International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 4 | ISSN : 2456-3307

Medical Papers Cluster Supported Medication/Symptom Names Using Multi Read Nonnegative Matrix Factorization

T. Gowthami
MCA Sri Padmavathi College of Computer Sciences and Technology Tiruchanoor, Tirupati, Andhra Pradesh, India

ABSTRACT

Clinical documents are rich free-text knowledge sources containing valuable medication and symptom data, that have a great potential to enhance health care. Existing system, a brand new convolutional neural network primarily based multimodal disease risk prediction algorithmic rule mistreatment structured and unstructured knowledge from hospital. To the simplest of our information, none of the existing work focused on each knowledge types within the space of medical big knowledge analytics. Compared to many typical prediction algorithms, the prediction accuracy of our proposed algorithm reaches with a convergence speed. Proposed system, we tend to build an integrating system for extracting medication names and symptom names from clinical notes. Then we apply nonnegative Matrix factorization (NMF) and multi-view NMF to cluster clinical notes into purposeful clusters supported sample-feature matrices. Our experimental results show that multi-view NMF could be a preferred methodology for clinical document cluster. Moreover, we discover that using extracted medication symptom names to cluster clinical documents outperforms simply using words.

Keywords: Nonnegative Matrix Factorization, multi-view NMF, medication symptom, clinical documents

I. INTRODUCTION

Data mining is a well-known approach for knowledge discovery in database systems. It is an efficient way of finding useful information from huge amount of data. It is a branch of artificial intelligence (AI) method, which is used to extract vital data from enormous amount of data. The knowledge that we get through this technique can be used for further innovation and collaboration. There are many applications 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. Clinical data mining is nothing but mining clinical data, so as to get essential data based on our requirement. Clinical documents contain textual data. By applying data mining technique on these data, we can fetch key information like medication names and symptom names from clinical narratives. Information extraction is important task in case of machine learning (ML) and natural language processing (NLP), as it involves significant data extraction from natural language text. Extraction of these essential data helps health care provider to advance health care system. Clinical document plays a vital role in analysis and diagnosis of disease. Mining of vital data in medical field involves, handling number of important tasks like recognition of medical related terms, recognition of attributes such as negation, severity, uncertainty and mapping words in document to concept in domain specific ontologies. The entire procedure depends on many different types of NLP processes such as tokenization, parsing, and part of speech tagging.

CSEIT1833410 | Received :15 March 2018 | Accepted :25 March 2018 | March-April-2018 [(3)4 : 438-441]

Nonnegative Matrix Factorization (NMF) has been widely applied to document clustering. Akata et al extended NMF towards joint NMF, which can jointly analyze different types of features for multi-view learning. Instead of fixing a common clustering solution for each view, Liu et al further formulated the process by finding a nearest consensus for each view. Multi-view NMF can integrate various sources of data and yield a better clustering result.

Our contributions in this paper are three-folds:

(1) we present a system for extracting symptom/medication names from clinical notes;

(2) we apply multi-view NMF to evaluate the effects of using medication/symptom names to improve the clinical documents clustering results;

(3) we compare the performances of NMF and multi-view NMF on clinical documents clustering.

Symptoms and medications are two important types of information that can be obtained from clinical notes. Symptom related information such as diseases, syndromes, signs, diagnose etc., can be used to analyze diseases for patients. In addition, valuable medication information is commonly embedded in unstructured text narratives spanning multiple sections in clinical documents. Medication information from clinical notes is often expressed with medication names and other signature information about drug administration, such as dosage, route, frequency, and duration. In this paper, we extract medication names from clinical notes. Other related medication information is also very important, and will be considered in future research. Recently, large volumes of clinical documents are generated by electronic health record systems. These clinical documents are unstructured or semi-structured. It is a difficult task to extract information from these documents. Symptom information and medication information extraction for clinical notes need sophisticated clinical language processing methods. Due to the individual diversity, it is a challenge problem to discover the underlying patterns from a corpus of clinical documents.

NON NEGATIVE MATRIX FACTORIZATION (NMF):

Basic NMF

NMF is a useful method 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 $k \times m$ matrix H, which can be expressed as the following optimization problem using the square of Euclidean distance:

$$\mathrm{Min}||A - WH||^{2}$$

The cost function can be minimized by applying the update rules as follows:

$$H_{a\mu} \leftarrow H_{a\mu} \frac{(W^T A)_{a\mu}}{(W^T W H)_{a\mu}}$$

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

Where k represents the number of clusters. It can be decided by referring from a consensus matrix and cophentic correlation coefficient.

Multi-View NMF

NMF has been extended to multi-view learning. Multi-view learning aims to identify latent components in different sub-matrices in a simultaneous manner. These sub-matrices can represent different features spaces. Akata et al extends the basic NMF to convex combination of p different views as following optimization problem:

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

Due to constraint that matrix H is fixed among multiple views, Liu et al[16] further extend to solving the following optimization problem:

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

This problem attempts to optimize $Ai \approx WiHi$ for each view i, and keep constraining each Hi will be similar.

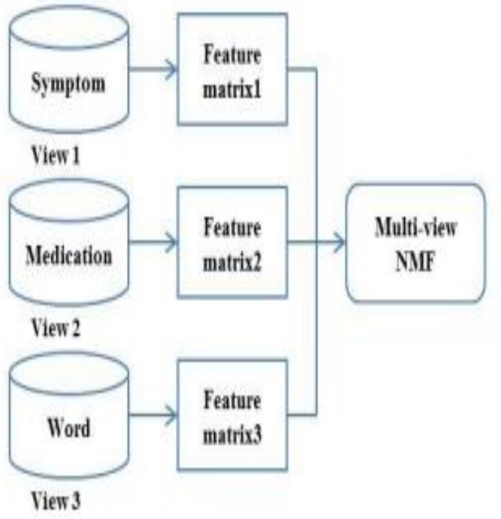

Figure 1. The Framework of Applying Multi-view NMF

II. CONCLUSION

In this paper, we build an integrating system to extract symptom/medication names from unstructured/semi-structured clinical notes. the system contains 5 parts: word/sentence annotator; section annotator; negation annotator; symptom name annotator; and medicine name annotator. we use the extracted symptom/medication names combined with words as three-views from clinical notes, so we apply multi-view NMF for documents cluster. we use two totally different datasets to match multi-view NMF with NMF. It showed that by exploitation symptom names and medicine names, the cluster performance will be improved. It additionally indicates that multi-view NMF are able to do higher results than NMF.

III. REFERENCES

[1]. J.M. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5):604-632, 1999.

[2]. C. Li, J. Han, G. He, X. Jin, Y. Sun, Y. Yu, and T. Wu. Fast computation of simrank for static and dynamic information networks. In EDBT, pages 465-476, 2010.

[3]. D. Lizorkin, P. Velikhov, M.N. Grinev, and D. Turdakov. Accuracy estimate and optimization techniques for simrank computation. PVLDB, 1(1):422-433, 2008.

[4]. P.A. McKee, W.P. Castelli, P.M. McNamara, and W.B. Kannel. The natural history of congestive heart failure: The framingham study. N Engl J Med., 285:1441-1446, 1971.

[5]. S. Meystre, G. Savova, K.K. Schuler, and J. Hurdle. Extracting information from textual documents in the electronic health record: A review of recent research. IMIA Yearbook of Medical Informatics Methods Inf Med 2008, 2008. 47 Suppl 1:128-44.

[6]. L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[7]. A. Pathak, S. Chakrabarti, and M.S. Gupta. Index design for dynamic personalized pagerank. In ICDE, pages 1489-1491, 2008.

[8]. Y. Sun, Y. Yu, and J. Han. Ranking-based clustering of heterogeneous information networks with star network schema. In John F. Elder IV, Françoise Fogelman-Soulié, Peter A. Flach, and Mohammed Javeed Zaki, editors, KDD, pages 797-806. ACM, 2009

[9]. H. Tong, C. Faloutsos, and J.Y. Pan. Random walk with restart: fast solutions and applications. Knowl. Inf. Syst., 14:327-346, March 2008.

[10]. H. Tong, S. Papadimitriou, P.S. Yu, and C. Faloutsos. Proximity tracking on time-evolving bipartite graphs. In SDM, pages 704-715, 2008.

[11]. Y. Wang. Annotating and recognising named entities in clinical notes. In ACL/AFNLP (Student Research Workshop), pages 18-26. The Association for Computer Linguistics, 2009.

[12]. H.Xu, S.P. Stenner, S. Doan, K.B. Johnson, L.R. Waitman, and J.C. Denny. Medex: a medication information extraction system for clinical narratives. Journal of American Medical Informatics Association, 17(1):19-24, Jan-Feb 2010.

[13]. Aronson, A.R., Metamap: Mapping text to the umls metathesaurus. Bethesda, MD: NLM, NIH, DHHS, 2006.

[14]. Sondhi, P., et al. SympGraph: a framework for mining clinical notes through symptom relation graphs. in Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. 2012. ACM.

[15]. Williamson, D.P., The primal-dual method for approximation algorithms. Mathematical Programming, 2002. 91(3): p. 447-478.

[16]. Mitchell, J.E., Branch-and-cut algorithms for combinatorial optimization problems. Handbook of Applied Optimization, 2002: p. 65-77.

[17]. Makhorin, A., GNU linear programming kit. Moscow Aviation Institute, Moscow, Russia, 2001. 38.

[18]. RinnooyKan, A. and J. Telgen, The complexity of linear programming. Statistica Neerlandica, 1981. 35(2): p. 91-107.

[19]. Uzuner, O., I. Solti, and E. Cadag, Extracting medication information from clinical text. Journal of the American Medical Informatics Association, 2010. 17(5): p. 514-518.

[20]. Davis, J. and M. Goadrich. The relationship between Precision-Recall and ROC curves. in Proceedings of the 23rd international conference on Machine learning. 2006. ACM.

[21]. Kim, M.-Y., et al., Patient Information Extraction in Noisy Tele-health Texts, in In the IEEE International Conference on Bioinformatics and Biomedicine (BIBM13)2013: Shanghai, China.

Author's Profile:

Ms. Thirolla Gowthami has received her graduation degree in BSc. Bachelor of Science from Sri Gnanambica Degree College, Madanapalli, Chittoor Affiliated to SV University in the year of 2012-2015 . At Present She is Pursuing Post graduate degree MCA, Master of Computer Applications from Sri Padmavathi College of Computer Sciences and Technology Affiliated to Sri Venkateswara University , Tirupati, AP, India.