Topic Modeling and Toolkits in Bioinformatics

Barman, Jhinuk
Univerisity of Illinois Urbana-Champaign
California, United States
jbarman2@illinois.edu

***Abstract*—In the branch of Natural Language Processing there are a wide range of applications. The method of topic modeling is an important machine learning method that can applied to biological datasets. This method allows biological datasets to be interpreted more easily and accurately. I will explore the applications of this method and the toolkits involved.**

Keywords—bioinformatics, classification, clustering, topic model,

I. Introduction

There is a new application of text retrieval, specifically topic modeling in the field of bioinformatics. One such application is the analogy between the pairs word-document and gene sample in studies on expression microarray data. This application is the basis of numerous bioinformatics techniques that creates an analogy between document-topic-word and a biological object. There are three main concepts in the tasks for topic modeling on biological data: clustering analysis, data classification and data feature extraction. In order to appropriately use these techniques on vast amounts of biological data, there are four main toolkits that come in handy: (1) Genism (2) Stanford topic modeling toolbox (TMT) (3) MALLET (4) Other open-source software. By comparing and contrasting these bioinformatic topic modeling toolkits, the functions of each toolkit can be examined in more depth.

II. Biological Topic Modeling concepts

A. Clustering Analysis

Topic models can perform clustering and classification of biological data. Unlike traditional clustering methods, a topic model allows data to come from a mixture of clusters instead of a single cluster. The Bag of Words (BoW) method is used to identify biological "clusters" on unlabeled data. The BoW is a word-document matrix. Table 1 shows four words (gene, protein, pathway and microarray) and six documents in the corpus. The value w_{ij} in the matrix refers to the frequency of the word i in document j. The BoW method is a "simplified representation of a corpus as the input of topic modelin" (Liu, Tang, Dong 2016). There are two main assumptions we can make in BoW, Probablistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA). The order of words does not affect representation which means the words can be arranged in any possible way. Additionally, the documents in a corpus are independent.

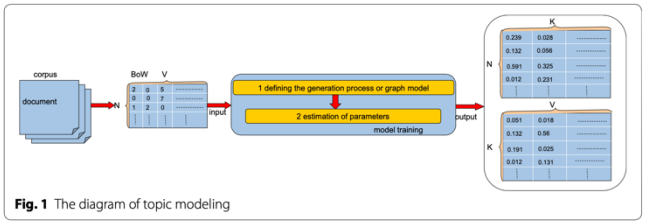

Fig. 1 The diagram of topic modeling

Table 1 An example of a BoW

	d_1	d_2	d_3	d_4	d_5	d_6
Gene	2	0	3	0	0	0
Protein	0	5	0	0	0	0
Pathway	1	2	0	0	0	0
Microarray	0	0	3	6	0	0

Fig. 1. The Diagram of Topic Modelling and Table of BoW

In addition to the BoW method, the LDA model which is completely supervised is a popular method used for clustering documents. To further examine the application of topic modeling to bioinformatics, the studies of researchers are significant. The researcher considers "genomic sequences to be documents, small fragments of a DNA string of size k to be words and the topics discovered by LDA are assigned taxonomic labels" (Liu, Tang, Dong 2016). This study, similar to that of another researcher, shows that genomic sequences correspond to a taxonomic label and the topic found in genomic sequence data has a probability distribution over words. The main topic models used in biological data are PLSA and LDA, both of which need improvements when used in biological contexts. There are many scenarios that violate the assumptions of these topic models, for instance, protein-protein interaction. One potential improvement is the "real-time" topic model (Yao, Hoffmen 2009). This model merges classic inference algorithms that are complex and accurate with new advanced algorithms.

B. Classification

While clustering is a key concept of biological data topic modeling, classification and is an important factor too. "The topic model classifies discoveries of biological 'topics' from a BoW of biological data" (Liu, Tang, Dong 2016). Compared to classification approaches such as support vector machines (SVM), "the classification result of a topic model under certain conditions shows competitive performance" (Rubin 2011). One example of these is the expression of microarray data that illustrates the analogy between word-document and

gene-sample. The Dirichlet allocation extends the LDA model to perform classification tasks. This model can be a foundation for other studies as well such as drug-pathway-gene relations and patient-related texts constructed from clinical and multidimensional genomic measurements.

C. Feature Extraction

Alongside classification, topic modeling is a key component of biological data feature extraction. Algorithms such as a classifier can generate the topic features after reducing dimensionality with methods such as Principal Component Analysis. One study of magnetic resonance imaging (MRI) "applied the PLSA model...they extracted a generative learning score from the learned model, which was used as an input of SVM for the classification task" (Castellani 2010). In this MRI study, the image represented a document, the image shapes were the words and the patterns of the brain surface were the topics. Another study for protein sequence data used a hierarchical latent Dirichlet allocation-random forest (LDA-RF) model that predicted human protein to protein interactions (Pan 2010). The feature values were detected by an LDA model and the probability of the protein-to-protein interaction was predicted by the random forest model. The K-means algorithm and K-nearest neighbor rules are also commonly used methods for feature extraction in bioinformatics.

III. Main Toolkits

Topic models are a common method in natural language processing and bioinformatics. As the models become more widespread, more toolkits are being developed to apply topic models to biology. The four main toolkits will be compared and contrasted to create full, detailed dive into the aspects of each of the toolkits.

A. Genism

Genism is a free Python library that guides the automatic extraction of semantic topics from documents. The input of Genism is a corpus of text documents. Genism has various algorithms such as LDA, LSI, and Random Projects to find the topics of documents. The text documents are compared against other documents for topic similarity.

- A key component of Genism is digital document indexing and similarity search.
- Another feature is the fast, memory-efficient, large-scale algorithms for Singular Value Decomposition and LDA. This consists of unsupervised, semantic analysis of plain text.
- Compared to other toolkits, Genism is easy to use in Python, performs a variety of functions and generates topic modeling visualizations.

B. Stanford Topic Modeling Toolbox (TMT)

Stanford TMT is created by the Stanford NLP group and written in Scala language. TMT is intended for social scientists and anyone who wants to perform analysis on datasets that have a substantial textual component. TMT has the ability to manipulate texts, train topic models, and generate outputs for tracking word usage across topics.

- One barrier to adopting enriched text modeling techniques is that existing toolkits require the user to have prior knowledge on basic text processing methods in order to convert documents into words. TMT tackles this issue by containing implementations of powerful NLP techniques such as LDA and Labeled LDA.
- Many toolkits avoid environments that most users are already familiar with such as spreadsheets and statistically programming languages like R. TMT combats this problem by reading the input text and generative the output as a comma-separated value (CSV) file. This CSV file can be readily analyzed and plotted for data analysis.
- Other features TMT includes are text manipulation pipelines, tokenization, filtering rare or common terms, removing words with too few words. Users can interact with the toolkit by using Scala scripts and can edit the examples provided by the software to cater their needs.
- One of the drawbacks to TMT is the requirement of writing in the Scala language to edit the models provides by the software. A wider range of users and audiences could use this technique if the language was expanded to Python, R and other common languages that more research scientist and data scientist are adept in.

C. MALLET

MALLET is a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction and other machine learning text applications.

- Tools for document classification include Naïve Bayes, Maximum Entropy and Decision Tree algorithms.
- There are various metrics used to evaluate classifier performance.
- There are tools for sequence-tagging such as Hidden Markov Models, Maximum Entropy Markov Models, and Conditional Random Fields.
- The topic modeling toolkit includes sample-based implementations of LDA, Pachinko Allocation and Hierarchical LDA.
- MALLET can also transform text documents into numerical representations using tokenizers, removing stopwords and converting sequences into count vectors.
- Compared to the other toolkits, MALLET has the ability to cover more topics such as classification, sequence-tagging, and transforming text documents in addition to topic modeling.

D. Open Source Software

There is a plethora of open-source software packages that are free and available to the public. For example, at Columbia

University, David Blei's Lab features many of these topic modeling packages on Github. There is a C implementation of Expectation Maximization for LDA, Bayesian modeling for LDA in Python, C++ implementations and more. BLAST is a well-known software library for sequence analysis consisting of parsers and data objects. BioPerl is a library of modules written in in the Perl language to connect bioinformatic applications and datasets for accelerated development of the application. The BioPython package functions similarly to BioPerl in the Python language. BioSQL is a project that presents a schema to represent biological sequence objects and features. In contrast to the previously mentioned bioinformatics toolkits, these open-source tools are simpler to integrate into a user's applications and pipelines (Stajich, Lapp 2006).

IV. Conclusion

Topic Modeling is heavily applicable to bioinformatics through the transformation of word-document methods to biological objects. The main concepts of clustering, classification and feature extraction are crucial to topic modeling and are significantly used on biological data. After conducting a thorough of analysis of the various topic modeling toolkits used in bioinformatics applications, we can see the efficiency, ease and growth of these software as biological data continues to rapidly grow. Toolkits in topic modeling span from Python, Perl, and C implementations that are targeted towards researchers, any user and open-source software that is free and accessible.

A. Authors and Affiliations

Jhinuk Barman

Acknowledgment

University of Illinois Urbana-Champaign, Graduate Data Science Department, *Text Information Systems Course*

References

[1] Liu, L., Tang, L., Dong, W., Yao, S., & Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, *5*(1), 1608. https://doi.org/10.1186/s40064-016-3252-8

[2] Ramage, D., Rosen, E., Chuang, J., Manning, C. D., & McFarland, D. A. (2009, December). Topic modeling for the social sciences. In *NIPS 2009 workshop on applications for topic models: text and beyond* (Vol. 5, pp. 1-4).

[3] Řehůřek, R., & Sojka, P. (2011). Gensim—statistical semantics in python. *Retrieved from genism. org*.

[4] Rogers S, Girolami M, Campbell C, Breitling R (2005) The latent process decomposition of cDNA microarray data sets. IEEE/ACM Trans Comput Biol Bioinf 2(2):143–156

[5] Rubin TN, Chambers A, Smyth P, Steyvers M (2011) Statistical topic models for multi-label document classification. Mach Learn 88(1–2):157–208.

[6] Stajich, J. E., & Lapp, H. (2006). Open source tools and toolkits for bioinformatics: significance, and where are we?. *Briefings in bioinformatics*, *7*(3), 287-296.