SAJET

South Asian Journal of Engineering and Technology Vol.2, No.17 (2016) 288–293

ISSN No: 2454-9614

Clinical Document Clustering using Multi-view Non-Negative Matrix Factorization

S. Viveka, S. Kalpana*, V. Kiruthika, S. Meiyazhagan, R. Nandha Kumar

Department of Information Technology,Velalar College of Engineering and Technology Tiruchengode,Tamilnadu, India.

*Corresponding Author: S. Viveka

E-mail: shiveka@gmail.com,

Received: 10/11/2015, Revised: 14/12/2015 and Accepted: 03/03/2016

Abstract

Clinical document contains vital information like symptom names, medication names, age, gender and some demographical information. These information can be used for giving quick relief from a disease. In existing system, they had built a system for clustering symptom names and medication names using Multi-View Non-Negative Matrix Factorization. While considering the clinical documents the factors like age, gender and some demographical information become important. In this paper, we build a system for clustering the clinical documents based on age, gender and some demographical information in addition to symptom names and medication names using Multi-View Non-Negative Matrix Factorization.

Reviewed by* **ICETSET'16 *organizing committee*

Keywords: Symptom names, medication names, demographical information.

1. Introduction

Clinical documents is a rich text-free data sources contains valuable information like symptom names, medication names, age, gender and some important demographical information(like environment in which they are living, etc.,). These give more information for the doctor for improving health conditions of the patients. These information forms a large volume of raw data. Maintaining of these large volumes of data become more difficult.

Nowadays, data mining become more powerful in maintain large volume of raw data. By using data mining it is easy to maintain large volumes of data. Data mining provides many different algorithms for cluster the related data. Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters. It helps users to understand the natural grouping or structure in a data set. Clustering helps in extracting the required data from the large number of data.

Non-Negative Matrix Factorization is a group of algorithms in multivariate analysis and linear algebra where a matrix V is factorized into two matrices have no negative elements. This non-negativity makes the resulting

matrices easier to inspect. NMF generates factors with significantly reduced dimensions compared to the original matrix. For example, if V is an m × n matrix, W is an m × p matrix, and H is a p × n matrix, w is an m × p matrix, and H is a p × n matrix the p can be significantly less than both m × n.

In our proposed system, the Multi-View NMF is applied to cluster the clinical documents based on age, gender, demographical information, in addition to symptom names and medication names. This following three sections contains, related works, Existing system and Experiment Results.

2. Related Work

2.1. Overview of NMF

Recently, there has been significant development in the use of non-negative matrix factorization(NMF) methods for clustering. NMF is used to factorize the input matrix into two nonnegative matrices. The factorized matrix mostly have lower rank. Although NMF can be used for conventional data analysis, the recent overwhelming interest mining and machine learning problems. In particular , NMF with the sum of squared error cost function is equivalent to a related K-mean clustering.

Let matrix V be the product of the matrices W and H,

V=WH.

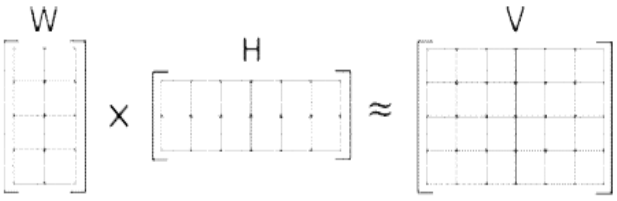

Fig.1

Matrix multiplication can be implemented as computing the column vectors of **V** as linear combinations of the column vectors in **W** using coefficients supplied by columns of **H.** That is each column of **V** can be computed as follows:

$$\mathbf{v}_i = \mathbf{W}\mathbf{h}_i ,$$

where $\mathbf{v}_i$ is the *i*-th column vector of the product matrix **V** and h is the *i*-th column vector of the matrix **H**.

2.2. Non-Negative Factorization for Clustering of Microarray Data

Non-Negative Matrix Factorization can be used for clustering microarray of data. Gene expression data are formed with the help of many number of samples. This gene expression data comprise relevant information as well as irrelevant information often interpreted as noise. This irrelevant information may reduce the efficiency of clustering. To avoid the irrelevant information from the extracted information is to provide data dimensionality reduction, in which data is decomposed into lower dimensional factors, so that those factors improves efficiency of original data.

2.3.Multi-View Nonnegative Matrix Factorization for Clothing Image Characterization

Due to the ambiguity in describing and discriminating between clothing images of different styles. It has been a challenging task to solve clothing image characterization problems. Based on the use of multiple types of visual features,, a novel multi-view nonnegative matrix factorization algorithm for solving the above task. Multi-view NMF not only observes image representations for describing clothing images in terms of visual appearances, an optimal combination of such features for each clothing image style would be learned, while the separation between different image styles can be conduct experiments on two image datasets, and confirmed that the proposed method produces satisfactory performance in terms of both clustering and categorization.

3. Existing System

In the existing system, they have built an integrated system for extracting medication names and symptom names from the large number of clinical documents. The overall system contains five parts for extracting clinical notes. They are: word/sentence annotator, section annotator, negation annotator, symptom annotator, and medication name annotator.

The word/sentence annotator is used to extract the important words and sentence from the clinical notes. The section annotator is used for the purpose of identifying different sections for clinical notes. The negation annotator is used to remove the negative words and negation symptoms and medication names from the extracted clinical notes. Now the extracted clinical document was ready for extracting symptom names and medication names. For this purpose, the symptom name annotator and medication name annotator are used. The symptom name annotator extracted symptom m=names from the clinical documents. Similarly, the medication name annotator was used for the purpose of extracting medication names from the clinical notes. Multi-View NMF has been applied on the extracted clinical notes. The different views used in the Multi-View NMF were: symptom, medication, word. .

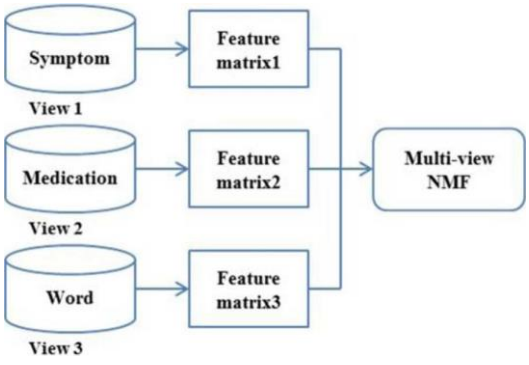

Fig.2 the framework of applying multi-view NMF

4. Proposed System

4.1. Pre-processing the Datase

Clinical document is the source of rich text which contains description about the disease, health status of the patient, and some demographical factors. Clustering of these data rich document may helps in improving the

health care. To cluster the clinical documents, the contents are to be extracted and pre-processed before clustering is done. Clustering based on multi-view may give better performance than doing in the single-view.

The overall system is dived into word/sentence annotator, section annotator, negation annotator, symptoms annotator, medication annotator, age annotator, climate annotator. The word/sentence annotator helps to extract the important words and meaningful sentence from the clinical document. It extracts the keywords from the text-rich clinical documents. The section annotator is used to identify the different sections from the extracted word/sentence from the clinical documents. The negation annotator is used for removing the negation words from the clinical notes. This gives better performance to the cluster.

The symptom name annotator and medication name annotator are used to extract the symptom names and medication names from the clinical notes. Then the age and climatic condition where the patient lives are extracted. This extracted information are used to cluster the clinical documents with the help of Multi-View NMF.

4.2. Clustering the Clinical Documents

The extracted information from the clinical documents are then clustered using Multi-View NMF. The result from both NMF and Multi-View NMF are compared and has been proved that Multi-View NMF produces better performance than NMF.

5. Result

#	Major Features	
	Symptom	Medication
1	Pain; meds (microcephaly, epilepsy, and diabetes syndrome); infections	Fluvastatin; nicardipine; methyldopa; amphotericin; thera; ammonia; hydroxyzine hcl
2	Congestive heart failure; coronary artery disease; secondaries (neoplasm metastasis); diabetes	Emtricitabine; potassium citrate; bicalutamide; mcp; dipyridamole
3	Ischaemia; nausea; congestive heart failure; symptoms	Procaine; hydroxyzine hcl; menthol; dextran 40; linezolid; clopidogrel bisulfate
4	Hypertension; obesity; asthmatics; pulmonary failure; gout; apnea, sleep apnea syndromes; mental depression; hepatitis b; diabetes mellitus; depressive disorder	-
5	Erythema; diarrhea; abdominal pain; haematocrit; obesity; wound; place (ocular myopathy with hypogonadism); vomiting	Beta blockers; emtricitabine

TABLE 1

#	Major Features	
	Symptom	Medication
1	Hyperlipidaemia; hypercholesterolaemia; polycythaemia; gerd; hypertensive disease	Aspirin; Lisinopril; furosemide; phencyclidine; metoprolol
2	Chest pain; constipation; facial hemiatrophy; pain; food-drug interactions	Heparin, porcine; digoxin; amiodarone; furosemide; warfarin
3	Place (ocular myopathy with hypogonadism); haematocrit; secondaries (neoplasm metastasis); pain; chest pain	Dextrose; insulin; metoprolol; aspirin; creatinine
4	Diabetes mellitus; glaucoma; hepatitis c; hepatitis c virus;	Prednisone; insulin, aspart, human/rdna; acetaminophen;

Table II

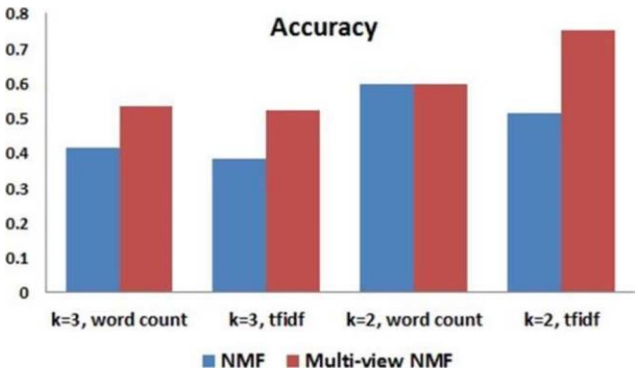

Fig 3

We choose $k=5$ to cluster documents into 5 groups. For each document clusters, the top 10 features with the highest weight are listed in Table I (NMF results) and Table IV (multi-view NMF results).

The fig 3 shows the comparison between NMF and Multi-View NMF. It is observed that the accuracy of Multi-view NMF is higher than NMF. Hence this proved that Multi-View NMF produces better result compared to NMF.

6. Conclusion

In this paper, we build an integrating system to extract symptom/medication names, age, and climatic condition of the patient from unstructured/semi I structured clinical notes. The overall system contains seven parts: word/sentence annotator; section annotator; negation annotator; symptom name annotator; medication name annotator, age annotator; climate annotator. We use the extracted symptom/medication names, climate age combined with words as five-views from clinical notes, and then we apply multi-view NMF for documents clustering.

In future, we planned to consider some other demographical information, gender of the patient for improving the cluster performance. This helps to improve the medication recommendation for the patients.

References

[1] Yuan Ling, Xuelian Pan, Guangrong Li*, and Xiaohua Hu.- Chua," Clinical Documents Clustering Based on Medication/Symptom Names Using Multi-View Nonnegative Matrix Factorization," ,2014.

[2] M.-Y. Kim *et al.*, "Patient information extraction in noisy tele-health 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 distribu-tional 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 tech-niques," in *California HealthCare Found.*, 2010 [Online]. Available: http://www.chcf.org/~/media/MEDIA%20LIBRARY%20Files/PDF/ C/PDF%20ClinicalDocumentationEHRDeploymentTechniques.pdf

[8] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-neg-ative matrix factorization," in *Proc. 26th Ann. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval ACM*, 2003.

[9] J. Liu *et al.*, "Multi-view clustering via joint nonnegative matrix fac-torization," in *Proc. SDM SIAM*, 2013.