Dirichlet Process Mixture Models for Verb Clustering

Andreas Vlachos AV308@CL.CAM.AC.UK
Computer Laboratory, University of Cambridge, Cambridge, CB3 0FD, UK

Zoubin Ghahramani ZOUBIN@ENG.CAM.AC.UK
Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, UK

Anna Korhonen ALK23@CAM.AC.UK
Computer Laboratory, University of Cambridge, Cambridge, CB3 0FD, UK

Keywords: non-parametric modelling, computational linguistics, clustering, semantics, semi-supervised learning

Abstract

In this work we apply Dirichlet Process Mixture Models to a learning task in natural language processing (NLP): lexical-semantic verb clustering. We assess the performance on a dataset based on Levin's (1993) verb classes using the recently introduced V-measure metric. In, we present a method to add human supervision to the model in order to to influence the solution with respect to some prior knowledge. The quantitative evaluation performed highlights the benefits of the chosen method compared to previously used clustering approaches.

1. Introduction

Bayesian non-parametric models have received a lot of attention in the machine learning community. These models have the attractive property that the number of components used to model the data is not fixed in advance but is actually determined by the model and the data. This property is particularly interesting for natural language processing (NLP) where many tasks are aimed at discovering novel, previously unknown information in corpora.

In this work, we apply the basic models of this class, Dirichlet Process Mixture Models (DPMMs) (Neal, 2000) to a typical unsupervised learning task in NLP: 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"). Such classes can provide important support for other NLP tasks and applications. Although some fixed classifications are available (e.g. Levin (1993)), these are not comprehensive and are inadequate for specific domains such as the biomedical one (Korhonen et al., 2006b).

The clustering algorithms applied to this task so far require the number of clusters as input (Schulte im Walde, 2006; Korhonen et al., 2006b). This is problematic as we do not know how many classes exist in the data. Even if the number of classes in a particular dataset was known (e.g. in the context of a carefully controlled experiment), a particular dataset may not contain instances for all the classes. Moreover, each class is not necessarily contained in one cluster exclusively, since the target classes are defined manually without taking into account the feature representation used. The fact that DPMMs do not require the number of target clusters in advance, renders them particularly promising for the many NLP tasks where clustering is used for learning purposes.

In addition to applying the standard DPMM to verb clustering we also present a method to add human supervision to the model in order to to influence the solution with respect to some prior intuition or some considerations relevant to the application in mind. We achieve this by enforcing pairwise clustering constraints on the solution discovered by the model. We evaluate these methods on two different datasets including general English and biomedical verbs, respectively. Our results compare favourably to earlier results reported with verb clustering and demonstrate the potential of DPMM based models for discovering novel information from natural language data.

Appearing in *ICML Workshop on Prior Knowledge for Text and Language Processing*, Helsinki, Finland, 2008.

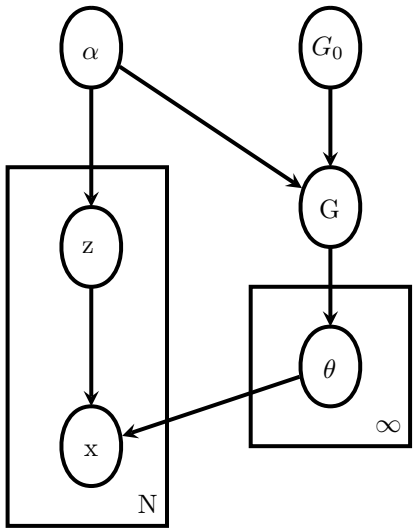

Figure 1. Graphical representation of DPMMs.

2. Unsupervised clustering with DPMMs

With DPMMs, as with other Bayesian non-parametric models, the number of mixture components is not fixed in advance, but is determined by the model and the data. The parameters of each component are generated by a Dirichlet Process (DP) which can be seen as a distribution over the parameters of other distributions. In turn, each instance is generated by the chosen component given the parameters defined in the previous step:

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

In Eq. 1, G_0 and G are probability distributions over the component parameters (θ), and $\alpha > 0$ is the concentration parameter which determines the variance of the Dirichlet process. We can think of G as a randomly drawn probability distribution with mean G_0. Intuitively, the larger α is, the more similar G will be to G_0. z_i is the component chosen for instance x_i, and θ_{z_i} its parameters. The graphical model is depicted in Figure 1.

The prior for assigning instance x_i to either an existing component z or to a new one z_{new} conditioned on the other component assignments (z_{-i}) is given by:

$$\begin{aligned} p(z_i = z|z_{-i}) &= \frac{n_{-i,z}}{N-1+\alpha} \\ p(z_i = z_{new}|z_{-i}) &= \frac{\alpha}{N-1+\alpha} \end{aligned} \tag{2}$$

where $n_{-i,z}$ is the number of instances assigned to component z excluding instance x_i and N is the total number of instances. A clustering of the instances is generated by assigning more than one instance to the same mixture component.

The prior in Eq. 2 exemplifies two main properties of the DPMMs. Firstly, the probability of assigning an instance to a particular component is proportionate to the number of instances already assigned to it ($n_{-i,z}$). In other words, DPMMs exhibit the "rich get richer" property. Secondly, the probability that a new cluster is created depends on the concentration parameter α.

A popular metaphor to describe DPMMs is the Chinese Restaurant Process. Customers (instances) arrive at a Chinese restaurant which has an infinite number of tables (components). Each customer chooses to sit at one of the tables that is either occupied ($p(z_i = z|z_{-i})$) or vacant ($p(z_i = z_{new}|z_{-i})$). Popular tables attract more customers.

An alternative view of DPMMs is the stick-breaking construction (Sethuraman, 1994). In this construction, the mixing proportions of the components (π_k) are produced as follows:

$$\begin{aligned} \pi_k &= \beta_k \prod_{j=1}^{k-1}(1-\beta_j) \\ \beta_k &\sim \mathcal{B}(1, \alpha) \end{aligned} \tag{3}$$

where $\mathcal{B}$ is the Beta distribution. It can be verified that $\sum_{k=1}^{\infty} \pi_k = 1$. Intuitively, the mixing proportion of each component is obtained by successively breaking a stick of unit length. As a result, the mixing proportion of a new component gets progressively smaller. In order to generate an instance x_i, the component z_i is chosen using a multinomial distribution parameterized by the mixing proportions π_k, and the instance is generated as in Eq. 1.

3. Evaluation

The evaluation of unsupervised clustering against a gold standard is not straightforward because the clusters found by the algorithm are not associated with the classes in the gold standard. Formally defined, the method partitions a set of instances $X = \{x_i|i = 1, ..., N\}$ into a set of clusters $K = \{k_j|j = 1, ..., |K|\}$. To evaluate the quality of the resulting clusters, we use an external gold standard in which the instances are partitioned into a set of classes $C = \{c_l|l = 1, ..., |C|\}$. The aim of a clustering algorithm is to find a partitioning of the instances K that resembles as closely as possible the gold standard C.

Most work on verb clustering has used F-measure or the Rand Index (Rand, 1971) for quantitative eval-

uation. However, Rosenberg and Hirschberg (2007) point out that F-measure assumes (the missing) mapping between c_l and k_j. Also, in their experimental assessment they show that when the number of clusters not representing a particular class was increased the Rand Index did not decrease. Another recently introduced metric, variation information (Meilă, 2007), while it avoids these problems, its value range depends on the maximum number of classes $|C|$ and clusters $|K|$ involved in the evaluation, rendering the performance comparisons between different algorithms and/or datasets difficult (Rosenberg & Hirschberg, 2007). 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. This is computed as the conditional entropy of the class distribution of the gold standard given the clustering discovered by the algorithm, $H(C|K)$, normalized by the entropy of the class distribution in the gold standard, $H(C)$. Completeness assesses the degree to which each class is contained in a single cluster. This is computed as the conditional entropy of the cluster distribution discovered by the algorithm given the class, $H(K|C)$, normalized by the entropy of the cluster distribution, $H(K)$. In both cases, we subtract the resulting ratios from 1 to associate higher scores with better solutions:

$$h = 1 - \frac{H(C|K)}{H(C)}$$
$$c = 1 - \frac{H(K|C)}{H(K)}$$
$$V = \frac{2 * h * c}{h + c} \quad (4)$$

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 (Rosenberg & Hirschberg, 2007). To demonstrate this bias for the dataset used in the following section, the clustering solution in which each verb is assigned to a singleton cluster achieves 100% homogeneity, 53.3% completeness and 69.5% V-measure, which are in fact higher than the scores achieved by any of the clustering methods evaluated in the following sections. While increasing $|K|$ does not guarantee an increase in V-measure (splitting homogeneous clusters would reduce completeness without improving homogeneity), it is easier to achieve higher scores when more clusters are produced. The lenience of V-measure towards such solutions reflects the intuition mentioned in the introduction that a single class is likely to be contained in more than one cluster given the representation used. As our method does not require the number of clusters in advance, it is worth keeping this bias in mind.

4. Experiments

Following Kurihara et al. (2007), we used variational inference in order to perform parameter estimation for the DPMMs. In particular, we approximated the infinite vector of the mixing proportions using a finite symmetric Dirichlet prior. The distributions generating the instances of each component were Gaussians with diagonal covariance. The initial number of components was set to 100 and the concentration parameter alpha was set to 1.[1]

After inferring the parameters of the DPMM from the data, for each instance we obtain a probability distribution over the components, in other words a "soft" clustering. In order to produce a clustering solution in which each instance is assigned to one cluster only, each instance is assigned to the component with the highest probability. As a result, the components of the mixture are considered to be the clusters of the clustering solution. However, the transformation described above can result in fewer clusters than components, since there may be components that are not the most probable ones for any instance of the dataset, resulting in empty clusters.

To perform lexical-semantic verb clustering we used the dataset of Sun et al. (2008). It contains 204 verbs belonging to 17 fine-grained classes in Levin's (1993) taxonomy so that each class contains 12 verbs. The classes and their verbs were selected randomly. In Sun et al.'s dataset, the features for each verb are their subcategorization frames (SCFs) and associated frequencies in corpus data, which capture the syntactic context in which the verbs occur in text. SCFs were extracted from the publicly available VALEX lexicon (Korhonen et al., 2006a). VALEX was acquired automatically using a domain-independent statistical parsing toolkit, RASP (Briscoe & Carroll, 2002), and a classifier which identifies verbal SCFs. As a consequence, it includes some noise due to standard text processing and parsing errors and due to the subtlety of argument-adjunct distinction.

As a pre-processing step, we used the logarithms of the frequencies instead of the frequencies themselves, to smooth the very skewed distributions that are typical to natural language. This has a down-scaling effect

[1] http://mi.cs.titech.ac.jp/kurihara/vdpmog.html

on extremely frequent features, without reducing them to the same scale as infrequent ones. Subsequently, the feature vector of each verb was normalized to unit length so that the frequency of the verb does not affect its representation.

Furthermore, dimensionality reduction was applied due to the large number of sparse features. The latter have similar distributions across verbs simply due to their sparsity. Since DPMMs do not weigh the features, a large number of sparse features is likely to influence inappropriately the clustering discovered. Nevertheless, sparse features incorporate useful semantic distinctions and have also performed well in some previous works. Therefore, rather than excluding them, we used principal component analysis (PCA) to reduce the dimensionality, employing the same cut-off point in all our experiments.

We evaluated the performance of the DPMMs in lexical-semantic clustering using the dataset of Sun et al. and experimented with various versions of the VALEX lexicon and the feature sets. In order to alleviate the effect of the random initialization, we ran each experiment 200 times. We achieved the best results with the cleanest version of the lexicon. Our performance was 69.5% homogeneity, 53.7% completeness and 60.5% V-measure, discovering 61.1 clusters on average. The best performance achieved in previous work was 59% in V-measure (Sun et al., 2008) using pairwise clustering (Puzicha et al., 2000). However, this result was achieved by setting the number of clusters to be discovered equal to the number of classes in the dataset, while DPMMs discover the number of clusters in the dataset.

5. Adding supervision

While the ability to discover novel information is attractive in NLP, in many cases it is also desirable to influence the solution with respect to some prior intuition or some considerations relevant to the application in mind. For example, while discovering finer-grained lexical-semantic classes than those included in the gold standard is useful, some NLP applications may benefit from a coarser clustering or a clustering targeted towards revealing some specific aspect of the dataset. 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.

As an extension to this work, we implemented a semi-supervised version of the DPMMs that enables human supervision to guide the clustering solution. The human supervision is modelled 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. Specifying the relations between the instances results in an indirect labeling of the instances. Such labeling is likely to be re-usable, since it defines relations between the datapoints rather than explicit labels. The former are more likely to be useful to multiple tasks which might not have the same labels but could still take advantage of relations between datapoints.

The constraints will be added to the model by taking them into account during parameter estimation. We built a Dirichlet process mixture model using a standard sampling inference scheme (algorithm 3 from Neal (2000)). We chose the multinomial distribution to model the components. Following Neal (2000), we integrated analytically over the parameters θ_{z_i} of the model (Eq. 1 in Section 2).

In order to add supervision to the Dirichlet Process model we sample from distributions that respect the constraints imposed. In more detail, if two instances are connected with a *cannot-link* constraint, we will sample only from distributions that keep them in different components. Therefore, we set to 0 the probability of assigning an instance to a component containing *cannot-link* instance(s). Accordingly, in case they are connected with a *must-link* constraint, we sample only from distributions that keep them in the same component. Therefore, we set to 1 the probability of assigning an instance to a component containing *must-link* instance(s).

The expectation is that such constraints will not only affect the participating instances but the overall clustering as well. By guiding the clustering solution in this manner the DPMMs may discover knowledge better suited to the user's needs.

6. Experiments with supervision

In order to experiment with this method of adding supervision to the DPMMs, we implemented the DPMM model described in the previous section. The α parameter was determined by using a Gamma prior in Metropolis sampling scheme, which was run after each sampling of a component assignment z_i (Eq. 1).

In this second set of experiments we used the dataset of Korhonen et al. (2006b). It consists of 193 medium

to high frequency verbs from a corpus of 2230 full-text articles from 3 biomedical journals. The features, as in the Sun et al. (2008) dataset, were the subcategorization frames (SCFs) and their associated frequencies in the corpus, which were extracted automatically, resulting in 439 preposition-specific SCFs.

A team of domain experts and linguists were involved in creating a gold standard for this dataset. The former analyzed the verbs requiring domain-knowledge and the latter the general English and/or scientific ones. This effort resulted in a three-level gold standard which exemplifies the need for human supervision in order to influence the clustering solution discovered by the DPMMs, since ideally we would like to be able to discover any of these solutions. The number of classes was 16, 34 and 50 at each level of granularity.

As in Section 4, the feature set was very sparse and therefore we applied dimensionality reduction. However, PCA could not be used, since we used the multinomial distribution to model the components which cannot accept negative values. Therefore, we applied non-negative matrix factorization (NMF) (Lin, 2007) which decomposes the dataset in two dense matrices with non-negative values. In order to simulate the process of obtaining human supervision, we generated random verb pairs which we labelled as *must-link* or *cannot-link* according the version of the gold standard we aimed for.

We used 35 dimensions for the NMF dimensionality reduction. The base measure G_0 used was the normalized mean of the dataset, the initial value for the α was 1 and all the instances were assigned to a single component initially. We generated 100 pairs of verbs and obtained their *must-links* or *cannot-links* for each of the three level of granularity of the gold standard. First, we ran the DPMM without any supervision, in order to adapt itself to the data without any constraints for 100 iterations of the Gibbs sampler. Then, we ran the model using the constraints to restrict the sampling for another 100 iterations and obtained the final component assignment.

The results from these experiments appear in Table 1. The rows labeled "vanilla" contain the results for the standard unsupervised model. The other rows are labelled according to the version of the gold standard followed by the number of links obtained from it. The number of clusters discovered by all the versions of the model did not vary substantially, being between 37 and 41. It can be observed that adding supervision to the model guides it to clustering closer to the version of the gold standard the supervision was obtained from. For example, adding 100 links from the coarsest version which contains 16 classes (row "link16_100" in the "16 classes" part of the table) improves the V-measure by 2.28% when evaluating on the same version. However, when evaluating on the finest grained version of the standard (containing 50 classes), then the V-measure remains identical and only the homogeneity and completeness scores change. On the other hand, adding 100 links from the latter version of the gold standard, improves the performance by 1.39% when evaluating on it. As expected though, the performance drops for the 16-class version, since the supervision guides the clustering to a different solution. Adding supervision from the 34-class version, improves the performance by 1.76% in v-measure (row "link34_100"). Overall, the model is adapted towards the clustering solution aimed for. In the rows labeled "gauss" we report the result with the DPMM using Gaussians used in the experiments of Section 4, which discovered 63.23 clusters on average. The new model outperforms it at all level of gold standard, even without using supervision.

	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%
gauss	78.54%	50.22%	61.26%
34 classes			
vanilla	70.24%	78.94%	74.34%
link34_100	73.19%	79.24&	76.10%
gauss	77.30%	66.79%	71.65%
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%
gauss	76.53%	74.49%	75.49%

Table 1. Results on the biomedical verb dataset.

It must be noted that the different levels of granularity could have been achieved by appropriate tuning of the concentration parameter α (Eq. 1). However, to a non-expert in non-parametric modelling we believe it could be easier to simply provide examples of verbs that he or she would consider appropriate to be clustered together or separately. Moreover, α would affect the granularity of the clustering globally, while in a given application one might prefer to influence it more locally, something that can be achieved with the inclusion of pairwise links.

7. Conclusions - Future work

This paper makes several contributions. We applied DPMMs to a typical NLP learning task (lexical-semantic verb clustering) where the ability to discover the number of classes from the data is highly attractive. We experimented with two different datasets including (i) general English and (ii) biomedical verbs. Our quantitative evaluation using the recently introduced V-measure shows that the method compares favorably to earlier verb clustering methods which all rely on a pre-defined number of target clusters. In addition, we demonstrated how such models can be adapted to different needs using supervision in the form of pairwise links between instances.

The results encourage to apply DPMMs to further datasets and tasks. For verb clustering, we plan to investigate hierarchical Bayesian non-parametric models (Heller & Ghahramani, 2005) and to extend our experiments to larger datasets. We plan to conduct a thorough investigation of the ability of DPMMs to discover novel information not included in gold standards. Our preliminary assessment showed that many "errors" are due to the DPMM identifying verbs which are in fact too polysemous to be classified in single classes in large un-disambiguated input data and discovering semantically related classes as well as sub-classes of existing fine-grained classes. With respect to adding supervision to the model, we intend to explore ways in which the DPMM would select the links between instances to be labelled as in Klein et al. (2002), instead of obtaining them at random. Finally, an extrinsic evaluation of the clustering provided by DPMMs as part of an NLP application is likely to be very informative on their practical potential.

References

Briscoe, T., & Carroll, J. (2002). Robust accurate statistical annotation of general text. *Proceedings of the 3rd International Conference on Language Resources and Evaluation*.

Heller, K. A., & Ghahramani, Z. (2005). Bayesian hierarchical clustering. *Proceedings of the 22nd International Conference on Machine Learning*.

Klein, D., Kamvar, S., & Manning, C. (2002). From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. *Proceedings of the 19th International Conference on Machine Learning*.

Korhonen, A., Krymolowski, Y., & Briscoe, T. (2006a). A large subcategorization lexicon for natural language processing applications. *Proceedings of the 5th International Conference on Language Resources and Evaluation*.

Korhonen, A., Krymolowski, Y., & Collier, N. (2006b). Automatic classification of verbs in biomedical texts. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*.

Kurihara, K., Welling, M., & Teh, Y. W. (2007). Collapsed variational Dirichlet process mixture models. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*.

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

Lin, C.-J. (2007). Projected gradient methods for non-negative matrix factorization. *Neural Computation*, *19*, 2756–2779.

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

Neal, R. M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, *9*, 249–265.

Puzicha, J., Hofmann, T., & Buhmann, J. (2000). A theory of proximity based clustering: Structure detection by optimization. *Pattern Recognition*, *33*, 617–634.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*, 846–850.

Rosenberg, A., & Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Schulte im Walde, S. (2006). Experiments on the automatic induction of german semantic verb classes. *Computational Linguistics*, *32*, 159–194.

Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica Sinica*, *4*, 639–650.

Sun, L., Korhonen, A., & Krymolowski, Y. (2008). Verb class discovery from rich syntactic data. *Proceedings of the 9th International Conference on Intelligent Text Processing and Computational Linguistics*.