Application of Lexical Topic Models to Protein Interaction Sentence Prediction

Tamara Polajnar*
Department of Computing Science
University of Glasgow
Glasgow, G12 8QQ, Scotland
tamara@dcs.gla.ac.uk

Mark Girolami
Department of Computing Science
University of Glasgow
Glasgow, G12 8QQ, Scotland
girolami@dcs.gla.ac.uk

Abstract

Topic models can be used to improve classification of protein-protein interactions (PPIs) by condensing lexical knowledge available in unannotated biomedical text into a semantically-informed kernel smoothing matrix. Detection of sentences that describe PPIs is difficult due to lack of annotated data. Furthermore, sentences generally contain a small percentage of the features, thus leading to sparse training vectors. By exploiting contextual similarity of words we are able to improve the classification performance. This contextual data is gathered from a large unannotated corpus and incorporated through a semantic kernel. We use Hyperspace Analogue to Language (HAL) and Bound Encoding of the Aggregate Language Environment (BEAGLE) semantic models to create the kernels. The modularity of the method lends itself to further exploration along several different avenues including experimentation with any number of word and topic models.

1 Introduction

Topic models such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Latent Semantic Analysis (LSA) (Landauer et al., 1998) have been used on a variety of text-based linguistic tasks (Blei et al., 2006; Papadimitriou et al., 2000; Zheng et al., 2006) (as well as for other applications (Yuan et al., 2005)). Here we examine two related models, Hyperspace Analogue to Language (HAL) (Lund and Burgess, 1996) and Bound Encoding of the Aggregate Language Environment (BEAGLE) (Jones and Mewhort, 2007). Whereas, in LDA and LSA words are generally grouped based on their co-occurrence in similar documents, in HAL and BEAGLE words are grouped based on their co-occurrence with other words. Like LSA, HAL and BEAGLE have been evaluated on a variety of psycho-linguistic tasks such as TOEFL word synonym examinations and semantic priming (Jones et al., 2006; Jones and Mewhort, 2007; Landauer et al., 1998; Lund and Burgess, 1996). Topic models such as these provide semantic knowledge that is lost through bag-of-words representation of documents. This property allows us to enrich classification kernels for data-poor applications such as PPI sentence classification. Our approach is similar to the semantic smoothing of kernels using Word Net or Wikipedia information (Minier et al., 2007). However, manually constructed ontological lexical information, such as this, is not available for biomedical words. We also gain inspiration from LDA and LSA semantic kernels which often used to smooth kernels based on word-document co-occurrence in the training data (Aseervatham, 2008; Cristianini et al., 2002).

Protein-protein interaction extraction is a key application of text mining to biological texts. This area of research is strongly motivated by the needs of biologists investigating sub-cellular functions of organisms. PPIs reported in biomedical journals are detected by large-scale biomedical experiments. The substantial number of results produced by these experiments, combined with the ease of access

*Inference Group *http://dcs.gla.ac.uk/inference*

to the digitised publications provided by the various publisher portals, has increased the number of results made available each day. Our goal is to automatically identify sentences that describe PPIs in biomedical texts as a way of aiding curators and researchers.

The most accurate approach is to model the user needs from samples annotated for relevance. We use the AImed dataset (Bunescu et al., 2005) that contains 1,964 sentences, 614 of which contain PPIs. This corpus yields approximately 3000 unique bag-of-word features. While this is a quite small data set for text classification, it is one of the largest high-quality corpora labelled for PPIs.

2 Method

In order to address the lack of data we propose a novel way of combining labelled and unlabelled data (semi-supervised learning), by integrating semantic information from unsupervised lexical semantic models trained on a larger, unannotated GENIA (Kim et al., 2003) corpus.

Firstly we use a semantic model to collect word co-occurrence information. We compare two such models: HAL (Lund and Burgess, 1996) and BEAGLE (Jones and Mewhort, 2007). Next, by applying the cosine and Gaussian kernels, which are often used for text classification, we get viable kernel matrices ($\mathbf{S}$) that can be integrated into the kernel classification methods. We further employ an LDA model of the HAL matrices in order to study the usage of biomedical words.

The AImed data ($\mathbf{X}$) is used for Gaussian process (GP) classifier training and testing, while the semantic information ($\mathbf{S}$) is integrated directly into the kernel $\mathbf{K} = (\mathbf{X} + \epsilon)\mathbf{S}(\mathbf{X} + \epsilon)^T$. The GP was previously shown to have a significantly higher AUC and F-score than the SVM and Naive Bayes on the AImed data (Polajnar et al., 2009). The GPs have an additional advantage of not having extra parameters, such as the SVM margin parameter. This reduces the search space when tuning the kernel hyperparameters for optimal performance.

The above approach has the effect of re-introducing the semantic information about the words that was lost in the bag-of-words representation used to encode the features. A small number $\epsilon < 1$ is added to the training data to allow semantic smoothing across the whole feature set.

3 Semantic models

The Hyperspace Analogue to Language (HAL) matrix, $\mathbf{H}$, is constructed by passing a window of fixed length, L, (context) across the corpus. The first word after the window is considered the target, or the word whose context we are gathering.The strength of the co-occurrence between a target and the context words depends on the distance between the two words, l, $1 < l < L$, within the window. The matrix produced $\mathbf{M}$ by this method is asymmetric, and records the context before the target in the columns and the context after the target in the rows. We use $\mathbf{H} = \mathbf{M} + \mathbf{M^T}$ in order to encode the whole context window of length $2L$.

Bound Encoding of the Aggregate Language Environment (BEAGLE) was proposed as a combined semantic space that incorporates word co-occurrence and word order. For the purpose of comparison with HAL, we only consider the word co-occurrence construction. BEAGLE differs from HAL in that it does not use the raw word counts directly. Instead, it represents each target t with a $1 \times D$ signal vector, $\mathbf{e}_t$, of points drawn from the Gaussian distribution $\mathcal{N}(0, \frac{1}{D}^2)$. This is a random indexing scheme that has been previously compared to principal component analysis (PCA) and LSA in the context of dimensionality reduction (Kaski, 1998; Papadimitriou et al., 2000). The suggested values for D are multiples of 1024 (Jones and Mewhort, 2007). The context in BEAGLE consists of the words occurring in the same sentence as the target word. The target vectors in the BEAGLE co-occurrence matrix, $\mathbf{B}$, are sums of the environmental vectors of the context words. The more times that a certain word is found in the same sentence as the target, the stronger its signal will be within the vector $\mathbf{B}_t$.

4 Results and discussion

By applying HAL and BEAGLE semantic smoothing we are able to improve the area under the receiver operator curve (AUC) and the predictive likelihood (PL) of the models. The best results for

HAL (AUC=90.99 $\pm$ 0.21, PL=0.2759 $\pm$ 0.0061) and for the BEAGLE (AUC=88.59 $\pm$ 0.23, PL = 0.2034 $\pm$ 0.0050) are compared to the best results for this dataset without semantic smoothing (AUC=89.41 $\pm$ 0.24 and PL=0.1934 $\pm$ 0.0040). They show that using semantic kernels generated from a larger external dataset can improve PPI sentence classification AUC by statistically significant amount of over 2%.

References

Aseervatham, S. (2008). A local latent semantic analysis-based kernel for document similarities. *Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, pages 214–219.

Blei, D. M., Franks, K., Jordan, M. I., and Mian, I. S. (2006). Statistical modeling of biomedical corpora: mining the caenorhabditis genetic center bibliography for genes related to life span. *BMC Bioinformatics*, 7:250–250.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Bunescu, R., Ge, R., Kate, R. J., Marcotte, E. M., Mooney, R. J., Ramani, A. K., and Wong, Y. W. (2005). Comparative experiments on learning information extractors for proteins and their interactions. *Artif Intell Med*, 33(2):139–155.

Cristianini, N., Shawe-Taylor, J., and Lodhi, H. (2002). Latent semantic kernels. *J. Intelligent Information Systems*, 18(2-3):127–152.

Jones, M. N., Kintsch, W., and Mewhort, D. J. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, 55(4):534–552.

Jones, M. N. and Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114:1–37.

Kaski, S. (1998). Dimensionality reduction by random mapping: fast similarity computation for clustering. In *Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on*, volume 1, pages 413–418 vol.1.

Kim, J. D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). GENIA corpus–semantically annotated corpus for bio-textmining. *Bioinformatics*, 19 Suppl 1:180–182.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers*, 28(2):203–208.

Minier, Z., Bodo, Z., and Csato, L. (2007). Wikipedia-based kernels for text categorization. In *SYNASC '07: Proceedings of the Ninth International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, pages 157–164, Washington, DC, USA. IEEE Computer Society.

Papadimitriou, C. H., Raghavan, P., Tamaki, H., and Vempala, S. (2000). Latent semantic indexing: a probabilistic analysis. *Journal of Computer and System Sciences*, 61(2):217–235.

Polajnar, T., Rogers, S., and Girolami, M. (2009). Classication of protein interaction sentences via Gaussian processes. In *Proceedings of 4th IAPR International Conference, Pattern Recognition in Bioinformatics*, pages 282–292. Springer Verlag.

Yuan, Y., Lin, L., Dong, Q., Wang, X., and Li, M. (2005). A protein classification method based on latent semantic analysis. *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the*, pages 7738–7741.

Zheng, B., McLean, D. C., and Lu, X. (2006). Identifying biological concepts from a protein-related corpus with a probabilistic topic model. *BMC Bioinformatics*, 7:58–58.