Semi-supervised Clustering using Combinatorial MRFs

Ron Bekkerman RONB@CS.UMASS.EDU
Department of Computer Science,
University of Massachusetts, 140 Governors Drive, Amherst MA 01002

Mehran Sahami SAHAMI@GOOGLE.COM
Google Inc., 1600 Amphitheatre Parkway, Mountain View CA 94043

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 learning by showing their relationship with two existing models. We then present a Comraf model for semi-supervised clustering that demonstrates superior results in comparison to an existing semi-supervised scheme (constrained optimization).

1. Introduction

Graphical models have proven themselves to be a useful machine learning framework, showing excellent results in information retrieval (Metzler & Croft, 2005), natural language processing (Sha & Pereira, 2003), computer vision (Freeman et al., 2000), and a variety of other fields (Jordan, 2004). One benefit of using graphical models is the availability of black-box inference mechanisms; once a model is designed, it is usually straightforward to apply an existing optimization procedure to make inferences in the model.

Unsupervised learning tasks are often performed using generative graphical models (such as Bayesian networks). Practitioners traditionally make assumptions about the structure of the model based on domain knowledge, the need for computational tractability, or both. Such assumptions can potentially be inappropriate and thus introduce undesired bias into the model. Moreover, while in some cases it may be possible to learn the structure of the model from data, this can easily become infeasible without significant restrictions on the class of models being considered.

Despite the fact that generative models are likely to be biased, they remain a popular approach for unsupervised learning within the graphical model framework. Other types of graphical models for unsupervised learning are emerging.[1] This paper proposes a combinatorial MRF (Comraf): a non-generative graphical model which is an instance of a Markov Random Field (MRF), an undirected graphical model (see, e.g. Li, 1995). We show how Comrafs can be applied to clustering in general and to semi-supervised clustering in particular.

Combinatorial MRFs realize four basic principles: (a) **data-driven:** unlike generative models, Comrafs do not prescribe the intrinsic structure of the data, thereby limiting the bias; (b) **multi-modal:** Comrafs exploit the fact that the data usually has multiple views; (c) **compact:** each modality of the data is represented with *one* random variable – only interactions between modalities are explicitly modeled, making *model learning* easier; (d) **general:** the Comraf is a unified model that can be applied *as is* to various tasks, such as unsupervised and semi-supervised clustering, transfer learning, ranking, etc.

The rest of the paper is organized as follows. In Section 2 we propose the Comraf model. In Section 3 we discuss clustering with Comrafs and in Section 4 we present a Comraf for semi-supervised clustering. Section 5 describes our experimental setup; Section 6 provides empirical results. We conclude in Section 7.

Appearing in *Proceedings of the ICML Workshop on Learning in Structured Output Spaces*, Pittsburgh, PA, 2006.

[1] Some work has been done on non-standard application of existing types of graphical models, e.g. Friedman et al. (2001) use a Bayesian network for describing dependencies between variables in an information-theoretic clustering system.

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 jargon 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. $|A| = O(2^{|B|})$. 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 out of 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 finite support. However, from the practical point of view, a combinatorial r.v. is different: in most real-world cases, the support of X^c is so large that the distribution $P(X^c)$ cannot be explicitly specified. Moreover, the *Most Probable Explanation (MPE)* task for combinatorial r.v.'s can be computationally hard. Considering an unfair lottery example, if the distribution of outcomes is flat (close to uniform), the problem of identifying the most probable outcome is hard because the number of possible outcomes is very large. For instance, if the probability of one outcome $\{7, 23, 29, 35, 48, 49\}$ is 0 and the probability of another outcome $\{4, 18, 28, 37, 39, 43\}$ is $2/\binom{49}{6}$, while the rest of the outcomes still have the probability $1/\binom{49}{6}$, then an $O(2^{|B|})$ long sampling process is required to detect the most probable outcome.

It is easy to come up with other examples of combinatorial r.v.'s: over all the possible translations of a sentence, orderings in a ranked list of retrieved documents, positions in a chess game, 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 in 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 denote interactions between these variables. P is a joint probability distribution defined over the nodes of G. The Markov property holds in this graph.

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

2.1. Inference in Comrafs

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

The famous Hammersley-Clifford theorem (Besag, 1974) states that the joint distribution over nodes of an MRF is a Gibbs distribution:

$$P(\mathbf{x}) = \frac{1}{Z_{\mathbf{f}}} \exp \sum_i f_i(\mathbf{x}), \tag{1}$$

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. Note that $Z_{\mathbf{f}}$ depends on the particular choice of f's and is a sum over all the possible configurations. It is often intractable to directly compute $Z_{\mathbf{f}}$, so many inference techniques such as mean field approximation, variational methods etc. (see, e.g. Wiegerinck, 2000) deal with approximating $Z_{\mathbf{f}}$, which is generally a difficult task. However, if the potentials f_i are predefined and fixed for each clique, the partition function $Z_{\mathbf{f}}$ becomes a constant and then $\log P(\mathbf{x}) \propto \exp \sum_i f_i(\mathbf{x})$, so for the MPE task it is sufficient to directly optimize:

$$\mathbf{x}' = \arg\max_{\mathbf{x}} P(\mathbf{x}) = \arg\max_{\mathbf{x}} \sum_i f_i(\mathbf{x}). \tag{2}$$

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. Clustering with Comrafs

We will demonstrate the basic principle of unsupervised learning with Comrafs on a classic application of data clustering. First, we define a combinatorial r.v. $\tilde{X}^c$ over a set of all clusterings of a given set. To illustrate this on a small example, let us consider three data points $\{x_1, x_2, x_3\}$ and define a discrete random variable X with an empirical probability distribution over this set. One of a few possible hard

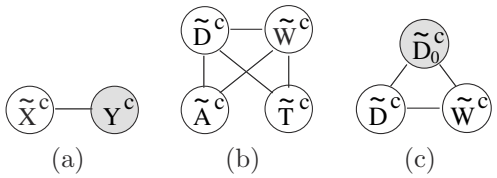

Figure 1. Comraf graphs for: (a) hard version of IB; (b) 4-way MDC; (c) semi-supervised clustering.

clusterings of this set is, say, $\{\{x_1, x_3\}, \{x_2\}\}$. We can define a random variable $\tilde{X}$ over this set of clusters, so that it can take two values: $\tilde{x}_1 = \{x_1, x_3\}$ and $\tilde{x}_2 = \{x_2\}$. There are five possible clusterings of $\{x_i\}$: $\tilde{x}_1^c = \{\{x_1, x_2, x_3\}\}$, $\tilde{x}_2^c = \{\{x_1\}, \{x_2, x_3\}\}$, $\tilde{x}_3^c = \{\{x_1, x_2\}, \{x_3\}\}$, $\tilde{x}_4^c = \{\{x_1, x_3\}, \{x_2\}\}$, and $\tilde{x}_5^c = \{\{x_1\}, \{x_2\}, \{x_3\}\}$. The combinatorial r.v. $\tilde{X}^c$ is then defined over the set $\{\tilde{x}_i^c\}$ of these clusterings: it will take one of the five possible values. Throughout this paper we will be consistent with the notation introduced above: x is a data point, $\tilde{x}$ is a cluster, and $\tilde{x}^c$ is a clustering; X is a r.v. over $\{x_i\}$, $\tilde{X}$ is a r.v. over $\{\tilde{x}_i\}$, $\tilde{X}^c$ is a (combinatorial) r.v. over $\{\tilde{x}_i^c\}$.

We then decide about interactions between combinatorial r.v.'s (possibly, with ordinary r.v.'s), and construct a Comraf graph. To use the objective from Equation (2), 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 complex potential functions (see Section 2.1), we use Mutual Information between r.v.'s defined over values of adjacent nodes: a potential between values of $\tilde{X}_i^c$ and $\tilde{X}_j^c$ is $I(\tilde{X}_i; \tilde{X}_j) = \sum_{\tilde{x}_i, \tilde{x}_j} P(\tilde{x}_i, \tilde{x}_j) \log \frac{P(\tilde{x}_i, \tilde{x}_j)}{P(\tilde{x}_i) P(\tilde{x}_j)}$. Recall that $\tilde{X}_i$ is defined over a clustering $\tilde{x}_i^c$ (a value of $\tilde{X}_i^c$).

This non-trivial choice of potential functions allows us to consider two important special cases of Comrafs:

A hard version of Information Bottleneck (IB) (Tishby et al., 1999) is a special case of a Comraf. In IB, a clustering $(\tilde{x}^c)'$ is constructed that maximizes information about the variable Y (and minimizes information about X), i.e. $(\tilde{x}^c)' = \arg\max_{\tilde{x}^c}(I(\tilde{X}; Y) - \beta I(\tilde{X}; X))$, where β is a Lagrange multiplier. The compression constraint $I(\tilde{X}; X)$ can be omitted if the number of clusters is fixed: $|\tilde{x}^c| = k$. Let us consider graph G in Figure 1(a), where a shaded Y^c represents an *observed* variable.[2] On the only clique in the graph we define one potential which is the Mutual Information $I(\tilde{X}; Y)$. The MPE objective is then defined as:

$$(\tilde{x}^c)' = \arg\max_{\tilde{x}^c} P(\tilde{x}^c, y^c) = \arg\max_{\tilde{x}^c} I(\tilde{X}; Y). \quad (3)$$

Multi-way distributional clustering (MDC) (Bekkerman et al., 2005) is a generalization of IB, where the data has a number of interdependent modalities (such as documents, words, authors, titles, etc.). Bekkerman et al. (2005) represent interactions between the modalities 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 $\tilde{\mathbf{X}}^\mathbf{c} = \{\tilde{X}_1^c, \ldots, \tilde{X}_m^c\}$ that are nodes in a graph G with edges $\mathbf{E}$. The MPE scheme is then:

$$(\tilde{\mathbf{x}}^\mathbf{c})' = \arg\max_{\tilde{\mathbf{x}}^\mathbf{c}} P(\tilde{\mathbf{x}}^\mathbf{c}) = \arg\max_{\tilde{\mathbf{x}}^\mathbf{c}} \sum_{(\tilde{X}_i^c, \tilde{X}_j^c) \in \mathbf{E}} I(\tilde{X}_i; \tilde{X}_j), \quad (4)$$

which is equivalent to the objective proposed by Bekkerman et al. (2005). An example Comraf graph for a 4-way MDC (that corresponds to simultaneously clustering documents, words, authors and titles) is shown in Figure 1(b). MDC demonstrates superior results with respect to two previous state-of-the-art document clustering algorithms, including information-theoretic co-clustering (Dhillon et al., 2003).

Due to unique characteristics of combinatorial r.v.'s, it is problematic to apply existing inference algorithms to Comrafs. Bekkerman et al. (2006) propose a simple and efficient inference algorithm specific for Comrafs, which is based on combinatorial optimization. They show that even such a simple algorithm consistently and significantly outperforms Latent Dirichlet Allocation (Blei et al., 2003) on document clustering.

In the case of a combinatorial MRF with more than one combinatorial r.v., this Comraf inference algorithm becomes a variation of the *Iterative Conditional Mode (ICM)* method (Besag, 1986). ICM optimizes each node of an MRF iteratively (in a round-robin fashion), given its Markov blanket. When an ICM iteration is applied to a node $\tilde{X}_i^c$, the MPE objective from Equation (4) with $O(|\mathbf{X}|^2)$ terms is reduced to:

$$(\tilde{x}_i^c)' = \arg\max_{\tilde{x}_i^c} \sum_{j:\ (\tilde{X}_i^c, \tilde{X}_j^c) \in \mathbf{E}} I(\tilde{X}_i; \tilde{X}_j) \quad (5)$$

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

4. Semi-supervised clustering with Comrafs

The Comraf model is a convenient framework for performing semi-supervised clustering. Prior to present-

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

ing details of a particular Comraf, let us define the concepts of hidden and observed states in Comrafs. A combinatorial r.v. is *hidden* if it can take any value from its support. A combinatorial r.v. is *observed* if its value is preset and fixed.

Semi-supervised clustering is the 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 when each labeled data point $x_i|_{i=1}^n$ belongs to only one category $l_j|_{j=1}^k$. We propose an *intrinsic* Comraf approach (introducing observed nodes to a Comraf graph) for incorporating labeled data into the clustering, and compare it with an existing *constrained optimization* scheme.

Intrinsic scheme. An advantage of Comrafs is that they offer a unique intrinsic method for incorporating labeled data which does not require significant changes in the model. First, note that labels define a natural partition of the labeled data: for each label l_j let $\tilde{x}_j$ be a subset of $\{x_i\}$ labeled with l_j, i.e. $\tilde{x}_j = \{x_i | l_i = l_j\}$. We now define a random variable $\tilde{X}_0$ over the set $\{\tilde{x}_j\}$, and we also define a combinatorial r.v. $\tilde{X}_0^c$ over all the possible partitionings of the set $\{x_i\}$. Since the partition $\tilde{x}_0^c = \{\tilde{x}_j\}$ 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 Figure 1(c). The objective function from Equation (5) and the Comraf inference procedure remain unchanged (with the only difference being 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, while values of the observed nodes are fixed by definition.

Constrained optimization. Previous research works (Wagstaff & Cardie, 2000; Basu et al., 2004) perform semi-supervised clustering with two types of boolean constraints: *must-link* (two data points must be in the same cluster during the course of the clustering algorithm) and *cannot-link* (two data points must not be in the same cluster). Formally, for a cluster $\tilde{x}$ and two data points x_i and x_j labeled by l_i and l_j respectively, a must-link constraint is:

$$ml(x_i, x_j) = \begin{cases} 0, & \text{if } (l_i = l_j) \wedge (x_i \in \tilde{x}) \wedge (x_j \in \tilde{x}) \\ 1, & \text{otherwise,} \end{cases}$$

and a cannot-link constraint is:

$$cl(x_i, x_j) = \begin{cases} 1, & \text{if } (l_i \neq l_j) \wedge (x_i \in \tilde{x}) \wedge (x_j \in \tilde{x}) \\ 0, & \text{otherwise.} \end{cases}$$

Note that in order to fairly compare two semi-supervised methods, for both of them we must use the same underlying clustering algorithm. So, we use the Comraf inference algorithm, where the objective function from Equation (5) is modified to incorporate the constraints. For each combinatorial r.v. $\tilde{X}^c$:

$$\begin{aligned}(\tilde{x}_i^c)' = \arg\max_{\tilde{x}_i^c} & \sum_{j:\ (\tilde{X}_i^c, \tilde{X}_j^c) \in \mathbf{E}} I(\tilde{X}_i; \tilde{X}_j) \\ & - \sum_j w_j\ ml_j - \sum_j w_j\ cl_j,\end{aligned} \tag{6}$$

where w_j are weights that we set at $+\infty$, which corresponds to the requirement that all constraints must be satisfied. Note that in the general case we are free to choose any non-negative weights.

5. Experimental setup

5.1. Datasets

We evaluate the Comraf models on six text datasets: the standard benchmark 20 Newsgroups dataset (20NG) and five real-world email directories. Three of them belong to participants in the CALO project[3] and the other two belong to former Enron employees.[4] For preprocessing steps and statistics on the data, see Bekkerman et al. (2005).

5.2. Evaluation criterion

Following Dhillon et al. (2003), we use *micro-averaged accuracy* for evaluation of our clustering methods. Let $\{x_i\}$ be the data and $\tilde{x}^c$ its clustering. 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 $\tilde{x}$, let $\gamma_T(\tilde{x})$ be the maximal number of $\tilde{x}$'s elements that belong to one category. We say that this category is *dominant* in $\tilde{x}$. Then, the accuracy $Acc(\tilde{x}, T)$ of $\tilde{x}$ with respect to T is defined as $Acc(\tilde{x}, T) = \gamma_T(\tilde{x})/|\tilde{x}|$. The micro-averaged accuracy of the entire clustering $\tilde{x}^c$ is:

$$Acc(\tilde{x}^c, T) = \frac{\sum_{i=1}^k \gamma_T(\tilde{x}_i)}{\sum_{i=1}^k |\tilde{x}_i|}. \tag{7}$$

The main drawback of clustering accuracy is that it does not penalize a split of a category over a number

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

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

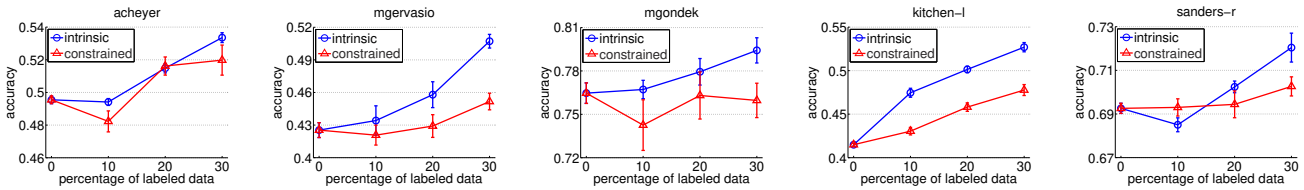

Figure 2. Clustering accuracy of the semi-supervised Comraf on five email datasets.

of clusters, as long as the category remains dominant in these clusters. In this aspect, clustering accuracy differs from (standard) classification accuracy, as defined, e.g., in Bekkerman et al. (2003), so clustering results cannot be directly compared with classification results. However, in the special case described below they can be compared.

Definition 3. *A clustering is called* well-balanced *if each category of the clustered data is dominant exactly in one cluster.*

Claim 1. *For well-balanced clusterings, the micro-averaged clustering accuracy (7) equals the standard micro-averaged classification accuracy.*

The proof of this claim is straightforward (based on the fact that the topic of a cluster is determined according to its dominant category). Claim 1 allows us to compare clustering and classification methods.

6. Semi-supervised clustering results

We report on the clustering accuracy averaged over ten independent runs on the email datasets and five runs on 20NG. We apply agglomerative clustering to documents and divisive clustering to words. Figure 1(c) 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 $\tilde{D}_0^c$, whose value $\tilde{d}_0^c$ is a given partitioning of the labeled documents.

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) versus 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, but the intrinsic method usually outperforms the constrained method.

On 20NG, we select 10% of data to be labeled. The constrained method obtains $74.8 \pm 0.6\%$ accuracy, while the intrinsic method obtains $78.9 \pm 0.8\%$ accuracy (over 5% and 9% absolute improvement to the unsupervised result, which is $69.5 \pm 0.7\%$). We note that having 10% labeled data from 20NG is actually 2,000 labeled documents – a number that should allow constructing a good classifier. We also note that three of the five independent runs of the intrinsic method on 20NG produced *well-balanced* clusterings, whose accuracy ($80.0 \pm 0.6\%$) can be directly compared to classification accuracy (Claim 1). We apply SVM (with linear kernel)[5] to the same three data splits and obtain $77.2 \pm 0.2\%$ accuracy, which is significantly inferior to the semi-supervised intrinsic Comraf results.

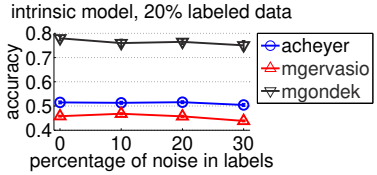

Figure 3. Resistance to noise in intrinsic semi-supervised Comraf.

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 3 shows the results: clustering accuracy remains almost the same for all three datasets.

7. Conclusion and future work

In this paper, we have presented combinatorial MRFs and empirically shown their utility on the problem of semi-supervised clustering of documents. The use of Comrafs is not limited to clustering problems only. We plan to apply Comrafs to ranking and machine translation, along with other important tasks. We also plan to apply Comraf clustering to other domains, such as to image clustering. 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. Learning such models is another interesting area of our future work.

[5] We use Thorsten Joachims' SVM^{light}. For details on our experimental setup, see Bekkerman et al. (2003).

Acknowledgements

Portions of this research were conducted while Ron Bekkerman was an intern at Google. Ron was also supported in part by the Center for Intelligent Information Retrieval and in part by the Defense Advanced Research Projects Agency (DARPA) under contract number HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

Basu, S., Bilenko, M., & Mooney, R. (2004). A probabilistic framework for semi-supervised clustering. *Proceedings of SIGKDD-10* (pp. 59–68).

Bekkerman, R., El-Yaniv, R., & McCallum, A. (2005). Multi-way distributional clustering via pairwise interactions. *Proceedings of ICML-22* (pp. 41–48).

Bekkerman, R., El-Yaniv, R., Tishby, N., & Winter, Y. (2003). Distributional word clusters vs. words for text categorization. *JMLR*, *3*, 1183–1208.

Bekkerman, R., Sahami, M., & Learned-Miller, E. (2006). Combinatorial Markov Random Fields. Under review.

Besag, J. (1974). Spatial interaction and statistical analysis of lattice systems. *Journal of the Royal Statistical Society*, *36*, 192–236.

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society*, *48*.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *JMLR*, *3*, 993–1022.

Dhillon, I. S., Mallela, S., & Modha, D. S. (2003). Information-theoretic co-clustering. *Proceedings of SIGKDD-9* (pp. 89–98).

Freeman, W., Pasztor, E., & Carmichael, O. (2000). Learning low-level vision. *IJCV*, *40*, 25–47.

Friedman, N., Mosenzon, O., Slonim, N., & Tishby, N. (2001). Multivariate information bottleneck. *Proceedings of UAI-17*.

Jordan, M. (2004). Graphical models. *Statistical Science*, *19*, 140–155.

Li, S. (1995). *Markov random field modeling in computer vision*. Springer Verlag.

Metzler, D., & Croft, W. (2005). A markov random field model for term dependencies. *Proceedings of SIGIR-28*.

Sha, F., & Pereira, F. (2003). Shallow parsing with conditional random fields. *Proceedings of HLT-NAACL*.

Tishby, N., Pereira, F., & Bialek, W. (1999). The information bottleneck method. Invited paper to the 37th Annual Allerton Conference.

Wagstaff, K., & Cardie, C. (2000). Clustering with instance-level constraints. *Proceedings of ICML-17*.

Wiegerinck, W. (2000). Variational approximations between mean field theory and the junction tree algorithm. *Proceedings of UAI-16* (pp. 626–633).