Dirichlet Process Mixture Models for Verb Clustering

Andreas Vlachos, Zoubin Ghahramani, Anna Korhonen
University of Cambridge

Introduction

In natural language processing (NLP) many tasks are aimed at discovering novel, previously unknown information in corpora. Bayesian non-parametric models are well-suited for this purpose, since the number of components used to model the data is not fixed in advance but is determined by the model and the data.

In this work, we apply Dirichlet Process Mixture Models (DPMMs) (Neal, 2000) to a typical unsupervised learning NLP task, lexical-semantic verb clustering. The task involves discovering classes of verbs similar in terms of their syntactic-semantic properties (e.g. MOTION class for the verbs "travel", "walk" and "run"). Although some fixed classifications are available (e.g. Levin, 1993), these are not comprehensive and are inadequate for specific domains.

The clustering algorithms applied to this task so far require the number of clusters as input. This is problematic because:

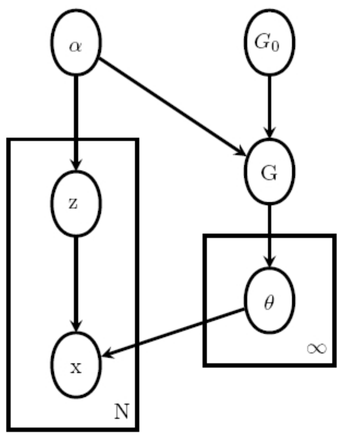
- *We do not know how many classes exist in the data.*
- *Even if we do (e.g. in the context of a carefully controlled experiment), the dataset might not contain instances for all of them.*
- *Each class is not necessarily contained in one cluster exclusively, since the target classes are defined manually without taking into account the feature representation.*

Unsupervised clustering with DPMMs

Dirichlet Process mixture model:

$$\begin{aligned} G|\alpha, G_0 &\sim DP(\alpha, G_0) \\ \theta_{z_i}|G &\sim G \\ x_i|\theta_{z_i} &\sim p(x_i|\theta_{z_i}) \end{aligned} \tag{1}$$

- *G_0 and G are probability distributions over the component parameters (θ), G being randomly drawn with mean G_0.*
- *$\alpha > 0$ is the concentration parameter which determines the variance of the DP.*
- *z_i is the component chosen for instance x_i, and θ_{z_i} its parameters.*
- *The probability of assigning an instance to a particular component is proportionate to the number of instances already assigned to it.*
- *The probability that a new cluster is created is proportionate to α.*



In the experiments reported, the components are modelled after the multinomial distribution, we used Gibbs sampling to infer the assignments z_i and the α parameter was determined using a Gamma prior in a Metropolis sampling scheme.

Adding supervision

In many cases it desirable to influence the solution with respect to some prior intuition or some considerations relevant to the application in mind. For example, in the task of verb clustering, "encompass" and "carry" could be in the same cluster if the aim is to cluster all verbs meaning INCLUSION together, but they could also be separated if the aspect of MOTION of the latter is taken into account.

We modelled human supervision as pairwise constraints over instances, as in Klein et al. (2002): given a pair of instances, either they should be clustered together (must-link) or not (cannot-link). This information can be obtained either from a human expert, or by appropriate manipulation of extant resources, such as ontologies. The expectation is that such constraints will not only affect the participating instances but the overall clustering as well. We incorporated these pairwise links by altering the standard sampling scheme of Neal (2000):.

$$\begin{aligned} P(z_i = z|z_{-i}, x_i) &= 0 \quad \text{if } x_i \text{ has a } \textit{cannot-link} \text{ with an instance in } z \\ P(z_i = z|z_{-i}, x_i) &= 1 \quad \text{if } x_i \text{ has a } \textit{must-link} \text{ with an instance in } z \end{aligned} \tag{2}$$

$$\begin{aligned} P(z_i = z|z_{-i}, x_i) &\sim \frac{n_{-i,z}}{n-1+\alpha} \int_\theta p(x_i|\theta) Dir(\theta|\alpha, G) \\ P(z_i = z'|z_{-i}, x_i) &\sim \frac{\alpha}{n-1+\alpha} \int_\theta p(x_i|\theta) Dir(\theta|\alpha, G_0) \end{aligned}$$

Biomedical verb dataset

In our experiments we used the dataset of Korhonen et al. (2006):

- *Instances: 193 medium to high frequency verbs from a corpus of 2230 full-text articles*
- *Features: 439 subcategorization frames (SCFs) and their associated frequencies automatically extracted from the corpus*
- *Three-level gold standard: 16, 34 and 50 classes*
- *Very sparse feature set, applied non-negative matrix factorization and kept 35 dimensions*

1 Have an effect on activity (BIO/29)	9 Report (GEN/30)
1.1 Activate / Inactivate	**9.1 Investigate**
1.1.1 Change activity: *activate, inhibit*	9.1.1 Examine: *evaluate, analyze*
1.1.2 Suppress: *suppress, repress*	9.1.2 Establish: *test, investigate*
1.1.3 Stimulate: *stimulate*	9.1.3 Confirm: *verify, determine*
1.1.4 Inactivate: *delay, diminish*	**9.2 Suggest**
1.2 Affect	9.2.1 Presentational: *hypothesize, conclude*
1.2.1 Modulate: *stabilize, modulate*	9.2.2 Cognitive: *consider, believe*
1.2.2 Regulate: *control, support*	**9.3 Indicate:** *demonstrate, imply*
1.3 Increase / decrease: *increase, decrease*	
1.4 Modify: *modify, catalyze*	

Evaluation

The evaluation of unsupervised clustering against a gold standard is not straightforward because the clusters found by the algorithms ($K = \{k_j | j = 1, ..., |K|\}$) are not associated with the classes in the gold standard ($C = \{c_l | l = 1, ..., |C|\}$). Frequently used metrics include (Rosenberg and Hirschberg, 2007):

- *F-measure (assumes the missing mapping between c_l and k_j)*
- *Rand Index (adding noisy clusters does not decrease the score)*
- *Variation of information (its value range depends on the maximum number of classes $|C|$ and clusters $|K|$)*

Rosenberg and Hirschberg suggest a new information-theoretic metric for clustering evaluation, V-measure. V-measure is the harmonic mean of homogeneity and completeness which evaluate the quality of the clustering in a complementary way. Homogeneity assesses the degree to which each cluster contains instances from a single class of C. Completeness assesses the degree to which each class is contained in a single cluster.

$$\begin{aligned} h &= 1 - \frac{H(C|K)}{H(C)} \\ c &= 1 - \frac{H(K|C)}{H(K)} \\ V &= \frac{2 * h * c}{h + c} \end{aligned} \tag{3}$$

We should note that V-measure favors clustering solutions with a large number of clusters (large $|K|$), since such solutions can achieve very high homogeneity while maintaining reasonable completeness.

Experiments

1. *Ran the DPMM without any supervision ("vanilla")*
2. *Generated 100 random verb pairs and labelled them as must-link or cannot-link according to the version of the gold standard targeted*
3. *Continued the sampling process using the links generated*
4. *Evaluated against all versions of the gold standard*

	hom	*comp*	*V*
16 classes			
vanilla	*77.09%*	*64.11%*	*70%*
link16_100	*82.16%*	*64.52%*	*72.28%*
link50_100	*77.53%*	*62.69%*	*69.32%*
34 classes			
vanilla	*70.24%*	*78.94%*	*74.34%*
link34_100	*73.19%*	*79.24&*	*76.10%*
50 classes			
vanilla	*69.07%*	*87.43%*	*77.17%*
link16_100	*70.87%*	*84.71%*	*77.17%*
link50_100	*71.19%*	*87.63%*	*78.56%*

Conclusions - Future Work

Contributions made:

- *Applied DPMMs to lexical-sematic verb clustering*
- *Showed how to adapt DPMMs to different needs using supervision in the form of pairwise links between instances*
- *Evaluated results using the newly introduced V-measure*

Future work:

- *Investigate hierarchical Bayesian non-parametric models for verb clustering*
- *Active selection of supervision links instead of random*
- *Extrinsic evaluation of the clustering produced in the context of an NLP application*