Multi-way Distributional Clustering (MDC) and Combinatorial Markov Random Fields (Comrafs)

The goal of the MDC project is to provide a comprehensive, unified solution to the problem of data clustering, both in unsupervised and semi-supervised learning frameworks. Clustering is a fundamental Machine Learning problem of constructing homogeneous groups of data instances, for example, of documents grouped by topic, of images grouped by foreground subject, or of genes grouped by function. There are various off-the-shelf toolkits in which clustering algorithms are implemented (e.g. WEKA), however, none of them has proved itself to be state-of-the-art. Our clustering method consistently and significantly outperforms best existing methods, both information-theoretic (e.g. sequential Information Bottleneck) and generative (Latent Dirichlet Allocation). We believe that MDC is one of the best clustering methods currently available.

MDC is an efficient implementation of the Comraf model. Comraf stands for Combinatorial Markov Random Fields, a novel type of undirected graphical models we have recently proposed. The model is grounded on three principles:

- We exploit multi-modality of the data, i.e. the fact that the data can be viewed from different angles or perspectives. For example, consider a data set of *documents* that should be clustered by topic. A different modality of this data set would be a set of *words* that are contained in the documents. Other modalities would be a set of authors' *names*, a set of documents' *titles* etc. Comraf allows to simultaneously cluster a number of data modalities, while each one potentially improves the quality of all the others.

- Most existing clustering methods are based on explicit definition of pairwise distance measure between data instances. Such a measure is usually chosen heuristically, and can in some cases be inappropriate for particular tasks. We refrain from such an explicit definition and instead optimize a global objective function over the entire data.

- Most existing clustering algorithms are either agglomerative (start with small clusters and merge them to compose larger clusters), divisive (start with one large cluster and split it to obtain smaller clusters), or flat, such as k-means (start with k clusters and rearrange data within these clusters). We propose to combine all the three approaches together, while choosing the most appropriate one for each data modality.

Detailed description of the Comraf model is given in Bekkerman et al. (ECML-2006); detailed description of the underlying clustering algorithm (MDC) is given in Bekkerman et al. (ICML-2005). Some modalities may not be clustered at all, see Bekkerman and Jeon (CVPR-2007) for details.

Code. The MDC toolkit is currently a stable beta, although it has not been tested on a large variety of clustering problems. Any bug report will be greatly appreciated. MDC is a console application, quite simple in use. To perform clustering, the following input is required:

- Pairwise contingency tables of data modalities (labeled data can be straightforwardly incorporated, which allows semi-supervised clustering).

- An initialization file that specifies a particular order of processing the modalities and a choice of clustering approach (agglomerative, divisive or flat) for each of them.

Publications:

R. Bekkerman and J. Jeon. **Multi-modal Clustering for Multimedia Collections.** In *Proceedings of **CVPR 2007***

R. Bekkerman, M.Sahami, and E. Learned-Miller. **Combinatorial Markov Random Fields.** In *Proceedings of **ECML 2006***

R. Bekkerman and M.Sahami. **Semi-supervised Clustering using Combinatorial MRFs.** In *Proceedings of ICML 2006 Workshop on Learning in Structured Output Spaces*

R. Bekkerman, R. El-Yaniv, and A. McCallum. **Multi-way Distributional Clustering via Pairwise Interactions.** In *Proceedings of **ICML 2005***