Please provide feedback

Please support the ScholarWorks@UMBC repository by emailing scholarworks-group@umbc.edu and telling us what having access to this work means to you and why it's important to you. Thank you.

FFTM: A Fuzzy Feature Transformation Method for Medical Documents

Amir Karami, Aryya Gangopadhyay
Information Systems Department
University of Maryland Baltimore County
Baltimore, MD, 21250
amir3@umbc.edu,gangopad@umbc.edu

Abstract

The vast array of medical text data represents a valuable resource that can be analyzed to advance the state of the art in medicine. Currently, text mining methods are being used to analyze medical research and clinical text data. Some of the main challenges in text analysis are high dimensionality and noisy data. There is a need to develop novel feature transformation methods that help reduce the dimensionality of data and improve the performance of machine learning algorithms. In this paper we present a feature transformation method named FFTM. We illustrate the efficacy of our method using local term weighting, global term weighting, and Fuzzy clustering methods and show that the quality of text analysis in medical text documents can be improved. We compare FFTM with Latent Dirichlet Allocation (LDA) by using two different datasets and statistical tests show that FFTM outperforms LDA.

1 Introduction

The exponential growth of medical text data makes it difficult to extract useful information in a structured format. Some important features of text data are sparsity and high dimensionality. This means that while there may be a large number of terms in most of the documents in a corpus, any one document may contain a small percentage of those terms (Aggarwal and Zhai, 2012). This characteristic of medical text data makes feature transformation an important step in text analysis. Feature transformation is a pre-processing step in many machine-learning methods that is used to characterize text data in terms of a different number of attributes in lower dimensions. This technique has a direct impact on the quality of text mining methods. Topic models such as LDA has been used as one of popular feature transformation techniques (Ramage et al., 2010). However, fuzzy clustering methods, particularly in combination with term weighting methods, have not been explored much in medical text mining.

In this research, we propose a new method called FFTM to extract features from free-text data. The rest of the paper is organized in the following sections. In the section 2, we review related work. Section 3 contains details about our method. Section 4 describes our experiments, performance evaluation, and discussions of our results. Finally we present a summary, limitations, and future work in the last section.

2 Related Work

Text analysis is an important topic in medical informatics that is challenging due to high sparse dimensionality data. Big dimension and diversity of text datasets have been motivated medical researchers to use more feature transformation methods. Feature transformation methods encapsulate a text corpus in smaller dimensions by merging the initial features. Topic model is one of popular feature transformation methods. Among topic models, LDA (Blei et al., 2003) has been considered more due to its better performance (Ghassemi et al., 2012; Lee et al., 2010).

One of methods that has not been fully considered in medical text mining is Fuzzy clustering. Although most of Fuzzy Clusterings work in medical literature is based on image analysis (Saha and Maulik, 2014; Cui et al., 2013; Beevi and Sathik, 2012), a few work have been done in medical text mining (Ben-Arieh and Gullipalli, 2012; Fenza et al., 2012) by using fuzzy clustering. The main difference between our method and other document fuzzy clustering such as (Singh et al., 2011) is that our method use fuzzy clustering and word weighting as a pre-processing step for

Proceedings of the 2014 Workshop on Biomedical Natural Language Processing (BioNLP 2014), pages 128–133,
Baltimore, Maryland USA, June 26-27 2014.

feature transformation before implementing any classification and clustering algorithms; however, other methods use fuzzy clustering as a final step to cluster the documents. Our main contribution is to improve the quality of input data to improve the output of fuzzy clustering. Among fuzzy clustering methods, Fuzzy C-means (Bezdek, 1981) is the most popular one (Bataineh et al., 2011). In this research, we propose a novel method that combines local term weighting and global term weighting with fuzzy clustering.

3 Method

In this section, we detail our Fuzzy Feature Transformation Method *(FFTM)* and describe the steps. We begin with a brief review of LDA.

LDA is a topic model that can extract hidden topics from a collection of documents. It assumes that each document is a mixture of topics. The output of LDA are the topic distributions over documents and the word distributions over topics. In this research, we use the topics distributions over documents. LDA uses term frequency for local term weighting.

Now we introduce FFTM concepts and notations. This model has three main steps including *Local Term Weighting (LTW), Global Term Weighting (GTM), and Fuzzy Clustering* (Algorithm 1). In this algorithm, each step is the out put of each step will be the input of the next step.

Step 1: The first step is to calculate LTW. Among different LTW methods we use term frequency as a popular method. Symbol f_{ij} defines the number of times term i happens in document j.We have n documents and m words.Let

$$b(f_{ij}) = \begin{cases} 1 & f_{ij} > 0 \\ 0 & f_{ij} = 0 \end{cases} \quad (1)$$

$$p_{ij} = \frac{f_{ij}}{\sum_j f_{ij}} \quad (2)$$

The outputs of this step are $b(f_{ij})$, f_{ij} , and p_{ij}. We use them as inputs for the second step.

Step 2: The next step is to calculate GTW. We explore four GTW methods in this paper including *Entropy, Inverse Document Frequency (IDF), Probabilistic Inverse Document Frequency (ProbIDF),* and *Normal*(Table 1).

IDF assigns higher weights to rare terms and lower weights to common terms (Papineni, 2001). ProbIDF is similar to IDF and assigns very low negative weight for the terms happen in every document (Kolda, 1998). In Entropy, it gives higher weight for the terms happen less in few documents (Dumais, 1992). Finally, Normal is used to correct discrepancies in document lengths and also normalize the document vectors. The outputs of this step are the inputs of the last step.

Step 3: Fuzzy clustering is a soft clustering technique that finds the degree of membership for each data point in each cluster, as opposed to assigning a data point only one cluster. Fuzzy clustering is a synthesis between clustering and fuzzy set theory. Among fuzzy clustering methods, Fuzzy C-means (FCM) is the most popular one and its goal is to minimize an objective func-

Algorithm 1 FFTM algorithm

Functions:E():Entropy;I():IDF;PI():ProbIDF; NO():Normal; FC():Fuzzy Clustering.
Input: Document Term Matrix
Output: Clustering membership value (μ_{ij}) for all documents and clusters.

Remove stop words
Step 1: Calculate LTW
for$i = 1$ to n**do**
  for$j = 1$ to m**do**
    Calculate f_{ij}, $b(f_{ij})$, p_{ij}
  endfor
endfor
Step 2: Calculate GTW
for$i = 1$ to n**do**
  for$j = 1$ to m**do**
    Execute E(p_{ij},n),I(f_{ij},n),PI($b(f_{ij})$,n), NO(f_{ij},n)
  endfor
endfor
Step 3: Perform Fuzzy Clustering
Execute FC(E),FC(I),FC(PI),FC(NO)

Table1: GTW Methods

Name	Formula
Entropy	$1 + \frac{\sum_j p_{ij} \log_2(p_{ij})}{\log_2 n}$
IDF	$\log_2 \frac{n}{\sum_j f_{ij}}$
ProbIDF	$\log_2 \frac{n - \sum_j b(f_{ij})}{\sum_j b(f_{ij})}$
Normal	$\frac{1}{\sqrt{\sum_j f_{ij}^2}}$

tion by considering constraints:

$$Min\ J_q(\mu, V, X) = \sum_{i=1}^{c}\sum_{j=1}^{n}(\mu_{ij})^q D_{ij}^2 \quad (3)$$

subject to:

$$0 \leq \mu_{ij} \leq 1; \quad (4)$$

$$i \in \{1, .., c\}\ and\ j \in \{1, ..., n\} \quad (5)$$

$$\sum_{i=1}^{c} \mu_{ij} = 1 \quad (6)$$

$$0 < \sum_{j=1}^{n} \mu_{ij} < n; \quad (7)$$

Where:

n= number of data
c= number of clusters
μ_{ij}= membership value
q= fuzzifier, $1 < q \leq \infty$
V= cluster center vector
$D_{ij} = d(x_j, v_i)$= distance between x_j and v_i

By optimizing eq.3:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c}(\frac{D_{ij}}{D_{kj}})^{\frac{2}{q-1}}} \quad (8)$$

$$v_i = \frac{\sum_{j=1}^{n}(\mu_{ij})^q x_j}{\sum_{j=1}^{n}(\mu_{ij})^q} \quad (9)$$

The iterations in the clustering algorithms continue till the the maximum changes in μ_{ij} becomes less than or equal to a pre-specified threshold. The computational time complexity is $O(n)$. We use μ_{ij} as the degree of clusters' membership for each document.

4 Experimental Results

In this section, we evaluate FFTM against LDA using two measures: document clustering internal metrics and document classification evaluation metrics by using one available text datasets. We use Weka[1] for classification evaluation, MALLET[2] package with its default setting for implementing LDA, Matlab fcm package[3] for implementing FCM clustering, and CVAP Matlab package[4] for clustering validation.

[1] http://www.cs.waikato.ac.nz/ml/weka/
[2] http://mallet.cs.umass.edu/
[3] http://tinyurl.com/kl33w67
[4] http://tinyurl.com/kb5bwnm

4.1 Datasets

We leverage two available datasets in this research. Our first test dataset called Deidentified Medical Text[5] is an unlabeled corpus of 2434 nursing notes with 12,877 terms after removing stop words. The second dataset [6] is a labeled corpus of English scientific medical abstracts from Springer website. It is included 41 medical journals ranging from Neurology to Radiology. In this research, we use the first 10 journals including: Arthroscopy, Federal health standard sheet, The anesthetist, The surgeon, The gynecologist, The dermatologist, The internist, The neurologist, The Ophthalmology, The orthopedist, and The pathologist. In our experiments we select three subsets from the above journals, the first two with 4012 terms and 171 documents, first five with 14189 terms and 1527 documents, and then all ten respectively with 23870 terms and 3764 documents to track the performance of FFTM and LDA by increasing the number of documents and labels.

4.2 Document Clustering

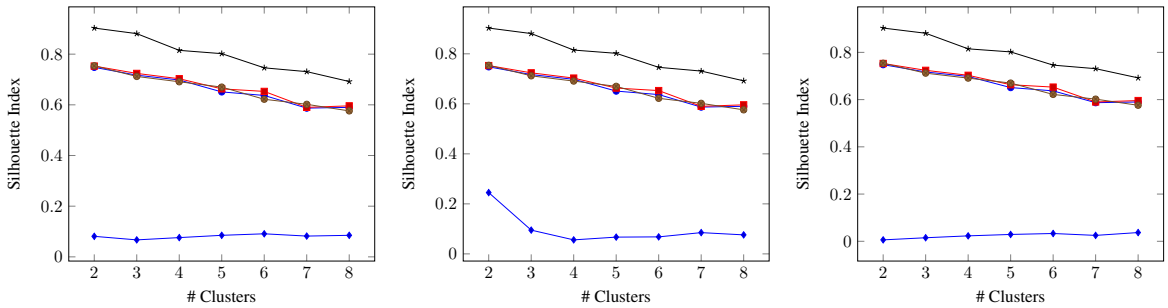
The first evaluation comparing FFTM with LDA is document clustering by using the first dataset. Internal and external validation are two major methods for clustering validation; however, comparison between these two major methods shows that internal validation is more more precise (Rendón et al., 2011). We evaluate different number of features (topics) and clusters by using two internal clustering validation methods including Silhouette index and Calinski-Harabasz index using K-means with 500 iterations. Silhouette index shows that how closely related are objects in a cluster and how distinct a cluster from other other clusters. The higher value means the better result.The Silhouette index (S) is defined as:

$$S(i) = \frac{(b(i) - a(i))}{Max\{a(i), b(i)\}} \quad (10)$$

Where $a(i)$ is the average dissimilarity of sample i with the same data in a cluster and $b(i)$ is the minimum average dissimilarity of sample i with other data that are not in the same cluster.

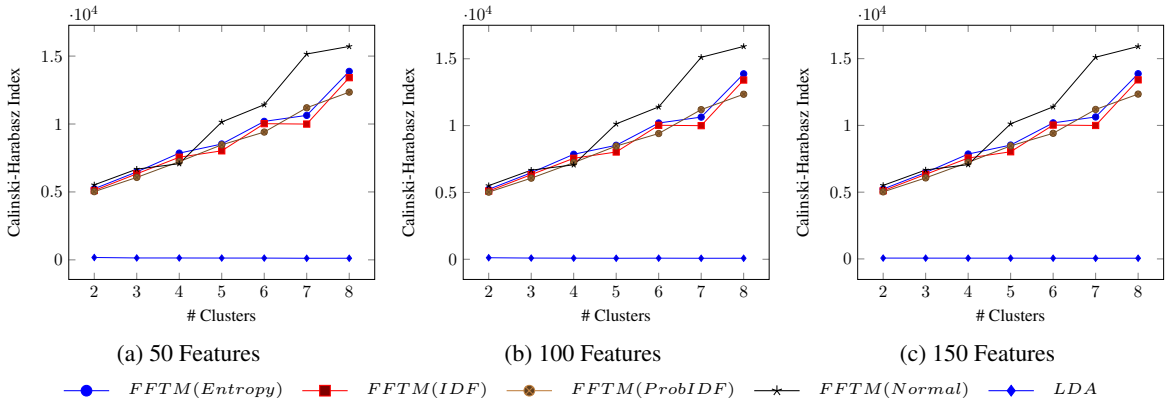
Calinski-Harabasz index (CH) valuates the cluster validity based on the average between- and within-cluster sum of squares.It is defined as:

[5] http://tinyurl.com/kfz2hm4
[6] http://tinyurl.com/m2c8se6



(a) 50 Features (b) 100 Features (c) 150 Features

Figure1: Clustering Validation with Silhouette Index



(a) 50 Features (b) 100 Features (c) 150 Features

Figure2: Clustering Validation with Calinski-Harabasz Index

$$CH = \frac{trace(S_B)}{trace(S_W)} . \frac{n_p - 1}{n_p - k} \qquad (11)$$

Where (S_B) is the between-cluster scatter matrix, (S_W) the internal scatter matrix, n_p is the number of clustered samples, and k is the number of clusters. Higher value indicates a better clustering. We track the performance of both FFTM and LDA using different number of clusters ranging from 2 to 8 with different number of features including 50, 100, and 150. Both Silhouette index and Calinski-Harabasz index show that FFTM is the best method with all ranges of features and clusters (Figures 1 and 2). The gap between FFTM and LDA does not change a lot by using different number of features and clusters. LDA has the lowest performance and Normal has the best performance among GTW methods in different ranges of features and clusters. According to the paired difference test, the improvement of FFTM over LDA is statistically significant with a $p - value < 0.05$ using the two internal clustering validation methods.

4.3 Document Classification

The second evaluation measure is document classification by using the second datasest. We evaluate different number of classes and features (topics) with accuracy, F-measure, and ROC using Random Forest. Accuracy is the portion of true results in a dataset. F-measure is another measure of classification evaluation that considers both precision and recall. ROC curves plot False Positive on the X axis vs. True Positive on the Y axis to find the trade off between them; therefore, the closer to the upper left indicates better performance. We assume more documents and classes have more topics;therefore, we choose 100 features for two classes, 150 features for five classes, and 200 features for ten classes. In addition, we use 10 cross validation as test option.

This experiment shows that FFTM has the best performance in different number of features and labels (Table 2). LDA has the lowest performance and the average performance of ProbIDF has the best among GTW methods in all ranges of features and clusters. According to the paired difference test, the improvement of FFTM over LDA is statistically significant with a $p - value < 0.05$.

Table2: The Second Dataset Classification Performance

Method	#Features	# Labels	Acc %	F-Measure	ROC
FFTM(Entropy)	100	2	96.49	0.959	0.982
FFTM(IDF)	100	2	98.24	0.982	0.996
FFTM(ProIDF)	100	2	97.66	0.977	0.987
FFTM(Normal)	100	2	92.39	0.912	0.971
LDA	100	2	90.06	0.9	0.969
FFTM(Entropy)	150	5	71.84	0.694	0.874
FFTM(IDF)	150	5	70.79	0.686	0.859
FFTM(ProIDF)	150	5	70.39	0.674	0.859
FFTM(Normal)	150	5	68.11	0.649	0.851
LDA	150	5	66.27	0.637	0.815
FFTM(Entropy)	200	10	51.06	0.501	0.828
FFTM(IDF)	200	10	51.73	0.506	0.826
FFTM(ProIDF)	200	10	53.72	0.525	0.836
FFTM(Normal)	200	10	50.05	0.485	0.815
LDA	200	10	47.68	0.459	0.792

5 Conclusion

The explosive growth of medical text data makes text analysis as a key requirement to find patterns in datasets;however, the typical high dimensionality of such features motivates researchers to utilize dimension reduction techniques such as LDA. Although LDA has been considered more recently in medical text analysis (Jimeno-Yepes et al., 2011), fuzzy clustering methods such as FCM has not been used in medical text clustering, but rather in image processing. In the current study, we propose a method called FFTM to combine LTW and GTM with Fuzzy clustering, and compare its performance with that of LDA. We use different sets of data including different number of features, different number of clusters, and different number of classes.The findings of this study show that combining FCM with LTW and GTW methods can significantly improve medical documents analysis. We conclude that different factors including number of features, number of clusters, and classes can affect the outputs of machine learning algorithms. In addition, the performance of FFTM is improved by using GTW methods. This method proposed in this paper may be applied to other medical documents to improve text analysis outputs. One limitation of this paper is that we use one clustering method, one classification method, and two internal clustering validation methods for evaluation. Our future direction is to explore more machine learning algorithms and clustering validation methods for evaluation and also other fuzzy clustering algorithms for feature transformation. The main goal of future research is to present an efficient and effective medical topic model using fuzzy set theory.

References

CharuC Aggarwal and ChengXiang Zhai. 2012. An introduction to text mining. In *Mining Text Data*, pages 1–10. Springer.

KMBataineh, MNaji, and MSaqer. 2011. A comparison study between various fuzzy clustering algorithms. *Jordan Journal of Mechanical & Industrial Engineering*, 5(4).

Zulaikha Beevi and Mohamed Sathik. 2012. A robust segmentation approach for noisy medical images using fuzzy clustering with spatial probability. *International Arab Journal of Information Technology (IAJIT)*, 9(1).

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

JamesC Bezdek. 1981. *Pattern recognition with fuzzy objective function algorithms*. Kluwer Academic Publishers.

DavidM Blei, AndrewY Ng, and MichaelI Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Wenchao Cui, YiWang, Yangyu Fan, Yan Feng, and Tao Lei. 2013. Global and local fuzzy clustering with spatial information for medical image segmentation. In *Signal and Information Processing (ChinaSIP), 2013 IEEE China Summit & International Conference on*, pages 533–537. IEEE.

Susan Dumais. 1992. Enhancing performance in latent semantic indexing (lsi) retrieval.

Giuseppe Fenza, Domenico Furno, and Vincenzo Loia. 2012. Hybrid approach for context-aware service discovery in healthcare domain. *Journal of Computer and System Sciences*, 78(4):1232–1247.

Marzyeh Ghassemi, Tristan Naumann, Rohit Joshi, and Anna Rumshisky. 2012. Topic models for mortality modeling in intensive care units. In *ICML Machine Learning for Clinical Data Analysis Workshop*.

Antonio Jimeno-Yepes, Bartłomiej Wilkowski, JamesG Mork, Elizabeth VanLenten, DinaDemner Fushman, and AlanR Aronson. 2011. A bottom-up approach to medline indexing recommendations. In *AMIA Annual Symposium Proceedings*, volume 2011, page 1583. American Medical Informatics Association.

TamaraG Kolda. 1998. Limited-memory matrix methods with applications.

Sangno Lee, Jeff Baker, Jaeki Song, and JamesC Wetherbe. 2010. An empirical comparison of four text mining methods. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, pages 1–10. IEEE.

Kishore Papineni. 2001. Why inverse document frequency? In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics.

Daniel Ramage, SusanT Dumais, and DanielJ Liebling. 2010. Characterizing microblogs with topic models. In *ICWSM*.

Eréndira Rendón, Itzel Abundez, Alejandra Arizmendi, and ElviaM Quiroz. 2011. Internal versus external cluster validation indexes. *International Journal of computers and communications*, 5(1):27–34.

Indrajit Saha and Ujjwal Maulik. 2014. Multiobjective differential evolution-based fuzzy clustering for mr brain image segmentation image segmentation. In *Advanced Computational Approaches to Biomedical Engineering*, pages 71–86. Springer.

VivekKumar Singh, Nisha Tiwari, and Shekhar Garg. 2011. Document clustering using k-means, heuristic k-means and fuzzy c-means. In *Computational Intelligence and Communication Networks (CICN), 2011 International Conference on*, pages 297–301. IEEE.