Topic Modeling in Bioinformatics

Bidisha Paul

Introduction

Topic modeling has been widely used in the field of computer science. However, in the recent years it has gained a lot attention by researchers especially in the field of genetics and bioinformatics. Topic model is a probabilistic generative model for building abstract topics occurring in a collection of documents. A document is said to be a mixture of topics. Topic models find out the themes and annotate documents according to those themes. Latent semantic indexing (LSI) is the basis for topic model development [1]. Various methods of topic modeling have been developed but latent Dirichlet Analyses is most popular as it very adaptive. However, it is not suitable for complex data relationships. In recent years we have observed an ever expanding growth of biological data. Exploring relationship among these data is vital for biomedical data analyses and other healthcare applications. Researchers are trying to utilize topic models for large datasets such as microarray data [2]. Topic Model can help to discover hidden topic that will be useful to assess the biological meaning of the large document.

Topic Modeling in Bioinformatics

The key steps of topic modeling include bag of words (BoW), model training, and model output. BoW is represented by a word-document matrix. In the field of biology, the matrix can be used to represent gene sequences [3]. For instance, genomic sequences can be considered as documents and DNA fragments of size k to be words. This matrix can be an input for next step of topic modeling. The order of words does not matter and there is no relation among the documents. These are the basic assumptions of topic model and PLSA and LDA use these assumptions. The next step is the model training, which use unsupervised algorithms and the topics are discovered during model training. For example, Table 1 illustrates how the different themes or 'topics' are discovered based on the words that are encountered in the biological documents. The probabilities are sorted in descending order of occurrence. Then a word distribution is created over each topic. Topic model are generally used for three main tasks in the field of biology- biological data classification, data extraction and biological data clustering analyses. Several studies have used latent process decomposition (LDA) for group structure across samples and genes [4]. A study on protein interaction data utilized infinite topic model to find functional gene modules combined with gene expression data. This topic model highlighted relation among documents and clustering of relational data [5].

Biological Research using Topic Modeling

Topic modeling using LDA has been used to assess genomic datasets and annotation of protein functions[6]. In case of microarray analyses, a microarray is considered to be a query and search result is a set of most similar microarray. For the BoW, count of differentially expressed genes in gene sets is equivalent to count of words. Query is encoded as vector consisting of differentially expressed genes [7]. Each experiment represented a document with various topics and each topic corresponded to a distribution over the gene sets. To recognize patterns in images, a study by Coelho used a large collection of fluorescent images and used LDA to identify subcellular localization patterns. An image is represented by mixture of multiple fundamental patterns or topics [8]. For gene sequencing studies, DNA sequences are represented by N-mer frequencies. A study by Chen et al used genome sequences as documents and N-mers as words and LDA model was used to find statistical patterns [9]. Another study by the same author, used LDA model with background distribution (LDA-B) to study microbial abundances to identify the taxa of microorganisms [10]. Topic model has also been used by researchers to study functional microRNA regulatory modules using latent Dirichlet allocation (Corr-LDA). Corr-LDA has been used to interpret images by caption words [11].

Topic models are also useful for classification of labelled data by matching topics to true biological labels. Studies by Perina et al. performed classification task using biologically aware latent Dirichlet allocation (BaLDA) [12]. Another study on classification of gene expression data focused on assessing drug-pathway-gene relations by drawing an analogy between drug-pathway- gene and document topic word. Genes were considered as words and a pathway as topic. The model was useful in predicting responsiveness of the pathway to new drug treatment [13]. PLSA models has been used in Magnetic resonance imaging to implement classification of normal vs schizophrenic patients [14]. To achieve this, images were regarded as documents and shape descriptor of images as visual words whereas geometric patterns of brain surface were visual topics. More recently, latent Dirichlet allocation has been used to evaluate individual genome similarity, which is critical to the study of ethnic groups and cause of genetic disorders [15]. Individuals have been described as mixtures over putative ancestries and each putative ancestry as a distribution over variants.

Topic model to analyze microbial community profile

Microbiome studies are extremely popular now as they play an important role in the study of almost every disease. Even very sterile organs such as liver and pancreas are now believed to have a microbiome of its known and they have been implicated to regulate diseases of the gut and even have indirect role in more lethal diseases such as cancer [16, 17]. Therefore, understanding large amounts of sequencing data and deciphering its relationship to diseases is pivotal. Generally, 16S rRNA sequencing is performed which generates OTU profiles. These OTU's are used to classify or cluster multiple samples. MetaTopics is an R package that has been developed to have a deep understanding of microbial communities and its role in a particular disease [18]. Additionally, it highlights the interactions between different microbial communities. MetaTopics is based off LDA model and Correlated topic model. A microbe sub-community that have a similar function in diseases can be considered a topic and can be interpreted by the probability distribution and the profile of bacteria. The next step is to identify the dominant microbe in each sub-community and these sub-communities can be visualized by the level of overlap to indicate community interaction [19]. The microbial sample is considered a document that are composed of mixture of function groups. Each function group are weight mixture of functional elements (words). Topic modeling helps to understand functional group in every sample [20].

Conclusion

Topic models such as LDA has been implemented in various aspects of bioinformatics studies such as microbiome composition analyses, text mining of bioinformatics literatures, pathway-drug-gene relationships, gene ontology, identifying substructures in metabolomics. Topic modeling is a new revolution in the field of bioinformatics as it is useful in revealing underlying semantic structure in a large collection of documents. Topic modeling can help to get a better understanding of healthcare data and can be used for better understand diseases and make therapeutic and clinical decisions.

Table 1: Frequent biological words and their topics

Topics	Epigenetics	Computation	Cancer
Words	Methylation	Algorithm	Oncogene
	Acetylation	Gene	Disease
	MiRNA	Data	Tumor
	Genetics	Model	Death
	DNA	Sequences	Malignant

References

1. Guha, R., *Exploring Information Retrieval by Latent Semantic Indexing and Latent Dirichlet Allocation Techniques.*
2. Yan, J., et al., *MetaTopics: an integration tool to analyze microbial community profile by topic model.* 2017. **18**(1): p. 1-5.
3. La Rosa, M., et al., *Probabilistic topic modeling for the analysis and classification of genomic sequences.* BMC Bioinformatics, 2015. **16**(6): p. S2.
4. Masada, T., et al. *Bayesian multi-topic microarray analysis with hyperparameter reestimation*. in *International Conference on Advanced Data Mining and Applications*. 2009. Springer.
5. Sinkkonen, J., et al. *A simple infinite topic mixture for rich graphs and relational data*. in *NIPS workshop on analyzing graphs: theory and applications*. 2008.
6. Liu, L., et al., *A partially function-to-topic model for protein function prediction.* 2018. **19**(10): p. 51-64.
7. Caldas, J., et al., *Probabilistic retrieval and visualization of biologically relevant microarray experiments.* 2009. **25**(12): p. i145-i153.
8. Coelho, L.P., T. Peng, and R.F.J.B. Murphy, *Quantifying the distribution of probes between subcellular locations using unsupervised pattern unmixing.* 2010. **26**(12): p. i7-i12.
9. Chen, X., et al., *Estimating functional groups in human gut microbiome with probabilistic topic models.* 2012. **11**(3): p. 203-215.
10. Chen, X., et al. *Inferring functional groups from microbial gene catalogue with probabilistic topic models*. in *2011 IEEE International Conference on Bioinformatics and Biomedicine*. 2011. IEEE.
11. Zhang, J., et al., *Inferring functional miRNA–mRNA regulatory modules in epithelial–mesenchymal transition with a probabilistic topic model.* 2012. **42**(4): p. 428-437.
12. Perina, A., et al. *Biologically-aware latent dirichlet allocation (balda) for the classification of expression microarray*. in *IAPR International Conference on Pattern Recognition in Bioinformatics*. 2010. Springer.
13. Pratanwanich, N., P.J.C.b. Lio, and chemistry, *Exploring the complexity of pathway–drug relationships using latent Dirichlet allocation.* 2014. **53**: p. 144-152.
14. Castellani, U., et al. *Brain morphometry by probabilistic latent semantic analysis*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2010. Springer.
15. Juan, L., et al., *Evaluating individual genome similarity with a topic model.* 2020. **36**(18): p. 4757-4764.
16. Pushalkar, S., et al., *The pancreatic cancer microbiome promotes oncogenesis by induction of innate and adaptive immune suppression.* 2018. **8**(4): p. 403-416.
17. Wang, R., et al., *Gut microbiome, liver immunology, and liver diseases.* 2021. **18**(1): p. 4-17.
18. Hosoda, S., et al., *Revealing the microbial assemblage structure in the human gut microbiome using latent Dirichlet allocation.* 2020. **8**(1): p. 1-12.
19. Yan, J., et al., *MetaTopics: an integration tool to analyze microbial community profile by topic model.* BMC Genomics, 2017. **18**(1): p. 962.
20. George, L.E. and L. Birla. *A Study of Topic Modeling Methods*. in *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*. 2018.