Active learning for constrained Dirichlet process mixture models

Andreas Vlachos
Computer Laboratory
University of Cambridge
av308@cl.cam.ac.uk

Zoubin Ghahramani
Department of Engineering
University of Cambridge
zoubin@eng.cam.ac.uk

Ted Briscoe
Computer Laboratory
University of Cambridge
ejb@cl.cam.ac.uk

Abstract

Recent work applied Dirichlet Process Mixture Models to the task of verb clustering, incorporating supervision in the form of *must-links* and *cannot-links* constraints between instances. In this work, we introduce an active learning approach for constraint selection employing uncertainty-based sampling. We achieve substantial improvements over random selection on two datasets.

1 Introduction

Bayesian non-parametric mixture models have the attractive property that the number of components used to model the data is not fixed in advance but is determined by the model and the data. This property is particularly interesting for NLP where many tasks are aimed at discovering novel information. Recent work has applied such models to various tasks with promising results, e.g. Teh (2006) and Cohn et al. (2009).

Vlachos et al. (2009) applied the basic model of this class, the Dirichlet Process Mixture Model (DPMM), to lexical-semantic verb clustering with encouraging results. The task involves discovering classes of verbs similar in terms of their syntactic-semantic properties (e.g. MOTION class for *travel, walk, run*, etc.). Such classes can provide important support for other tasks, such as word sense disambiguation, parsing and semantic role labeling. (Dang, 2004; Swier and Stevenson, 2004) Although some fixed classifications are available these are not comprehensive and are inadequate for specific domains.

Furthermore, Vlachos et al. (2009) used a constrained version of the DPMM in order to guide clustering towards some prior intuition or considerations relevant to the specific task at hand. This supervision was modelled as pairwise constraints between instances and it informs the model of relations between them that cannot be recovered by the model on the basis of the feature representation used. Like other forms of supervision, these constraints require manual annotation and it is important to maximize the benefits obtained from it. Therefore it is natural to consider active learning (Settles, 2009) in order to focus the supervision on clusterings on which the model is uncertain.

In this work, we propose a simple yet effective active learning method employing uncertainty based sampling. The effectiveness of the AL method is demonstrated on two datasets, one of which has multiple gold standards.

2 Constrained DPMMs for clustering

In DPMMs, the parameters of each component are generated by a Dirichlet Process (DP) which can be seen as a distribution over distributions. Each instance, represented by its features, is generated by the component it is assigned to. The components discovered correspond to the clusters. The prior probability of assigning an instance to a particular component is proportionate to the number of instances already assigned to it, in other words, the DPMM exhibits the "rich get richer" property. A popular metaphor to describe the DPMM which exhibits an equivalent clustering property is the Chinese Restaurant Process (CRP). Customers (instances) arrive at a Chinese restaurant which has an infinite number of tables (components). Each customer sits at one of the tables that is either occupied or vacant with popular tables attracting more customers.

Following Navarro et al. (2006), parameter estimation is performed using Gibbs sampling by sampling the assignment z_i of each instance x_i given all the others z_{-i} and the data X:

$$P(z_i = z|z_{-i}, X) \propto$$
$$p(z_i = z|z_{-i})P(x_i|z_i = z, X_{-i}) \quad (1)$$

In Eq. 1 $p(z_i = z|z_{-i})$ is the CRP prior and $P(x_i|z_i = z, X_{-i})$ is the distribution that generates instance x_i given it has been assigned to component z. This sampling scheme is possible because the assignments in the model are exchangeable, i.e. their order is not relevant.

The constrained version of the DPMM uses pairwise constraints over instances in order to adapt the clustering discovered. Following Wagstaff & Cardie (2000), a pair of instances is either linked together (*must-link*) or not (*cannot-link*). For example, *charge* and *run* should form a *must-link* if the aim is to cluster MOTION verbs together, but they should form a *cannot-link* if we are interested in BILL verbs. All links are assumed to be consistent with each other. In order to incorporate the constraints in the DPMM, the Gibbs sampling scheme is modified so that *must-linked* instances are generated by the same component and *cannot-linked* instances always by different ones. Following Vlachos et al. (2009), for each instance that does not belong to a *linked-group*, the sampler is restricted to choose components that do not contain instances *cannot-linked* with it. For instances in a *linked-group*, their assignment is sampled jointly, again taking into account their *cannot-links*. This is performed by adding each instance of the *linked-group* successively to the same component. In terms of the CRP metaphor, customers connected with *must-links* arrive at the restaurant and choose a table jointly, respecting their *cannot-links* with other customers.

3 Active Constraint Selection

In active learning, the model selects the supervision to be provided by a human expert. In the context of the DPMMs, the model chooses a pair of instances for which a *must-link* or a *cannot-link* must be provided. To select the pair, we employ the simple but effective idea of uncertainty based sampling. We consider the most informative link as that on which the model is most uncertain, more formally the link between instances l^*_{ij} that maximizes the following entropy:

$$l^*_{ij} = \arg\max_{i,j} H(z_i = z_j) \qquad (2)$$

If we consider clustering as binary classification of links into *must-links* and *cannot-links*, it is equivalent to selecting the pair with the highest label entropy. During the sampling process used for parameter inference, component assignments vary between samples and the components themselves are not identifiable, i.e. one cannot match the components of one sample with those of another. Furthermore, the conditional assignments estimated during Gibbs sampling (Eq. 1) they do not capture the uncertainty of the assignments z_{-i} on which they condition. Therefore, we resort to generating a set of samples from the (possibly constrained) DPMM and pick the link on which these samples maximally disagree, i.e. we approximate the distribution in Eq. 2 with the probability that instances i, j are in the same cluster or not. Thus, in a given set of samples the most uncertain link would be the one between two instances which are in the same cluster in exactly half of these samples. Using multiple samples allows us to take into account the uncertainty in the assignments of the other instances, as well as the varying number of components.

Compared to standard pool-based AL, when clustering with constraints the possible links between two instances (ignoring transitivity) are $C(N, 2) = N(N-1)/2$ (N is the size of the dataset) and there is an equal number of candidate queries to be considered, as opposed to N queries in a supervised classification task. Another interesting difference is that the the AL process can be initiated without any supervision, since the DPMM is unsupervised. On the other hand, in the standard AL scenario a (usually small) labelled seed set is used. Therefore, we rely exclusively on the model and the features to guide the constraint selection process. If the model combined with the features is not appropriate for the task then the constraints chosen are unlikely to be useful.

4 Datasets and Evaluation

In our experiments we used two verb clustering datasets, one from general English (Sun et al., 2008) and one from the biomedical domain (Korhonen et al., 2006). In both datasets the features for each verb are its subcategorization frames (SCFs) which capture the syntactic context in which it occurs. They were acquired automatically using a domain-independent statistical parsing toolkit, RASP (Briscoe and Carroll, 2002), and a classifier which identifies verbal SCFs. As a consequence, they include some noise due to standard text processing and parsing errors and due to the subtlety of the argument-adjunct distinction. The general English dataset contains 204 verbs

belonging to 17 fine-grained classes in Levin's (Levin, 1993) taxonomy so that each class contains 12 verbs. The biomedical dataset consists of 193 medium to high frequency verbs from a corpus of 2230 full-text articles from 3 biomedical journals. A team of linguists and biologists created a three-level gold standard with 16, 34 and 50 classes. Both datasets were pre-processed using non-negative matrix factorization (Lin, 2007) which decomposes a large sparse matrix into two dense matrices (of lower dimensionality) with non-negative values. In all experiments 35 dimensions were kept. Preliminary experiments with different number of dimensions kept did not affect the performance substantially.

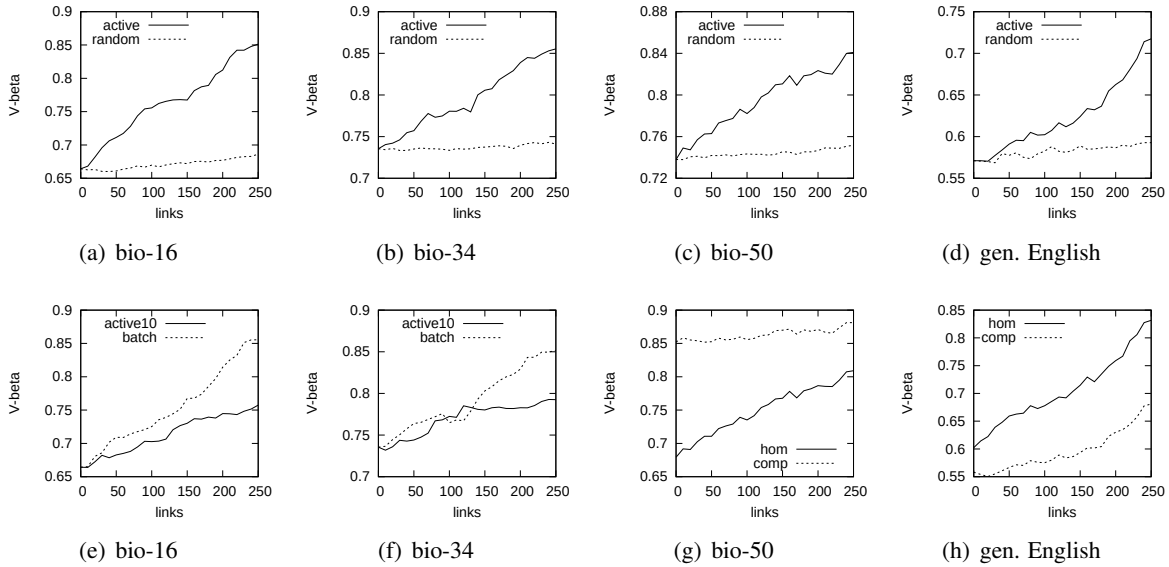
We evaluate our results using three information theoretic measures: Variation of Information (Meilă, 2007), V-measure (Rosenberg and Hirschberg, 2007) and V-beta (Vlachos et al., 2009). All three assess the two desirable properties that a clustering should have with respect to a gold standard, homogeneity and completeness. Homogeneity reflects the degree to which each cluster contains instances from a single class and is defined as the conditional entropy of the class distribution of the gold standard given the clustering. Completeness reflects the degree to which each class is contained in a single cluster and is defined as the conditional entropy of clustering given the class distribution in the gold standard. V-beta balances these properties explicitly by taking into account the ratio of the number of cluster discovered over the number of classes in the gold standard. While an ideal clustering should have both properties, naively improving one of them can be harmful for the other. Compared to the more commonly used F-measure (Fung et al., 2003), these measures have the advantage that they do not assume a mapping between clusters and classes.

5 Experiments

We performed experiments in order to assess the effectiveness of the AL algorithm for the constrained DPMM comparing it to random selection. In each AL round, we run the Gibbs sampler for the (constrained) DPMM five times, using 100 iterations for burn-in, draw 20 samples from each run with 5 iterations lag between samples and select the most uncertain link to be labeled. Following Navarro et al. (2006), the concentration parameter is inferred from the data using Gibbs sampling. The performances were averaged across the collected samples. Random selection was repeated three times. The three levels of the biomedical gold standard were used independently and together with the general English dataset result in four experimental setups.

The comparison between AL and random selection for each dataset is shown in graphs 1(a)-1(d) using V-beta, noting that the observations made hold with all evaluation metrics used. Constraints selected via AL improve the performance rapidly. Indicatively, the performance reached using 1000 randomly chosen constraints is obtained using only 110 actively selected ones in the *bio-50* dataset. AL performance levels out in later stages with performance superior to the one achieved using random selection with the same number of constraints. The poor performance of random selection is expected, since the unsupervised DPMM predicts more than 90% of the binary links correctly. Another interesting observation is that, during AL, homogeneity increased faster than completeness (graphs 1(g) and 1(h)). This suggests that the features used lead the model towards finer-grained clusters, which is further confirmed by the fact that the highest scores on the biomedical dataset are achieved when comparing against the finest-grained version of the gold standard. While it is possible to choose constraints to the model that would increase completeness with respect to the gold standard, we argue that this would not allow us to obtain obtain insights on the model and the features used.

We also noticed that the choice of batch size has a significant effect on the learning rate of the model. This phenomenon occurs in varying degrees in many applications of AL. Manual inspection of the links chosen at each round revealed that batches often contained links involving the same instances. This is expected due to transitivity: if the link between instances A and B is uncertain but the link between instances B and C is certain, then the link between A and C will be uncertain too. While reducing the batch size leads to better learning rates, it requires estimating the model more often. In order to ameliorate this issue, after obtaining the label of the most uncertain link, we remove the samples that disagreed with it and re-calculate the uncertainty of the remaining links given the remaining samples. This is repeated until the intended batch size is reached. Thus, we



(a) bio-16 (b) bio-34 (c) bio-50 (d) gen. English

(e) bio-16 (f) bio-34 (g) bio-50 (h) gen. English

Figure 1: (a)-(d): Constrained DPMM learning curves comparing random selection and AL. (e),(f): Batch selection comparison. (g),(h): Homogeneity and completeness curves during AL.

avoid selecting links involving the same instance, unless their uncertainty was not reduced by the constraints added. A consideration that arises is that by reducing the number of samples used for uncertainty estimation, progressively we are left with fewer samples to rank the remaining links. Each labeled link reduces the number of samples approximately by half since the most uncertain link is likely to be a *must-link* in half the samples and a *cannot-lnk* in the remaining half. As a result, for a batch with size $|B|$ the uncertainty of the last link will be estimated using $|S|/2^{|B|-1}$ samples. A crude solution would be to generate enough samples for the desired batch size. However, obtaining a very large number of samples can be computationally expensive. Therefore, we set a threshold for the minimum number of samples to be used to estimate the link uncertainty and when it is reached, more samples are generated using the constraints selected. In graphs 1(e) and 1(f) we demonstrate the effectiveness of the batch selection method proposed (labeled "batch") compared to naive batch selection (labeled "active10").

6 Discussion and Future Work

We presented an AL method for constrained DPMMs employing uncertainty based sampling. We applied it to two different verb clustering datasets with 4 gold standards in total and obtained very good results compared to random selection. The idea, while explored in the context of verb clustering with the constrained DPMM, is likely to be applicable to other models that can incorporate *must-links* and *cannot-links* in MCMC sampling.

Most literature on AL for NLP considers supervised methods for classification or sequential tagging. However, AL for clustering is a relatively under-explored area. Klein et al. (2002) incorporated actively selected constraints in hierarchical agglomerative clustering. Basu et al. (2006) have applied AL to obtain *must-links* and *cannot-links* however, the clustering framework used requires the number of clusters to be known in advance which restricts counter-intuitively the clustering solutions that are discovered. Moreover, semi-supervised clustering is a form of semi-supervised learning and in this light, our approach is related to the work of Zhu et al. (2003).

With respect to the practical application of the AL method suggested, it is worth noting that in all our experiments the constraints were obtained for the respective gold standard of the dataset at question and consequently they are all consistent with each other. However, this assumption might not hold in case human experts are employed for the same purpose. In order to use such feedback in the framework suggested, it is necessary to filter the constraints provided in order to obtain a consistent subset. To this end, it would be interesting to investigate the potential of using "soft" constraints, i.e. constraints that are provided with relative confidence.

References

Sugato Basu, Mikhail Bilenko, Arindam Banerjee, and Raymond J. Mooney. 2006. Probabilistic semi-supervised clustering with constraints. In O. Chapelle, B. Schoelkopf, and A. Zien, editors, *Semi-Supervised Learning*, pages 73–102. MIT Press.

Ted Briscoe and John Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 1499–1504.

Trevor Cohn, Sharon Goldwater, and Phil Blunsom. 2009. Inducing compact but accurate tree-substitution grammars. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 548–556.

Hoa Trang Dang. 2004. *Investigations into the role of lexical semantics in word sense disambiguation*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, USA.

Benjamin C. M. Fung, Ke Wang, and Martin Ester. 2003. Hierarchical document clustering using frequent itemsets. In *Proceedings of SIAM International Conference on Data Mining*, pages 59–70.

Dan Klein, Sepandar D. Kamvar, and Christopher D. Manning. 2002. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 307–314.

Anna Korhonen, Yuval Krymolowski, and Nigel Collier. 2006. Automatic classification of verbs in biomedical texts. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 345–352.

Beth Levin. 1993. *English Verb Classes and Alternations: a preliminary investigation*. University of Chicago Press, Chicago.

Chih-Jen Lin. 2007. Projected gradient methods for nonnegative matrix factorization. *Neural Compuation*, 19(10):2756–2779.

Marina Meilă. 2007. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895.

Daniel J. Navarro, Thomas L. Griffiths, Mark Steyvers, and Michael D. Lee. 2006. Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, 50(2):101–122, April.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 410–420.

Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Lin Sun, Anna Korhonen, and Yuval Krymolowski. 2008. Verb class discovery from rich syntactic data. In *Proceedings of the 9th International Conference on Intelligent Text Processing and Computational Linguistics*.

Robert S. Swier and Suzanne Stevenson. 2004. Unsupervised semantic role labelling. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 95–102.

Yee Whye Teh. 2006. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 985–992, Sydney, Australia, July.

Andreas Vlachos, Anna Korhonen, and Zoubin Ghahramani. 2009. Unsupervised and Constrained Dirichlet Process Mixture Models for Verb Clustering. In *Proceedings of the EACL workshop on GEometrical Models of Natural Language Semantics*.

Kiri Wagstaff and Claire Cardie. 2000. Clustering with instance-level constraints. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 1103–1110.

Xiaojin Zhu, John Lafferty, and Zoubin Ghahramani. 2003. Combining Active Learning and Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions. In *ICML workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, pages 58–65.