International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 04 | Apr -2017 www.irjet.net p-ISSN: 2395-0072

A Clustering Based Collaborative Approaches for Health Care System Using Clinical Document

Mohanadevi M[1]

[1]*Research Scholar, Dept. of Computer Science, Dr.N.G.P Arts and Science College, Tamilnadu, India*

Abstract - *Clinical documents contain huge of medical information and large free-text data source accommodate consisting manifestation and pharmaceutical details, which have an enormous possibilities to upgrading health care of patients. We define an approach to build a system that firstly pre-processes the clinical documents. The clinical documents are obtained from various hospitals and websites. We have used different tools like MedEx and MetaMap as annotators for extracting the manifestation names and the pharmaceutical names from clinical historical data. For clustering the fetched data we are using the NMF, parallel view NMF and to get Accuracy from k-means, which is a clustering technique.*

Key Words: **Parallel view, Document clustering, k-means clustering, Non-negative matrix approximation.**

1. INTRODUCTION

Data mining is to finds valuable information hidden in large volumes of data. There are many types of data mining. One of the among them is the clustering techniques. Clustering in data mining is the grouping of a particular set of objects based on their characteristics, aggregating them according to their similarities. It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, bioinformatics, data compression, document clustering and computer graphics. There are many application of data mining in medical field, as it has wide spread use in medical area. It is getting great pace in medical research as well as in clinical practice; thus saving time, money, and life.

A Clinical note which is clinical documents contains a large amount of valuable details regarding patients, such as Reponses (diagnoses) and medication conditions (medical symptoms, etc.,).Clinical data mining is the process of applying the data mining techniques on the obtained textual clinical documents. Rich text data source of clinical document contain information about pharmaceutical and manifestation. Extracting this information has proved beneficial so as to help refine the health care system. Clinical documents are widely used for future analysis and diagnosis of the disease[1]. The clinical notes have a great use in pharmacy store so to reduce forgery and prevent drug abuse. This is done primarily to locate important patterns[2].Ashort time ago, huge capacity of clinical documents is created by EHRs (Electronic Health Record systems).Information extraction of theses task in case of machine learning (ML) and Language Processing (NLP) as it involves significant data extraction form Natural language text.

K-means[3] clustering is well known widespread clustering techiniques.It is mostly used to cluster the numerical data. Clustering can therefore be formulated as a mutli-objective optimization problem. The appropriate clustering algorithm and parameters settings (include values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. MedEx[6] tool on clinical note to extract the pharmaceutical name .This tool is used to obtain the pharmaceutical related details such as dose volume, drug name, intake time of medicine etc. Simultaneously we are applying the MetaMap[7] to get the name of the manifestation.

Non-Negative Matrix factorization is group of algorithms in multivariate analysis and linear algebra where a matrix V is factorized into two matrices have no negative elements. Non negative matrix approximation (NMF) has been extensively applied to document clustering. We have built an integrating system which consists of following modules:

1) System for Extraction of manifestation names / pharmaceutical names from clinical notes.
2) Apply Parallel View NMF[8] to estimate the results of using manifestation names/pharmaceutical names in improving the clinical notes clustering. This helps in analyzing the class of features of patterns.
3) Compared the experimental results of parallel view NMF and NMF.
4) Preprocessing of textual data will amplify the performance of clustering .then we apply the k-means clustering on the pre-processed notes.
5) Nonnegative matrix approximation has been extensively applied to document clustering.
6) The Overall system is dived into word/sentence annotator, section annotator, negation annotator,

symptoms annotator, medication annotator, age annotator, climate annotator.

2. RELATED WORK

2.1 Patient Record

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

RECORD #1

505128233 | RH | 36002733 | | 8399692 | 10/18/2005 12:00:00 AM | SMALL BOWEL OBSTRUCTION | Signed | DIS | Admission Date: 10/18/2005 Report Status: Signed
Discharge Date: 11/14/2006
ATTENDING: DEMMER , ROBT MD
PRIMARY CARE PHYSICIAN: Dustin Theurer , MD
CARDIOLOGIST: Hassan Gowins , MD
GASTROENTEROLOGIST: Coy Spurgers , MD
PRINCIPAL DIAGNOSES:
1. Small-bowel obstruction.
2. Congestive heart failure.
3. Dilated idiopathic cardiomyopathy.
HISTORY OF PRESENT ILLNESS: Mrs. Aswegan is a 74-year-old woman , recently admitted to the Kimonte on 7/8/05 for CHF exacerbation and UTI (Gram-negative rods Klebsiella) who presents to the Vide Tutenoke Rahbrier Healthcare with one day of abdominal pain , nausea , vomiting , and decreased ostomy output. The patient has a history of multiple abdominal surgeries including a total colectomy for ulcerative colitis with colostomy , ventral hernia repairs in 1998 , 2003 , and 2004 , revision of her colostomy in 1996 , 1997 , and 2003. As a result , she has had a chronic left lower quadrant hernia. The patient denies any recent fevers , chills , melena , or hematochezia from the ostomy. She has stable dyspnea on exertion after 12 feet, she uses a walker at home. She denies PND. She has two-pillow orthopnea. She has had stable lower extremity edema in the past day and no dysuria.

Fig. 1. An example of selected sections from Medical record.[19]

History of Present Illness, Discharge Medications, Hospital Course, and Hospital Course by System, Brief Resume of Hospital Course and Hospital Course by Problem are the most frequent sections consisting of both Pharmaceutical and Manifestation names.

2.2 Preprocessing

We define an approach to build a system that firstly pre-processes the clinical documents. An example of Pre-Processing- Block diagram a shown in Fig. 2.

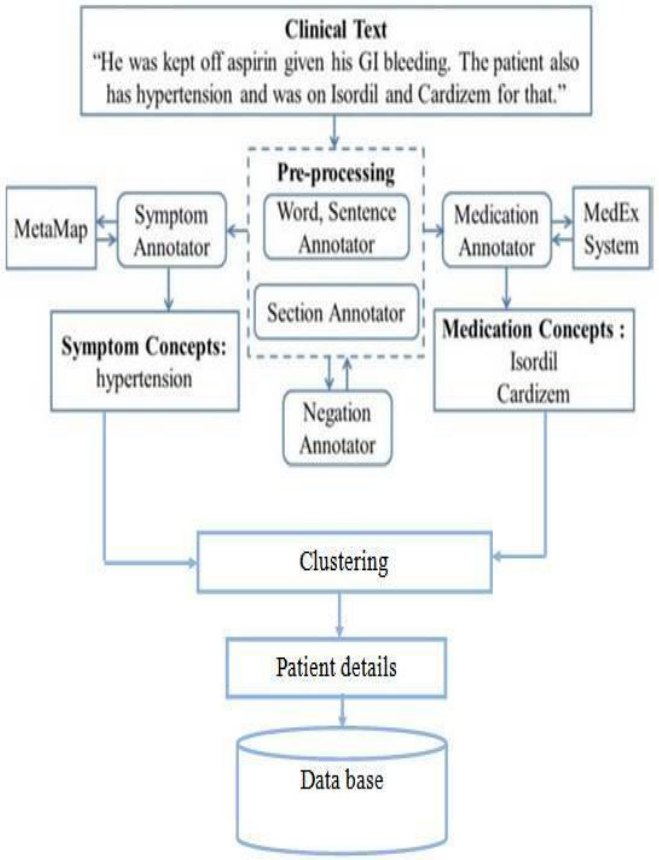

Fig. 2. A general overview of manifestation [symptom]/Pharmaceutical [medical term] extraction from Clinical Notes.[4]

The Fig.2 shows the block diagram of a manifestation and pharmaceutical names extraction from clinical notes. We have clinical text "Patient was suffering from ankle pain and he was taking antihistamine and thiazide". In this clinical text, manifestation (symptom) name is ankle pain. There are two medicines, namely, antihistamine and thiazide. Initially, Section annotator is used to find different sections in clinical document. In order to fetch these pharmaceutical and manifestation names, first we need to remove irrelevant words, using pre-processing. The pre-processing also improves quality of data. StandfordCore NLP tool is used to isolate words and sentences from clinical document. During pre-processing, we have to take out stop words and stem words which are nothing but most common words in English language, such as, this, that, she, he, etc. The output of stop words and stem words removal module is the medical terms along with the negation words. Negation annotator module is used to eliminate negation words like avoided, denies, ruled out etc. The output of negation annotator is only medical related term. The output of pre-processed data is fed into MedEx and MetaMap systems to get medical related terms.

The medical related terms include pharmaceuticals and manifestations. After removing unnecessary content from clinical notes, we are clustering medical related data. We have used multi-view NMF for clustering. Multi-view NMF finds latent components in sub matrices. When a user enters the problem statement containing symptom name, proposed method provides pharmaceutical names using MedEx system and manifestation names using MetaMap. Finally, system provides Ankle pain as manifestation name and pharmaceutical names, such as Antihistamine and thiazide.

3. METHODOLOGY

3.1 Overview of NMF

Non-Negative Matrix factorization is group of algorithms in multivariate analysis and linear algebra where a matrix V is factorized into two matrices have no negative elements. Non-Negative matrix approximation (NMF) has been extensively applied to document clustering. .The Non-negative Matrix Factorization (NMF) analyzed in this thesis can be mathematically described using matrices. This section gives a formal definition for Nonnegative Matrix Factorization problems, and defines the notations used throughout the vignette. Let V be a W×H non-negative matrix. Non-negative Matrix Factorization (NMF) consists in finding an approximation.

$$V \approx W H$$

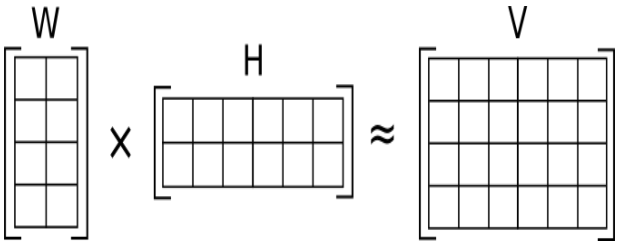

Fig.3. Nonnegative Matrix – NMF

3.2 Parallel view of NMF [Multi View]

A Parallel view NMF is to integrate information from multiple views in the unsupervised setting, multi-view clustering algorithms have been developed to cluster multiple views simultaneously to derive a solution which uncovers the common latent structure shared by multiple views. A novel NMF based multi-view clustering algorithm by searching for a factorization that gives compatible clustering solutions across multiple views.

3.3 Datasets

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. We use this as standard to calculate the cluster performance from Parallel-view NMF[14].

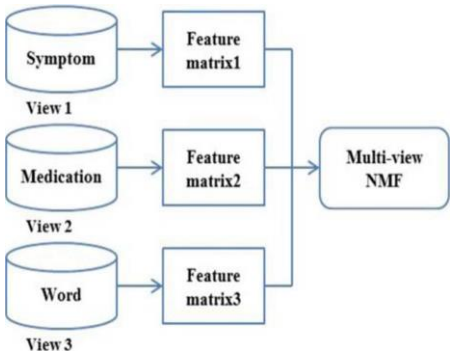

Fig. 4. The structure of Parallel-view NMF.

3.4 K-Means Clustering

Clustering[5] is a widely studied data mining technique in the area of text domain. It has diverse applications in real world scenarios. It is one of the main experimental researches in the field of data mining. Clustering of textual data is a way of directing and summarizing the content present in the document which paid heeds. In our statistical analysis we have chosen dataset which contains 100 such medical prescriptions.

After pre-processing the documents, we apply K-means Clustering on the processed documents in the following manner. K-means clustering is usually applied on numerical data which does not need considerations of other computations. But to apply K-means algorithm on the textual data which is in unstructured/semi structured format[14] is to be converted to the numerical form. This can be done by converting the documents as vectors. To do this we compute the term frequency-inverse document frequency and creation of document vectors is to be done. This will help in mapping the most frequent words in the documents and indicates the how essential the word is in corpus. These numerical data is considered in K-means algorithm.

Before applying k-means on the text documents, these documents are represented as mutually comparable vectors

using the tf-idf (term frequency-inverse document frequency) value[15]. It ranks the importance of a term in the textual document corpus. Term frequency is calculated as a normalized frequency i.e. it is a ratio of the frequency (number of occurrences) of a word in the document to the total number of words in that document.

The inverse document frequency is the log of the ratio of the number of documents in the corpus to the number of textual documents holding that term. These two metrics when multiplied together gives tf-idf value, stating the importance of a term frequent and rare in the corpus. Then tf–idf is computed as

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \times \mathrm{idf}(t, D)$$

Cosine similarity[16] is the calculation of the similarity between two documents. After converting the documents into document vectors by the previous calculation (tf-idf step) we can determine the similarity metric based on the cosine of the angle between the two document vectors. Each term in the document has its own axis. The formula given below finds the similarity between any two documents.

$$Cosine\ Similarity\ (d1, d2) = \frac{(d1, d2)}{\|d1\| * \|d2\|}$$

After obtaining the two vectors we divide them by the product of their magnitudes. The angle so calculated is a good indicator for the measure of similarity. The cosine of 0° is 1, and for any other angles it is less than 1.

After computing this, K-means clustering is performed. Initially the number of clusters (K) to be formed is decided. The value of K can be chosen as per the requirement. In our experiment we have chosen the value of K as 3. In a wide sense, K-means works by randomly initializing the k number of clusters as centroids. Here we apply an iteration of K-means on the dataset. The following method is employed here, by picking up the K objects and placing the centroid in the same place as those objects. Assignment of each object to it's closely cluster centroid. In the final step we need to update the average value of the cluster to the cluster centroid. Updating and allotment step is done is done periodically until the optimum solution is achieved hereby reducing fitting error.

4. 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[18].

TABLE I

2015 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

2016 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***

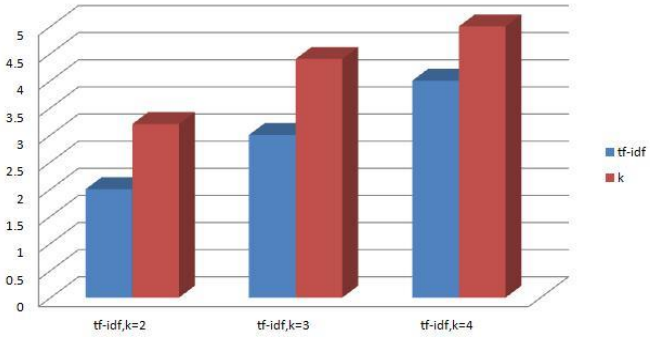

Fig. 5. Accuracy of pharmaceuticals (medications) to the related manifestations (symptoms) on the basis of cluster

During pre-processing, timer was set and processing speed for different clinical data set was recorded.

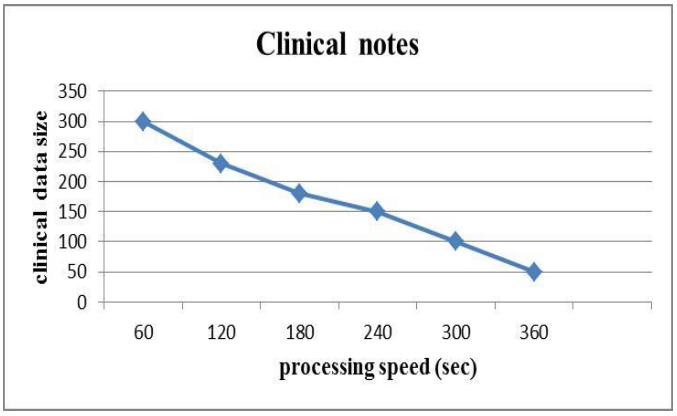

Chart -1: Processing speed with respect to clinical data set

Finally, we conclude that the processing speed increases as the number of clinical documents decreases, which is shown in Fig.3. If we have less number of clinical documents then we get less efficient results. In order to get more efficient results, we need to have more number of clinical documents.

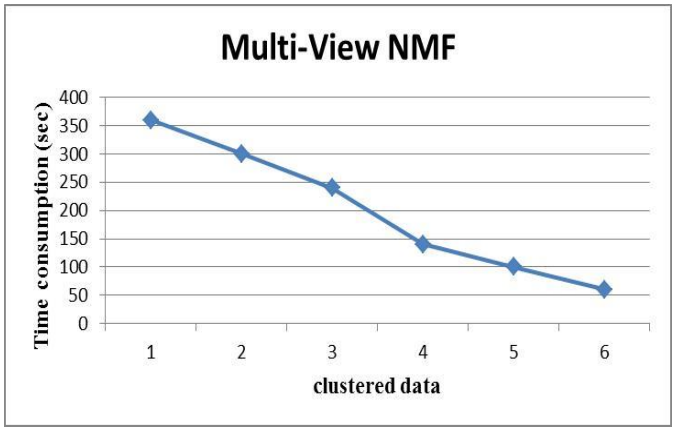

Chart -2: Time consumption with respect to clustered data[19]

4. CONCLUSION

In this paper we have built a system to know the accuracy of pharmaceutical associated with each manifestation clinical notes to build profile for independent patient. The whole system consists of 7 parts: section annotator; word/sentence annotator; negation annotator; pharmaceutical (medication) name annotator; manifestation (symptom) name annotator, Age annotator, Climate annotator.

We use the extracted manifestation/pharmaceutical names integrated with words as three-views from clinical notes, and then we have applied parallel-view (Multi-View) NMF and k-means clustering of documents. We have plan for comparing the two different dataset to compute the accuracy by applying both Parallel-View NMF and K-means NMI as evaluation metrics to compare results. It showed that, the clustering performance can be increases by using pharmaceutical names and manifestation names. It also specify that parallel-view NMF can achieve better results than K-means techniques.

REFERENCES

[1] 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.

[2] 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.

[3] Tapas Kanungo, Senior Member, IEEE, David M. Mount, Member, IEEE, Nathan S. Netanyahu, Member, IEEE, Christine D. Piatko, Ruth Silverman,And Angela Y. Wu, Senior Member, "An Efficient K-Means Clustering Algorithm: Analysis And Implementation" IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 24, No. 7, July 2002.

[4] Hung Chim And Xiaotie Deng, Senior Member, IEEE "Efficient Phrase-Based Document Similarity For Clustering", IEEE Transactions On Knowledge And Data Engineering, Vol. 20, No. 9, September 2008.

[5] Wen Zhang, Taketoshi Yoshida, Xijin Tang, "TFIDF,LSI and Multi-word in Information Retrieval and Text Categorization", IEEE International Conference on Systems, Man and Cybernetics, 2008.SMC 2008.

[6] Wen Zhang, Taketoshi Yoshida, Xijin Tang, "TFIDF,LSI and Multi-word in Information Retrieval and Text Categorization", IEEE International Conference on Systems, Man and Cybernetics, 2008.SMC 2008.

[7] HR. Aronson, "Metamap: Mapping text to the UMLS metathesaurus," pp. 1–26, 2006 [Online]. Available: http://skr.nlm.nih.gov/papers/references/metamap06.pdf

[8] Yuan Ling, Xuelian Pan, Guangrong Li*, and Xiaohua Hu, Member, IEEE, "Clinical Documents Clustering Based on Medication/Symptom Names Using Multi-View Nonnegative Matrix Factorization",IEEE Transactions On Nano bioscience, Vol. 14, No. 5, July 2015.

[9] Y. Ling, Y. An, And X. Hu, "A Matching Framework For Modelling Symptom And Medication

Relationships From Clinical Notes," In Proc. IEEE Int. Conf. IEEE Bioinformat. Biomed.(Bibm), 2014.

[10] 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

[11] 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.

[12] Ö. Uzuner, I. Solti, and E. Cadag, "Extracting medication information from clinical text,"J. Amer. Med. Informat. Assoc., vol. 17, no. 5, pp. 514–518, 2010.

[13] Shady Shehata, Member, IEEE, Fakhri Karray, Senior Member, IEEE, and Mohamed S. Kamel, Fellow,IEEE,"An Efficient Concept-Based Mining Model For Enhancing Text Clustering", IEEE Transactions On Knowledge And Data Engineering, Vol. 22, No. 10, October 2010.

[14] Manjunath varchagall and Gurubai rampure,"A Nonnegative Matrix Approximation Based Multi-View NMF Medication and Symptom Names Extraction for Clustering of Cinical Documents," in ijirset vol.5,special issue10,may 2016.

[15] Effat Naaz,Divya Sharma,D Sirisha,Venkatesan M ," Enhanced K-means Clustering Approaches for Health Care Analysis Using Clinical Documents ",ISSN-0975 1556,SCSE Department of VIT,Vellore Institute of Technology,Vellore.

[16] Ms.Ramys.S.Bhat,Mrs.A.Rafega Beham ,"Medical Records Clustering Based on the Text Fetched from Records",IJCSMC, Vol.5,Issue.5,May 2016,Pg.771-782.

[17] S.Viveka,S.Kalpana,V.Kiruthika,S.Meiyazhagan,R.Nandha Kumar",Clincal Document Clustering using Multi-view Non-Negative Matrix Factorization," SAJET Vol.2,No.17(2016)288-293.

[18] Manjunath Varchagll,Gurubai Rampure,"A Nonnegative Matrix Appromation Based Multi-View NMF Medication and Symptom Names Extraction for Clustering of Clinical Documents",IJIRSET Vol.5,Special Issue 10,May 2016.

[19] Farhat Mulla,Praksas H.Unki,"Clustering of Medical Documents Using Symptoms /Medication Names",IJARSE ISSN2319-8354,Vol.No.5,Issue No.07,July 2016.