Learning Clusterwise Similarity with First-Order Features

Aron Culotta and Andrew McCallum
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
{culotta, mccallum}@cs.umass.edu

Abstract

Many clustering problems can be reduced to the task of partitioning a weighted graph into highly-connected components. The weighted edges indicate pairwise similarity between two nodes and can often be estimated from training data. However, in many domains, there exist higher-order dependencies not captured by pairwise metrics. For example, there may exist soft constraints on aggregate features of an entire cluster, such as its size, mean or mode. We propose *clusterwise* similarity metrics to directly measure the cohesion of an entire cluster of points. We describe ways to learn a clusterwise metric from labeled data, using weighted, first-order features over clusters. Extending recent work equating graph partitioning with inference in graphical models, we frame this approach within a discriminatively-trained Markov network. The advantages of our approach are demonstrated on the task of *coreference resolution*.

1 Clusterwise Similarity

The input to a clustering algorithm is often a weighted undirected graph, where the weight between two nodes indicates their similarity. The goal of clustering is to partition the graph into components with heavy intra-cluster edges and light inter-cluster edges. Recently, these weights have been estimated from training data using maximum likelihood [1, 2].

While effective, by factoring the similarity metric into a set of pairwise functions, this approach sacrifices expressivity for tractability. For many domains, there exist *clusterwise* soft constraints that cannot be represented by pairwise functions. Consider the task of *coreference resolution* (also called identity uncertainty or deduplication), in which mentions referring to the same underlying object are clustered together. For example, given a database of research paper citations, we would like to cluster together citations referring to the same paper. Examples of clusterwise constraint include (1) a paper is rarely referenced more than 100 times or (2) an author's name is unlikely to be misspelled 5 different ways.

To represent these clusterwise constraints, we propose using similarity metrics to measure the compatibility of a cluster of nodes, rather than simply pairs of nodes. The challenge lies in constructing efficient methods to estimate these metrics from training samples and to partition the resulting graph.

Given a deduplicated training database, we estimate the clusterwise metric by sampling positive and negative example clusters. We then specify a set of *first-order* predicates as features to describe the compatibility of the nodes in each cluster. Each of these predicates has an associated weight, which is estimated by maximizing the conditional log-likelihood of the training data. We approximate the optimal partitioning of a graph with an agglomerative algorithm that greedily merges clusters based on their predicted compatibility scores.

We apply our technique to deduplicate authors and citations in a publications database and find that the clusterwise metric achieves higher F1 scores than the pairwise metric on 5 of 7 datasets. For more details, we refer the reader to our technical report [3].

2 Related Work

Inference in certain types of undirected graphical models can be reduced to graph partitioning [4]. Our technique is a type of undirected model that is parameterized by cliques over sets of mention variables, for which inference is approximated with a graph partitioning algorithm inspired by *correlation clustering* [5]. Our method can also be viewed as a type of Markov logic network [7] in which first-order predicates are defined over clusters. Thus, it is an extension of recent coreference research that parameterizes an undirected model with mention pairs [1, 6].

Milch et al. [8] present a *generative* model that can represent the clusterwise features we describe here; however, the transition to generatively-trained models sacrifices some of the attractive properties of the discriminative models in McCallum and Wellner [1] and Parag and Domingos [6], such as the ability to easily incorporate many overlapping features of the observed mentions. In contrast, generative models are constrained either to assume the conditional independence of these features or to explicitly model their interactions. Our work is an attempt to incorporate the attractive properties of McCallum and Wellner [1] and Milch et al. [8], resulting in a discriminatively-trained model to reason about objects.

References

[1] A. McCallum and B. Wellner. Toward conditional models of identity uncertainty with application to proper noun coreference. In *IJCAI Workshop on Information Integration on the Web*, 2003.

[2] Mikhail Bilenko and Raymond J. Mooney. Adaptive duplicate detection using learnable string similarity measures. In *9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003)*, pages 39–48, 2003.

[3] Aron Culotta and Andrew McCallum. Practical markov logic containing first-order quantifiers with application to identity uncertainty. Technical Report IR-430, University of Massachusetts, September, 2005.

[4] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *In IEEE transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 23(11):1222–1239, 2001.

[5] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. *Machine Learining*, 56:89–113, 2004.

[6] Parag and Pedro Domingos. Multi-relational record linkage. In *Proceedings of the KDD-2004 Workshop on Multi-Relational Data Mining*, pages 31–48, August 2004.

[7] M. Richardson and P. Domingos. Markov logic networks. Technical report, University of Washington, Seattle, WA, 2004.

[8] Brian Milch, Bhaskara Marthi, and Stuart Russell. BLOG: Probabilistic models with unknown objects. In *IJCAI*, 2005.