A Fuzzy Approach Model for Uncovering Hidden Latent Semantic Structure in Medical Text Collections

Amir Karami, University of Maryland Baltimore County
Aryya Gangopadhyay, University of Maryland Baltimore County
Bin Zhou, University of Maryland Baltimore County
Hadi Kharrazi, Johns Hopkins University

Abstract
One of the challenges for text analysis in the medical domain including the clinical notes and research papers is analyzing large-scale medical documents. As a consequence, finding relevant documents has become more difficult and previous work has also shown unique problems of medical documents. The themes in documents help to retrieve documents on the same topic with and without a query. One of the popular methods to retrieve information based on discovering the themes in the documents is topic modeling. In this paper we describe a novel approach in topic modeling, FATM, using fuzzy clustering. To assess the value of FATM, we experiment with two text datasets of medical documents. The quantitative evaluation carried out through log-likelihood on held-out data shows that FATM produces superior performance to LDA. This research contributes to the emerging field of understanding the characteristics of the medical documents and how to account for them in text mining.

Keywords: Medical Text Analysis; Fuzzy Clustering; Latent Features; Topic Model
Citation: Karami, A., Gangopadhyay, A., Zhou, B., Kharrazi, H. (2015). A Fuzzy Approach Model for Uncovering Hidden Latent Semantic Structure in Medical Text Collections. In *iConference 2015 Proceedings.*

Acknowledgements: I appreciate the feedback and help provided by Dr. Jordan Boyd-Graber for some ideas to better present and describe the purpose of this research.
Contact: amir.karami@umbc.edu

1 Introduction

In the past several years, the medical data have been growing explosively. For example, the number of papers published in PubMed was increased from 112,177 in 1960 to 2,019,238 in 2013 and the annual average number of discharges between 2007 and 2010 is around 35 million[1]. Recently, various text mining techniques have been introduced into the medical domain. One fundamental objective of those techniques is to process the unstructured medical data into a proper format for better utilization to recognize explicit facts. Topic Modeling with Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) is a popular unsupervised method for discovering latent semantic structure of a document collection. Topic modeling has been applied on medical data for different purposes, such as medical document categorization (Sarioglu, Yadav, & Choi, 2013) and medical document retrieval (Huang et al., 2014).

Despite the usefulness of topic models for medical data analysis, existing topic models such as LDA still suffer from several critical issues. One issue of those existing topic models is their computational complexity of the model. Almost all uses of topic models require probabilistic inference, which is arguably hard to achieve without approximate inference algorithms such as Gibbs sampling. Another issue of those existing topic models is their expressive power of representing medical documents. The performance of various tasks such as document classification modeling using topic models is still not satisfactory.

In this paper, we propose to model medical documents using fuzzy set theory. Fuzzy set theory models membership of objects using a possibility distribution. To the best of our knowledge, this is the first study in the medical domain that has been done to use fuzzy set theory to express semantic properties of words and documents in terms of topics. Compared with existing topic models such as LDA, the fuzzy set theory is computationally efficient. We develop several efficient strategies to model medical documents using fuzzy set theory. Regarding the expressive power, we adopt real medical document collections and compare the performance of our proposed method with LDA by considering document modeling. The experimental results showed major improvements.

[1]http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/6222/version/1

The remainder of this paper is organized as follows: In Section 2, we review some relevant studies including topic models and fuzzy set theory. In Section 3, we present our fuzzy set theory based model in detail. An empirical study was conducted to verify the effectiveness of our method and the results are provided in Section 4. Finally, we present a summary and future directions in Section 5.

2 Related Work

There are two major research areas in mining medical documents. The first one tracks concepts by looking for frequency of words (Poulin et al., 2014). The second area categorizes the concepts to find latent variables in medical documents (Lin, Karakos, Demner-Fushman, & Khudanpur, 2006). The first approach leads to high sparse dimensionality data (Aggarwal & Zhai, 2012); therefore, researchers have been motivated to use the second approach such as topic modeling. Among topic models, LDA is a popular and effective unsupervised topic model (Halpern, Horng, Nathanson, Shapiro, & Sontag, 2012).

In the medical domain, LDA has been leveraged in a wide range of applications. For example, Arnold et al. (2010) used LDA for comparing the topics of patient notes (Arnold, El-Saden, Bui, & Taira, 2010), and Bisgin et al. (2011) used LDA in FDA drug side effects labels to cluster drugs (Bisgin et al., 2012).

One of the methods that has not been fully considered in medical text mining is fuzzy set theory. Fuzzy set theory has been used to model systems that are hard to be defined precisely. It incorporates imprecision and subjectivity of human decision making into the model formulation and solution process (Karami & Guo, 2012). Since Bellman and Zadeh (Bellman & Zadeh, 1970) described the decision-making method in fuzzy environments, an increasing number of studies have dealt with uncertain fuzzy problems by applying fuzzy set theory (Karami & Guo, 2012; Karami, Yazdani, Beiryaie, & Hosseinzadeh, 2010). Some work have been done in medical text mining using fuzzy clustering to cluster or classify the documents without any knowledge about the latent semantic of documents (Ben-Arieh & Gullipalli, 2012; Fenza, Furno, & Loia, 2012). In addition, we recently used fuzzy clustering as a feature transformation (dimension reduction) approach for medical text data (Karami & Gangopadhyay, 2014). Among fuzzy clustering methods, Fuzzy C-means (FCM) (Bezdek, 1981) is the most popular one (Bataineh, Naji, & Saqer, 2011). In this research, we propose a novel method using fuzzy clustering to extract latent semantic features from medical documents.

3 FATM

In this section, we detail our *Fuzzy Approach Topic Model (FATM)* and describe the steps. In this algorithm, the output of each step is the input of the next step(s).

Step 1: The first step is to calculate Local Term Weighting (LTW). Among different LTW methods we use term frequency as a popular method. There are n documents and m terms; therefore, the output of this step is $n \times m$ document-term frequency matrix.

Step 2: The next step is to calculate Global Term Weighting (GTW) for document-term frequency matrix. We explore four GTW methods including *Entropy, Inverse Document Frequency (IDF), Probabilistic Inverse Document Frequency (ProbIDF),* and *Normal*. The output of this step is weighted document-term frequency matrix.

Step 3: Fuzzy set theory has been used to model systems in order to assign an instance to a set (Karami & Guo, 2012). Fuzzy clustering is a soft clustering technique that finds the degree of membership for each data point with respect to each cluster or topic (T), as opposed to assigning a data point only one cluster or topic. We use FCM whose goal is to minimize an objective function by considering constraints. The output of FCM is the degree of membership which is between 0 and 1 like probability (P). We run FCM on weighted document-term matrices of step 2 to find the membership degrees or $P(T_k|D_j)$.

Step 4: In this step we use the weighted document-term matrices of step 2 to find $P(D_j)$ using:

$$P(D_j) = \frac{\sum_{i=1}^{m} (W_i, D_j)}{\sum_{i=1}^{m} \sum_{j=1}^{n} (W_i, D_j)} \quad (1)$$

Step 5: The next step is to find $P(D_j|T_k)$. Fist, we calculate:

$$P(D_j, T_k) = P(T_k|D_j) \times P(D_j) \quad (2)$$

Then we normalize $P(D, T)$ in each topic:

$$P(D_j|T_k) = \frac{P(D_j, T_k)}{\sum_{j=1}^{n} P(D_i, T_k)} \quad (3)$$

Step 6: We do a similar calculation in step 5 to find $P(W_i|D_j)$:

$$P(W_i|D_j) = \frac{P(W_i, D_j)}{\sum_{i=1}^{m} P(W_i, D_j)} \quad (4)$$

Step 7: The final step is to find $P(W_i|T_k)$ by:

$$P(W_i|T_k) = \sum_{j=1}^{"\backslash"} P(W_i|D_j) \times P(D_j|T_k) \quad (5)$$

4 Experimental Results

To evaluate the value of FATM , we conduct a comparison of FATM to LDA according to document modeling. Topic models are learned using five methods on the same training sets: LDA and FATMs with Entropy, IDF, Normal, and ProbIDF. We use a Matlab package for Chib-style estimation[1], MALLET package[2] with its default setting for implementing LDA, and Matlab fcm package[3] with its default setting for implementing FCM clustering. We leverage two available datasets in this research. The first dataset[4] is a labeled corpus of English scientific medical abstracts from the Springer website with 5 journals including: Arthroscopy, Federal Health Standard sheet, The Anesthetist, The Surgeon, and The Gynecologist with 1527 documents and 14411 terms. The second dataset called Deidentified Medical Text[5] is an unlabeled corpus of 1607 nursing notes with 11,059 terms. We remove the MALLET list of stop words from each corpus.

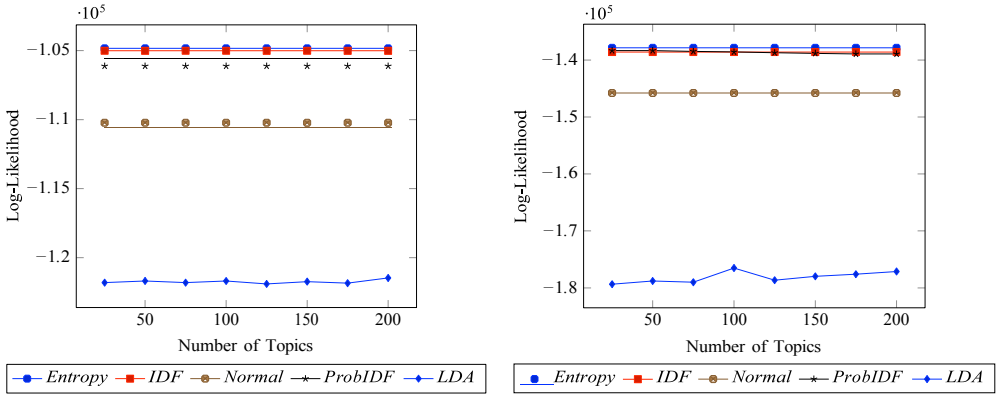
Document modeling produces the topic models on the training sets to produce topic assignments for the held-out documents. Then we learn topics from the larger set and calculate log-likelihood for the smaller set, $P(D_{test}|T)$. The documents in the corpora are treated as unlabeled; thus, our goal is density estimation to achieve high likelihood on a held-out test set. We split the first and the second dataset into two sublets with 90% and 10% of the dataset respectively.

There are different methods to calculate log-likelihood; Among them Chib-style estimation shows better performance (Wallach, Murray, Salakhutdinov, & Mimno, 2009). We compare FATMs with LDA and the result shows that FATMs have a better performance over LDA with different number of topics. Figures 1.a and 1.b present the log-likelihood using Chib-style estimation for each model on both corpora for different number of topics.

5 Conclusion

Analyzing a large volume of medical data is important to advance state-of-the-art healthcare. Due to the unstructured nature of free-text format for the medical data, text mining techniques such as topic modeling are widely adopted to extract latent semantic properties of a medical corpus. Despite the usefulness of topic models for medical data analysis, existing topic models such as LDA still suffer from several critical issues, such as extremely high computational complexity and unsatisfactory performance for data analytical tasks. In this paper, we proposed to use fuzzy set theory, the fuzzy clustering technique in particular, for modeling

[1] http://www.cs.umass.edu/~wallach/code/etm/
[2] http://mallet.cs.umass.edu/
[3] http://www.mathworks.com/help/fuzzy/fcm.html
[4] http://muchmore.dfki.de/resources1.htm
[5] http://physionet.org/



(a) Likelihood for the First Dataset (b) Likelihood for the Second Dataset

Figure 1: Document Modeling Evaluation

unstructured medical documents. The experiments on FATMs and LDA indicate that FATMs yields the best performance from a quantitative standpoint.

There are several interesting directions to explore in future including prediction of stages of various diseases in healthcare, as well as applying our FATM method on large-scale medical data to provide accurate predictions for patients.

References

Aggarwal, C. C., & Zhai, C. (2012). An introduction to text mining. In *Mining text data* (pp. 1–10). Springer.

Arnold, C. W., El-Saden, S. M., Bui, A. A., & Taira, R. (2010). Clinical case-based retrieval using latent topic analysis. In *Amia annual symposium proceedings* (Vol. 2010, p. 26).

Bataineh, K., Naji, M., & Saqer, M. (2011). A comparison study between various fuzzy clustering algorithms. *Jordan Journal of Mechanical & Industrial Engineering*, *5* (4).

Bellman, R. E., & Zadeh, L. A. (1970). Decision-making in a fuzzy environment. *Management science*, *17* (4), B–141.

Ben-Arieh, D., & Gullipalli, D. K. (2012). Data envelopment analysis of clinics with sparse data: Fuzzy clustering approach. *Computers & Industrial Engineering*, *63* (1), 13–21.

Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithms*. Kluwer Academic Publishers.

Bisgin, H., Liu, Z., Kelly, R., Fang, H., Xu, X., & Tong, W. (2012). Investigating drug repositioning opportunities in fda drug labels through topic modeling. *BMC bioinformatics*, *13* (Suppl 15), S6.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3* , 993–1022.

Fenza, G., Furno, D., & Loia, V. (2012). Hybrid approach for context-aware service discovery in healthcare domain. *Journal of Computer and System Sciences*, *78* (4), 1232–1247.

Halpern, Y., Horng, S., Nathanson, L. A., Shapiro, N. I., & Sontag, D. (2012). A comparison of dimensionality reduction techniques for unstructured clinical text. In *Icml 2012 workshop on clinical data analysis.*

Huang, Z., Dong, W., Ji, L., Gan, C., Lu, X., & Duan, H. (2014). Discovery of clinical pathway patterns from event logs using probabilistic topic models. *Journal of biomedical informatics*, *47* , 39–57.

Karami, A., & Gangopadhyay, A. (2014). Fftm: A fuzzy feature transformation method for medical documents. *ACL 2014* , 128.

Karami, A., & Guo, Z. (2012). A fuzzy logic multi-criteria decision framework for selecting it service providers. In *45th hawaii international conference on system science (HICSS)* (pp. 1118–1127).

Karami, A., Yazdani, H., Beiryaie, H., & Hosseinzadeh, N. (2010). A risk based model for is outsourcing vendor selection. In *the 2nd ieee international conference on information and financial engineering (ICIFE)* (p. 250-254).

Lin, J., Karakos, D., Demner-Fushman, D., & Khudanpur, S. (2006). Generative content models for structural analysis of medical abstracts. In *Proceedings of the hlt-naacl bionlp workshop on linking natural language and biology* (pp. 65–72).

Poulin, C., Shiner, B., Thompson, P., Vepstas, L., Young-Xu, Y., Goertzel, B., ... McAllister, T. (2014). Predicting the risk of suicide by analyzing the text of clinical notes. *PLOS ONE*, *9* (1), e85733.

Sarioglu, E., Yadav, K., & Choi, H.-A. (2013). Topic modeling based classification of clinical reports..

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning* (pp. 1105–1112).

Table of Figures