ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 10, May 2016

A Nonnegative Matrix Approximation Based Multi-View NMF Medication and Symptom Names Extraction for Clustering of Clinical Documents

Manjunath Varchagall [1], Gurubai Rampure [2]

P.G. Student, Department of Computer science Engineering, AMC Engineering College, Bannerughatta, Karnataka, India[1]

P.G. Student, Department of Software Engineering, AMC Engineering College, Bannerughatta, Karnataka, India [2]

ABSTRACT: Clinical documents are large free-text data sources accommodate consisting of symptom and medication details, which have an enormous possibilities to upgrading health care of patients. In this paper, we construct an integrating system for bring out symptom names and medication names from clinical notes. Then we put in an application nonnegative matrix approximation (NMF) and parallel-view NMF to cluster clinical notes into significant clusters based on sample-feature matrices. Compare the results of parallel-view NMF and NMF results which is a better suitable technique for clinical document clustering. Moreover, we find that using extracted symptom/medication names to cluster clinical documents more profitable than just using words.

KEYWORDS: Clinical document, clinical notes, document clustering, Parallel-view, nonnegative matrix approximation.

I INTRODUCTION

A clinical note which is clinical documents contains a large amount of valuable details regarding patients, such as responses (procedures, diagnoses and drugs) and medication conditions (injuries, diseases and medical symptoms, etc.). These underutilized assets have a enormous possibilities to enhance patient physical fitness care. These kind of valuable details bring out from clinical notes that can be used to construct outline for independent patients, come across disease association, and improve the patient responsibility. Medications and Symptoms are two essential varieties of details these can be acquired from clinical notes. Symptom associated details such as syndromes, diseases, diagnoses and signs etc. These can be used to examine sickness for patients. In inclusion, valuable medication details are frequently inserted in unstructured notes such as narratives. These narratives in the form of, for admission notes, instances or treatment plans, which can be written by clinicians and play a key task in allowing clinicians to communicate observations and reasoning in adaptable manner. These narratives can have spanning multiple segments in clinical documents.

Medication details from clinical notes are frequently demonstrated with medication names and other identification details regarding medical drug management, such as the size or frequency of a dose of a medicine, route which is a method of transmitting a disease or of administering a remedy, duration and frequency. In this project, we bring out medication labels from clinical notes. Other associated medication details is also extremely essential, and will be think about in upcoming investigation. A short time ago, huge capacity of clinical documents is created by EHRs (Electronic Health Record systems). The created clinical documents are semi-structured or unstructured. It is a complicated job to bring out details from those created documents. Symptom related details and medication related details bring out for clinical notes require knowledgeable clinical language processing technique.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 10, May 2016

Due to the independent assortment, it is a competitive difficulty to come across the underlying patterns of sequence or framework from a collection of clinical documents. Document clustering approach as a systematic process of navigating and recapitulating documents have accepted a large amount of observations.

Clustering of clinical documents have been exploring for categorize clinical documents into significant clusters, in order to come across patterns of sequences or framework and significant characteristics. Patterson et al (this is name of reference paper author) grouping a data set which is composed of 17 clinical note varieties utilize an unsupervised clustering techniques and express dissimilar clinical territory use dissimilar semantic and lexical patterns.

Then reference paper of Doing-Harris recognize medical specialty covering institution by differentiate linguistic characteristics of clinical notes from dissimilar organization using document clustering method. Then another reference paper Han appoint latent semantic indexing to group of clinical notes and found that latent semantic indexing was a powerful technique for calculating the equivalence of clinical notes. Another reference paper of Zhang evaluated 9 semantic equivalence measures of ontology-related expression for medical document clustering.

We compute the effects of integrating medication/symptom names for clinical documents clustering. Nonnegative matrix approximation (NMF) has been extensively applied to document clustering. Akata(reference author name) enlarged NMF towards joint NMF, which can together examine dissimilar kinds of characteristics for parallel-view learning. Alternatively to placing a usual clustering solution for each vision, another paper of Liu additionally constructs the procedure by detecting an adjacent concurrence for each view. Parallel-view NMF can consolidate different sources of data and produce a superior clustering solution. In this paper Our involvement consisting are 3-folds: 1) we present a system for bring out medication/symptom names from medical record; 2) Then we apply parallel-view NMF to determine the effects of using symptom/medication names to enhance the clustering of clinical documents solution; 3) Then we differentiate the performances of NMF and parallel-view NMF on clustering of clinical documents.

The remaining of the paper is organized as follows: 2 introduce the general review of extracting medication names and symptom names from medical records. 3 describe NMF and parallel-view NMF.

II. EXTRACTION OF SYMPTOM/ MEDICATION NAME

A. Clinical Documents

Medical record is an essential part of patient records in an unstructured or semi- structured free-text format. An example of a medical record with a few selected sections is shown in Fig. 1.

PRINCIPAL DIAGNOSIS: *Cellulitis*.
LIST OF PROBLEMS/DIAGNOSES: 1. *Inflammatory breast cancer*. 2. *Type II diabetes*. 3. *Hypertension*. 4. *Hypercholesterolemia*. 5. *CHF* with preserved systolic function. 6. *Obstructive sleep apnea*. 7. *Asthma*. 8. *Spinal stenosis* with herniated discs.
MEDICINES: 1. *Lantus* 40 units nightly. 2. *Aspirin* 81 mg daily. 3. *Lipitor* 40 mg daily. 4. *Zestril* 2.5 mg daily. 5. *Cardizem ER* 240 mg daily. 6. *Lasix* 20 mg daily. 7. *Procrit* 40000 units weekly. 8. *Pamidronate*. 9. *Dexamethasone* with chemotherapy. 10. AC chemo. 11. *Neulasta*. 12. *Ativan p.r.n*. 13. *Multivitamin*. 14. *Iron sulfate*. 15. *Isosorbide dinitrate* 10 mg t.i.d. 11. *Allegra* 60 , 000 mg b.i.d.

Fig. 1. An example of selected sections from Medical record.

History of Present Illness, Discharge Medications, Hospital Course, Hospital Course By System, Brief Resume of Hospital Course and Hospital Course By Problem are the most frequent sections consisting of both medication names and symptom names.

B. Extraction of Name

An general overview of extracting medications and symptom from medical records is showed in Fig. 2. We bring out the symptom names such as “hypertension” and medication names such as “Cardizem, and Isordil” from the medical

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 10, May 2016

texts "He was kept off aspirin given his GI bleeding. The patient also has hypertension and was on Isordil and Cardizem for that.

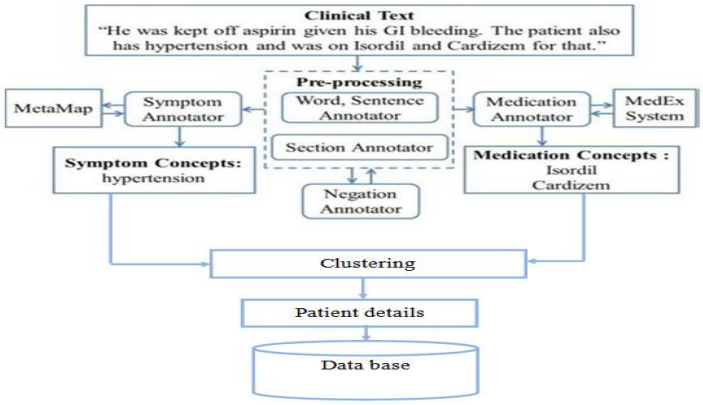

Fig. 2. An general overview of symptom/medical term extraction from Clinical Notes.

In the first step, we pre-process medical record to recognize sentences and words from medical records by using Stanford Core Natural Language Processing (NLP) Tool (http:/ /nlp.stanford.edu/downloads/). Throughout the pre-processing step, the task of section annotator to recognize dissimilar sections for each medical record. The section annotator dependent on the section header details from medical notes. Negation sections, such as "ALLERGIES" or "Family History," are excluded.

For example, "He is allergic to MORPHINE" from the section "ALLERGIES," the medication name "MORPHINE" is a negation medication name, so we exclude it.

The negation symptom and medication names can be removed by the help of negation annotator. An example is that "The patient was taking told to avoid aspirin or any other NSAIDs given his GI bleed," we remove "aspirin" and "NSAIDs" because of the pre-negation words "avoid." Pre-negation and post-negation are describe in Negation maker (NegEx: http://www.dbmi.pitt.edu/chapman/NegEx.html). Negation words like free, was ruled out, and so on. Pre-negation is negation words like cannot, without, avoid, deny, and so on.

After pre-process, the extraction of symptom names from medical notes by the help of symptom annotator based on the MetaMap. Meanwhile, the extraction of medication names from clinical notes with the help of medication annotator based on MedEx System. The extraction of symptom names from clinical notes with the use of MetaMap. The program that maps biomedical texts to concepts in the UMLS Meta-thesaurus is called MetaMap (http://nls3.nlm.nih.gov). Since Metamap returns all types of concepts, we only keep these concepts associated to symptom names, such as idea labeled as "sosy," which constitutes "sign and symptom." The associated kinds of concepts include: acab, inpo, lbtr, mobd, anab, comd,sosy, neop, dsyn,bact, dsyn, cgab, virs.

Extraction of medication names from clinical notes with the use of the MedEx system. MedEx system is a natural language processing system to extract medication details from clinical notes. In medical notes, medication data are frequently indicate in medication names and identification details about drug management.

The system such as MedEx extracts multiple semantic classification of medication detection from clinical notes, such as DrugName, Frequency, Route, Strength, Dose Amount, Form, Duration, IntakeTime, Refill, Dispense Amount, and Necessity. Here we are using the Drug Name as medication name.

III. NONNEGATIVE MATRIX APPROXIMATION (NMF).

a. Basic NMF

NMF is an essential technique to factorize a $n \times m$ nonnegative matrix A into the product of two lower dimensional nonnegative matrices: a $n \times k$ matrix W and a $m \times k$ matrix H, which can be expressed as the following optimization problem using the square of Euclidean distance:

$$\min ||A - WH||^2$$

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 10, May 2016

The minimization of cost function applying the update rules as follows:

$$Ha\mu \leftarrow Ha\mu \frac{(W^T A)_{a\mu}}{(W^T WH)_{a\mu}}$$

$$\text{Wia} \leftarrow \text{Wia} \frac{(AH^T)_{ia}}{(WHH^T)_{ia}}$$

Where K constitutes the number of clusters. It can be decided by referring from consensus matrix and cophenetic correlation coefficient .

b. Parallel-View NMF

Parallel-view learning is the extension of NMF. Parallel-view learning objective to recognize latent ingredients in dissimilar sub-matrices in a concurrent manner. These sub-matrices can represent dissimilar features spaces.

Extends the fundamental NMF to convex combination of different views as following optimization problem:

$$min_{W^i}, H \geq 0 \sum_{i=1}^{p} \lambda_i ||A^i - w^i H||^2$$

$$\sum_{i=1}^{p} \lambda_{i=1,} \quad \lambda_i \geq 0.$$

Due to constraint that matrix is fixed among multiple views, additionally extend to resolving the following optimization problem:

$W^i, H^{i^{min}}, H^* \geq 0 \sum_{i=1}^{p} ||A^i - W^i H^i||^2 + \sum_{i=1}^{p} \lambda_i ||H^i - H^*||^2$.

This problem attempts to optimize $A^i \approx W^i H^i$ for each view i, and keep constraining each H^i will be similar.

IV. DATASETS AND PREPROCESSING

A. Datasets Description

Clinical notes dataset contains thousands of clinical notes. After pre-processing, each patient has about 3–5 records. Compared with clinical notes of another dataset, this dataset was applied for the risk factor identification, for example heart disease track. All the risk factors are annotated in these records. We categorize these risk factors into medication names or symptom names. The genuine records have standard to specify 3 kinds of patients. The 1 type is patients who evolve coronary artery disease (CAD); the 2 type is patients who have CAD in their 1 records; and the 3 kind is patients who never progress CAD. We use this as standard to calculate the cluster performance from Parallel-view NMF.

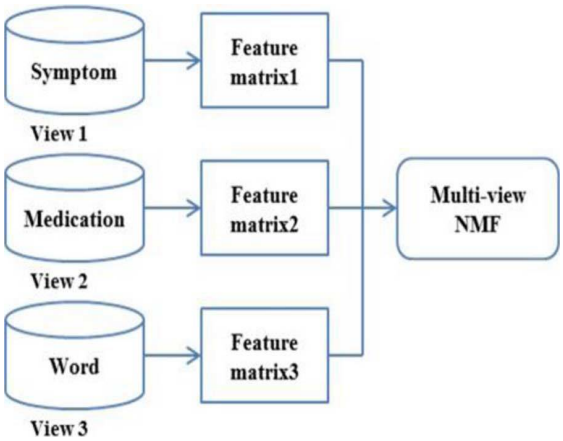

Fig. 3. The skeleton of applying Parallel-view NMF.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 10, May 2016

B. Preprocessing

We pre-process the dataset to create the sample-feature matrices as shown in Fig. 3. For each dataset, we process each medical record as a sample. While for each dataset, each patient is organized as a sample. Each sample can be represented from 3 views: medication names, symptom names and words.

For the words set, we separate repeated stop words and clean the data. We create features from these 3 views using word count or term frequency-inverse document frequency (TF-IDF). After preprocessing, we get these sample-feature matrices.

V. EXPERIMENTAL RESULTS

In both Tables I and II, we use expression checks What's more TF-IDF Concerning illustration features to produce those characteristic matrices. Utilizing manifestation names and prescription names have better correctness what's more NMI over barely utilizing expressions. Utilizing every last bit 3 sees (words, manifestation names, and solution names) together might accomplish those most elevated execution. Those comes about of utilizing the sum three perspectives would compared the middle of NMF Furthermore multi-view NMF are demonstrated Previously, fig. 4. When, utilizing expressions check as characteristic demonstrates that multi-NMF accomplishes around 12% higher precision over NMF. It needs 14% higher correctness the point when utilizing TF-IDF Concerning illustration Characteristics.

TABLE I
2014 DATASET RESULTS (k=3)

Feature Type	Views	Accuracy (%)	NMI
Count	Words	40.54	0.0228
	Symptom/Medication	52.03	0.1273
	All 3 views	***53.38***	***0.1459***
TF-IDF	Words	35.47	0.0020
	Symptom/Medication	52.36	0.1606
	All 3 views	***52.36***	***0.1711***

TABLE II
2014 DATASET RESULTS (k=2)

Feature Type	Views	Accuracy (%)	NMI
Count	Words	57.77	0.0198
	Symptom/Medication	55.07	0.0924
	All 3 views	***59.80***	***0.1751***
TF-IDF	Words	53.38	0.0034
	Symptom/Medication	73.31	0.1844
	All 3 views	***75.00***	***0.2283***

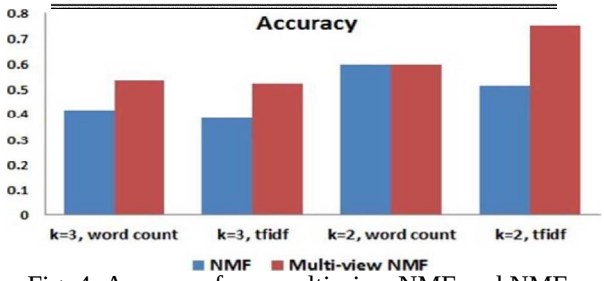

Fig. 4. Accuracy from multi-view NMF and NMF.

VI. CONCLUSION AND FUTURE WORK

In this paper, we are extracting symptom and medication names from semi-structured or unstructured clinical notes to build profile for independent patient. The whole system consists of 5 parts: section annotator; word/sentence annotator; negation annotator; medication name annotator; and symptom name annotator.

We use the extracted symptom/medication names integrated with words as three-views from clinical notes, and then we apply parallel-view(Multi-View) NMF for clustering of documents. We have plan for comparing the 2

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 10, May 2016

different dataset to compute the accuracy by applying both NMF and Parallel-View NMF and NMI as evaluation metrics to compare results. It showed that, the clustering performance can be increases by using medication names and symptom names. It also specify that parallel-view NMF can achieve better results than NMF.

In our further work, we may examine using other details, such as patient's gender/age/demographical details, to upgrading the clustering performance; and also explore intrinsic correspondence among different views. We also plan to use the clustering of clinical document results to improve medication guidance as discussed in our former work. We may consider upgrading the performance of the clinical clustering mechanism so that the computational complexity of the paper can be reduced as well as the load on the server also can be reduced concurrently.

REFERENCES

[1] K. Roberts and S. M. Harabagiu, "A flexible skeleton for deriving assertions from electronic medical records," J. Amer. Med. Informat.Assoc., vol. 18, no. 5, pp. 568–573, 2011.
[2] M.-Y. Kim et al., "Patient information extraction in noisy texts," in Proc. IEEE Int. Conf. Bioinformat. Biomed. (BIBM'13).
[3] F. S. Roque et al., "Using electronic patient records to discover disease correlations and stratify patient cohorts," PLoS Comput. Biol., vol. 7,no. 8, p. E1002141, 2011.
[4] G. Hripcsak et al., "Mining complex clinical data for patient safety research: a framework for event discovery," J. Biomed. Informat., vol. 36, no. 1, pp. 120–130, 2003.
[5] S. V. Pakhomov, A. Ruggieri, and C. G. Chute, "Maximum entropy modeling for mining patient medication status from free text," in Proc. AMIA Symp. Amer. Med. Informat. Assoc., 2002.
[6] A. Henriksson, "Semantic spaces of clinical text: Leveraging distributional semantics for natural language processing of electronic health records," Lic. degree, Dept. Comput. Syst. Sci., Stockholm Univ., Stockholm, Sweden, 2013.
[7] S. Kushinka, "Clinical documentation: EHR deployment techniques," in California HealthCare Found., 2010 [Online].Available:http://www.chcf.org/~/media/MEDIA%20LIBRARY%20Files/PDF/C/PDF%20Clinical DocumentationEHRDeploymentTechniques.pdf
[8] W. W. Chapman et al., "Overcoming barriers to NLP for clinical text:the role of shared tasks and the need for additional creative solutions,"J. Amer. Med. Informat. Assoc., vol. 18, no. 5, pp. 540–543, 2011.
[9] F. H. Saad, B. d. l. Iglesia, and D. G. Bell, "A comparison of two document clustering approaches for clustering medical documents," in Proc. Conf. Data Mining (DMIN), 2006.
[10] O. Patterson and J. F. Hurdle, "Document clustering of clinical narratives: a systematic study of clinical sublanguages," in Proc. AMIA Annu. Symp. Proc. Amer. Med. Informat. Assoc., 2011.