^c$ is:

$$Acc(\tilde{x}^c, T) = \frac{\sum_{j=1}^k \gamma_T(\tilde{x}_j)}{\sum_{j=1}^k |\tilde{x}_j|} = \frac{\sum_{j=1}^k \gamma_T(\tilde{x}_j)}{|\mathcal{X}|}. \qquad (4)$$

We evaluate the Comraf models on six text datasets. In addition to the standard benchmark 20 Newsgroups dataset (20NG) we use five real-world email

Table 1. Left 3 columns: statistics on the datasets. Right 3 columns: clustering accuracies (with standard error of the mean) for LDA and two Comraf algorithms. We report on only one of the two lengthy 20NG experiments with Comrafs.

Dataset	Size (num	Num of dis-	Num of	LDA	Comraf	Comraf
	of docs)	tinct words	classes		(sequent)	(random)
acheyer	664	2863	38	44.3±0.4	47.8±0.4	47.1±0.4
mgervasio	777	3207	15	38.5±0.4	42.4±0.4	44.0±1.0
mgondek	297	1287	14	68.0±0.8	75.9±0.6	75.5±0.5
kitchen-l	4015	15579	47	36.7±0.3	42.4±0.6	41.6±0.8
sanders-r	1188	5966	30	63.8±0.4	67.4±0.3	67.6±0.3
20NG	19997	39764	20	56.7±0.6	69.5±0.7	

directories. Three of them belong to participants in the CALO project[7] and the other two belong to former Enron employees.[8] We preprocess the data following Bekkerman et al. [5]. Table 1 provides basic statistics on the six datasets.

We report on the clustering accuracy averaged over ten independent runs on the email datasets and five runs on 20NG. For the (unsupervised) clustering task we use the Comraf graph from Figure 1(b), with $\tilde{X}^c$ for document clusterings and $\tilde{Y}^c$ for word clusterings. We apply agglomerative clustering to documents and divisive clustering to words. We compare two Comraf algorithms proposed in Section 2.1 with Latent Dirichlet Allocation [8], a popular generative clustering model. We use Xuerui Wang's LDA implementation [13] that applies Gibbs sampling with 10000 sampling iterations.[9] As shown in Table 1, both Comraf algorithms outperform LDA on all five email datasets and by more than 12% on an absolute scale on 20NG. Interestingly, both Comraf algorithms show almost identical results which suggests that the method of constructing advantageous paths does not matter a lot, as soon as the number of iterations is the same.

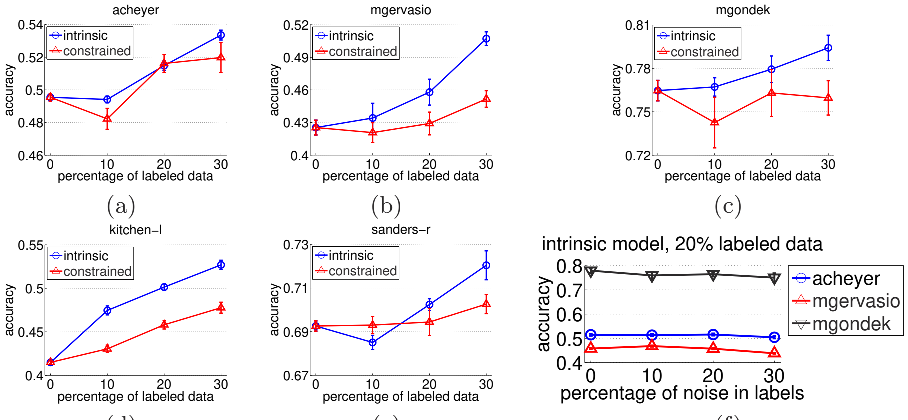
Figure 1(d) shows a Comraf graph for the intrinsic scheme of semi-supervised clustering (see Section 4). Together with a node $\tilde{D}^c$ over document clusterings and a node $\tilde{W}^c$ over word clusterings, we introduce an observed node $\tilde{D}^c_0$, whose value $\tilde{d}^c_0$ is a given partitioning of labeled documents. Our objective derived from Equation (2) is: $(\tilde{d}^{c*}, \tilde{w}^{c*}) = \arg\max_{\tilde{d}^c_j, \tilde{w}^c_j} I(\tilde{D}_j; \tilde{W}_j) + I(\tilde{D}_j; \tilde{D}_0) + I(\tilde{W}_j; \tilde{D}_0)$.

We conduct the following experiment: for each email dataset, we uniformly at random select 10%, 20%, or 30% of the data and refer to it as labeled examples while the rest of the data is considered unlabeled. We apply both intrinsic and constrained methods on the three setups and plot the accuracy (calculated on unlabeled data only) vs. the percentage of labeled data used. The results are shown in Figure 2. As we can see from the figure, both methods unsurprisingly improve the unsupervised results, while the intrinsic Comraf method usually

[7] `http://www.ai.sri.com/project/CALO`

[8] The preprocessed Enron email datasets can be obtained from `http://www.cs.umass.edu/~ronb/enron_dataset.html`.

[9] We also tried David Blei's LDA-C [8] that implements variational approximation and obtained significantly inferior accuracy.



(a) (b) (c)

(d) (e) (f)

Fig. 2. Plots (a)-(e): comparing accuracies of the semi-supervised Comraf and the constrained optimization method on five email datasets. Plot (f): the semi-supervised Comraf's resistance to noise in labeled data.

outperforms the constrained method. On 20NG, we select 10% of data to be labeled. The constrained method obtains 74.8±0.6% accuracy, while the intrinsic method obtains $78.9 \pm 0.8\%$ accuracy (over 5% and 9% absolute improvement to the unsupervised result, respectively). For another experiment with a semi-supervised Comraf, see [14].

The intrinsic scheme is resistant to noise. To show this, we conduct the following experiment: on CALO datasets with 20%/80% labeled/unlabeled split, we arbitrarily corrupt labels of 10%, 20% and 30% of the labeled data. Figure 2(f) shows that clustering accuracy remains almost unchanged for all three datasets.

Our *transfer learning* experiments are set up as follows. We notice that in two of the CALO datasets (ACHEYER and MGERVASIO) similar topics are discussed. Our hypothesis is that *known* categories of one dataset can improve the clustering results on another dataset. To test this hypothesis, we first consider one dataset to be labeled, while the other one is unlabeled, and then vice versa. However, since the two datasets do not consist of the *same* documents, we decide to use *word* clusters of the labeled dataset. We first cluster words distributed over categories of the labeled dataset, as described in [15]. Then we introduce the constructed clustering as an observed node $\tilde{W}_0^c$ into the Comraf graph (see Figure 1(e)) and perform the inference. Using this scheme we improve the clustering accuracy on MGERVASIO by 3% absolute over unsupervised clustering. However, we do not see any change in the results on the ACHEYER dataset.

6 Conclusion and future work

In this paper, we have presented combinatorial MRFs and empirically shown their utility on fundamental problems of unsupervised clustering, semi-supervised clustering, and transfer learning. In our future work, we aim at applying Comrafs to non-textual domains, such as computer vision. The use of Comrafs is not limited to clustering problems only. We plan to apply Comrafs to ranking, filtering and other tasks. Another interesting research problem is *model learning* in Comrafs. While model learning is often infeasibly expensive in graphical models with thousands or millions of nodes, we have shown that useful Comraf models can still be extremely compact, which makes model learning feasible.

Acknowledgements

We are grateful to Polina Golland, Leslie Pack-Kaelbling, Victor Lavrenko, Uri Lerner, Chris Pal and Yoram Singer for fruitful discussions. Portions of this research were conducted while Ron Bekkerman was an intern at Google Inc. Also, this work was supported in part by the Center for Intelligent Information Retrieval, in part by the Defense Advanced Research Projects Agency (DARPA) under contract number HR0011-06 C-0023, and in part by NSF under a CAREER Award #0546666. Ron thanks his wife Anna for her constant support.

References

1. McGurk, H., MacDonald, J.: Hearing lips and seeing voices. Nature **264**(5588) (1976) 746–748
2. de Sa, V.: Unsupervised Classification Learning from Cross-Modal Environmental Structure. PhD thesis, University of Rochester (1994)
3. Friedman, N., Mosenzon, O., Slonim, N., Tishby, N.: Multivariate information bottleneck. In: Proceedings of UAI-17. (2001)
4. Bickel, S., Scheffer, T.: Multi-view clustering. In: Proceedings of ICDM-4. (2004)
5. Bekkerman, R., El-Yaniv, R., McCallum, A.: Multi-way distributional clustering via pairwise interactions. In: Proceedings of ICML-22. (2005) 41–48
6. Li, S.: Markov random field modeling in computer vision. Springer Verlag (1995)
7. Besag, J.: Spatial interaction and statistical analysis of lattice systems. Journal of the Royal Statistical Society **36**(2) (1974) 192–236
8. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. JMLR **3** (2003) 993–1022
9. Tishby, N., Pereira, F., Bialek, W.: The information bottleneck method (1999) Invited paper to the 37th Annual Allerton Conference.
10. Dhillon, I.S., Mallela, S., Modha, D.S.: Information-theoretic co-clustering. In: Proceedings of SIGKDD-9. (2003) 89–98
11. Besag, J.: On the statistical analysis of dirty pictures. Journal of the Royal Statistical Society **48**(3) (1986)
12. Wagstaff, K., Cardie, C.: Clustering with instance-level constraints. In: Proceedings of ICML-17. (2000)
13. McCallum, A., Corrada-Emmanuel, A., Wang, X.: Topic and role discovery in social networks. In: Proceedings of IJCAI-19. (2005) 786–791
14. Bekkerman, R., Sahami, M.: Semi-supervised clustering using combinatorial MRFs. In: Proceedings of ICML-23 Workshop on Learning in Structured Output Spaces. (2006)
15. Bekkerman, R., El-Yaniv, R., Tishby, N., Winter, Y.: Distributional word clusters vs. words for text categorization. JMLR **3** (2003) 1183–1208