

A System for Recommendation of Medication Using Gaussian Naïve Bayes Classifier

Amitha, Merin Meleet

Abstract- As the medical data keeps growing day by day, it is difficult to search for relevant information from the huge data. Improper medications may lead to serious health risks and even may result in the death of the patient. The recommender systems can be used to provide suggestions based on the health status. This approach aims to develop an efficient recommendation system which is responsible for recommending medicines for the disease based on the symptoms. This system would help the doctors in prescribing medications correctly without medication errors.

Index Terms— Clinical Notes, Extraction, Classification, Recommendation, Naive Bayes Classifier

I. INTRODUCTION

Nowadays, the World Wide Web is growing rapidly, the medication information on these websites have also increased. There are many websites that provide personalized recommendations to the user allowing choosing their topics of interest. As the medical records keep increasing, it is difficult to search through these records. Since the clinical data is unstructured, the recommender systems are essential and training should be done to generate desired results. Recommender systems are used to filter out the information based on the user's profile [11]. The recommender systems are broadly classified as content based and collaborative filtering. The collaborative filtering makes automatic predictions and filtering based on the interests of the user by collecting preferences and taste of the user by collaborating various user's choices [13]. In content based filtering, user ratings are considered. The items are matches based on the user's preferences [12]. Medication recommender systems are similar systems that give medicines as recommendation to the user for the identified disease based on the symptoms.

The clinical document is given as an input to the system. Clinical documents contain useful information about the patient's health status which can be symptoms, diseases, medicines, reason for the illness, resume of the course recommended in the hospital, injuries and procedures or test undergone. This valuable information is used for building the profile of each patient. Medications and symptoms related information need to be extracted from these clinical notes. Some of the symptom related concepts most commonly found in the clinical notes are signs, diseases, syndrome and laboratory or test results. Medication names in the clinical notes are often associated with some valuable information which can be dosage level, strength, necessity, intake time, route, duration and frequency. It is difficult to parse the sentences that contain medications. Sometimes, one sentence would contain multiple medication names. Understanding this information and processing these becomes difficult. Hence there is need for the natural language processing techniques to process this information. MedEx is one of the natural language processing systems used to extract medication names and other signature information from these discharge summaries. As the clinical documents are unstructured, it is difficult to identify the symptom names and medications... Meta Map is used here to identify all the concepts from which the symptoms are extracted. It uses Unified Medical Language System (UMLS) and it consists of met thesaurus, Semantic Network and the lexicon tools [14]. The Gaussian Naïve Bayes classifier is used for training where the medications corresponding the disease is identified and is given as a suggestion to the user. The electronic medical record (EHR) contains patient's medical history and can be accessed only by the authorized users [15]. It is used by the doctors in taking decisions which could be the diseases recognized or medications to be provided for the patient. A large volume of medical reports are generated from EHR systems [16]. This system can be used to exchange of health information electronically in patients health care. Extracting medical information from clinical notes becomes difficult when it is unstructured. Natural language processing methods can be used to extract symptoms and medication information. Medication recommender system extracts medication and symptoms from the clinical document, predicts the disease associated with the symptoms and suggests the medication names as recommendation to the users.

Manuscript received May 22, 2019

Amitha, Department of Information Science and Engineering, R V College of Engineering, Bangalore, India
(e-mail:amitha.1933@gmail.com)

Merin Meleet, Department of Information Science and Engineering, R V College of Engineering, Bangalore, India

II. LITERATURE SURVEY

Ayrine John et al., [1] proposed a unified extraction system to extract medical information from clinical notes. In this approach, K- Means clustering is used to cluster the diseases. A collaborative filtering method is used to find out the desired medication. The four major steps involved in building the medication recommendation system are:

- i) Analyzing the clinical document
- ii) Retrieving clinical terminologies
- iii) Clustering of similar medication names and symptoms names
- iv) Recommendation of medicines to the user.

This approach can also be used as a tool by the doctors for the diagnosis of a disease.

Yuan Ling et al., [2] explained an approach for extracting medical terms using Natural Language Processing (NLP) techniques. Once the medical information is extracted, the clinical documents are clustered using Nonnegative Matrix Factorization (NMF) and Multi-View NMF. This approach is categorized into five major parts: word or sentence tokenization, section header identification, negation sentence identification, symptoms and medication names extraction. Here, NMF and Multi-View NMF techniques are applied in order to cluster the documents. The results obtained from these techniques prove that the accuracy of Multi- View NMF is more than NMF.

Hui Yang [3] projected an approach for extracting medicine names from the clinical notes. The dataset used here is i2b2's2009 Medication Extraction Challenge. The terms and tokens are matched to identify medicine information in the discharge summaries. The rule-based method can be used to identify numerous entities like medications. Some annotated documents are used for training a model using machine learning.

Scott Hal grim et al., [4] discussed method which uses machine learning for training a model and modules that contain set of rules for extracting medication names. The set of classifiers were cascaded together for detecting various fields and achieved better performance. When the results were compared with the other systems of i2b2's challenge, the performance of statistical approach is better than systems with many sophisticated rules.

Hua Xu et al., [5] proposed MedEx system which can be used in extracting medicine information in the discharge summaries. Discharge summaries usually contain medication data as free text. This system performed well in identifying drug names as well as other information which can be dosage amount, form, strength, intake amount, duration, necessity and frequency.

Alan R Aronson [6] developed a program called MetaMap which is customizable and can be used for mapping each word to the UMLS Meth thesaurus. It is one of the biggest lexicons in biomedical field. Some of applications that make use of MetaMap program are decision support systems and for managing the patient records. MetaMap uses various approaches such as natural language processing techniques and linguistic approaches.

Thierry Hamon and Natalia Grabar [7] presented a linguistic approach where drug names and other associated information can be extracted from the discharge summaries. A module was proposed to identify the new drug names from the various semantic forms. The results of this system for drug information extraction performed well where drug names, dchioma.awodiji@futo.edu.ng dosage level and frequency were clearly extracted. But duration and reasons were not extracted correctly.

Sharique Hasan et al., [8] proposed various techniques to automatically detect omissions in medication lists i.e. identifying the medications that the patients are taking which are not present in the patient's medication lists. The collaborative filtering approach can be used to provide solutions by including nearest neighbor and co-occurrence approaches. The results have been proven that this framework is a valuable tool for medication reconciliation.

Olga Patterson and John F Hurdle [9] built a feature space which can be used for both in patients and out patients which uses vocabulary and semantic resources. The main aim of clustering the documents is to trace the group of patterns from the huge set of unlabeled data. Later put these documents together based on the similarity. The experimental results have proven that unsupervised clustering of clinical notes result in clearly differentiated document clusters.

Miss Shreya B.Ahire, Ms. Harmeet Kaur Khanuja [10] proposed a framework for healthcare recommendation. It makes use of semantic web technology for examining the user's preferences. With this data, the user's profile about the health is built and this profile is used to group the associated knowledge which helps the users in inquiring about the food and exercise. For extracting the information from the database, decision tree algorithm is used here.

III. METHODOLOGY

The medication recommendation system accepts clinical notes as input. These clinical notes contain the health related information about the patient such as symptoms, laboratory test results, injuries, diagnosis and drugs. From this raw data, the symptoms and medications are extracted and classification is done based on the diseases after which the desired medication list is suggested to the user. The major steps for building this system are

- Pre-processing of clinical documents
- Extraction of symptoms ,diseases and medications
- Classification and recommendation of medications

The architecture diagram of the medication recommendation system is shown in the figure 1.

A. Pre-processing

In this work, the dataset is obtained from 2009 i2b2 Medication challenge. In the preprocessing step, word and sentence tokenization, stop words removal, the section identification is done to identify different sections in the clinical note. The most commonly used stop words in the clinical notes are 'the', 'is' and 'are'.

The sections that provide symptoms related information are diagnosis after admission, reason for illness, previous health history, course suggested in the hospital, resume of course recommended in the hospital, medications suggested at the time of discharge, physical examination, course suggested in the hospital by system and problem.

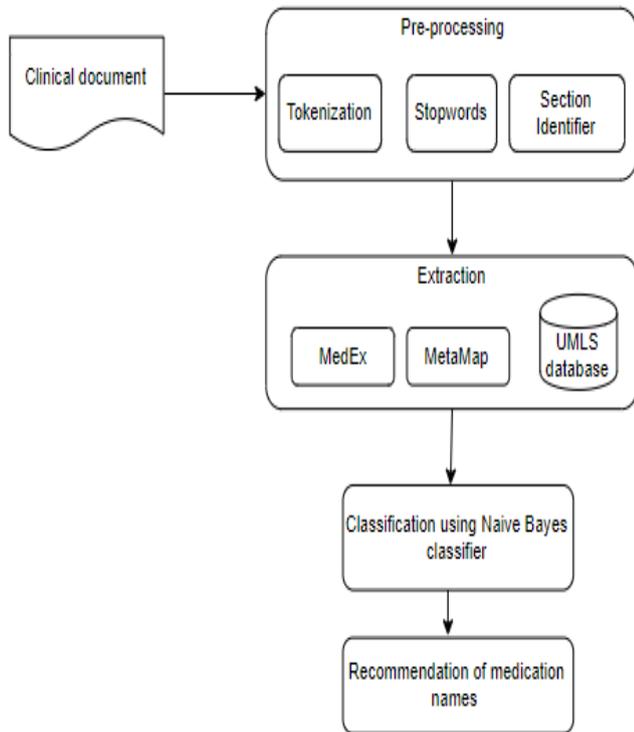


Fig. 1: Architecture of medication recommendation system

The sections that contain medication names are medications prescribed by the doctor at the time of discharge, hospital course, reason for illness, potentially serious interaction, medications on admission, resume of course recommended in the hospital, medications, medications on discharge, course suggested in the hospital by system and problem. Some of the sections that contain both symptoms and medication names are medications prescribed by the doctor at the time of discharge, reason for illness, hospital course, resume of course recommended in the hospital, course suggested in the hospital by system and problem. These symptoms related sections and medications related sections are passed to the extraction module which is built using natural language processing tools.

B. Extraction

After pre-processing, the symptom annotator is used to identify the various concepts. MetaMap is used for this purpose which maps text with the UMLS Met thesaurus. The Met thesaurus is a database text which provides information about various concepts including medical concepts. Hence MetaMap returns various types of concepts. Here only the symptom related concepts such as sign or symptom, disease or syndrome, neoplastic process, fungus, bacterium, virus, congenital abnormality, acquired abnormality, laboratory or test result, injury or poisoning, mental or behavioral dysfunction, cell or molecular dysfunction, anatomical abnormality are extracted. Extraction of drug names from the clinical documents is done through the MedEx system. It is a natural language processing system used for extracting medication information from the discharge summaries. It not only extracts medication names but also medication information such as dosage amount frequency and duration. Here drug names are treated as medication names.

C. Classification and recommendation

Medicine details for various diseases like myocardial infarction, coronary artery disease, incision hernia, and dementia are trained and are done using Gaussian Naive Bayes classifier. Diseases dictionary is created and is compared with the training data. Each entry in diseases dictionary is compared with the generated result and matching ratio is calculated. Most matching disease is identified and the corresponding medications are finalized as the result. For this identified disease, medicines are suggested to users.

The steps involved in the Naïve Bayes classifier are

Input:

Training dataset T,

$F=(f_1, f_2, f_3, \dots, f_n)$ // value of predictor variable

Output:

A class of testing dataset.

Steps:

1. Read the training dataset T;
2. Calculate the mean and standard deviation of predictor variables in each class;
3. Repeat

Calculate the probability of f_i using the gaussian density equation in each class;

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$$

Until the probability of all predictor variables ($f_1, f_2, f_3, \dots, f_n$) is calculated.

4. Calculate the likelihood for each class;
5. Get the greatest likelihood;

IV. EXPERIMENTAL RESULTS

Experiment is done on various transcriptions. Confusion matrix is used to measure the performance of a classifier based on the test data. The terms used in representing the confusion matrix are True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN). The two important metrics for evaluating an algorithm performance are precision and recall. On the basis of analysis done on 20 diseases, retrieved medications for 15 diseases are relevant. Hence the accuracy of the system is 75%. The Figure 1 shows the result of the medication recommendation system. This system gives disease name and the medication details as output. The precision, recall and accuracy is calculated as follows:

Precision= $TP/(TP+FP)$

Recall= $TP/(TP+FN)$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$



Fig. 2: Medications list for the disease

V. SUMMARY AND CONCLUSION

In this work, an efficient medication recommendation system is built for suggesting medications for various diseases using Gaussian Naïve Bayes Classifier. This system can be used for assessing the doctors for providing medications for the diseases identified based on the symptoms. This system also gives dosage amount and the duration information along with the medication names as the output. The medicines allergic to the patient are also considered for recommendation. This is the limitation of this approach. In the future work, the accuracy of the system can be increased by increasing the training data. Patient’s age, gender and other demographic information can be added to increase the efficiency in the training phase.

REFERENCES

[1] Ayrine John, MuhammedIyas H, VeenaVasudevan, “Medication Recommendation System Based on Clinical Documents”, IEEE International Conference of Information Science (ICIS) 2016, 180-184.

[2] Yuan Ling, Xuelian Pan, Guangrong Li, Xiaohua Hu, “Clinical Document Clustering Based on Medication/Symptom names using Multi-view Non-negative Matrix Factorization”, IEEE Transactions on NanoBioScience, 2015.

[3] Hui Yang, “Automatic extraction of medication information from medical discharge summaries”, JAMIA symposium proceedings 2010, 545-548.

[4] Scott Halgrim, Fei Xia, Imre Solti, EithonCadagOzlemUzuner, “Extracting Medication Information from Discharge Summaries”NAACL HLT Second Louhi Workshop on Text and Data Mining of Health Documents, 2010, pages 61–67.

[5] Hua Xu, Shane P Stenner, Son Doan, Kevin B Johnson, Lemuel R Waitman, Joshua C Denny, “Medex: a medication information extraction system for clinical narratives”, JAMIA symposium proceedings, 2009.

[6] Alan R Aronson, “Effective mapping of bio-medical text to the UMLS Metathesaurus: The MetaMap program”, In Proc AMIA Symp. 2001, 17-21.

[7] Thierry Hamon, Natalia Grabar, “Linguistic approach for identification of medication names and related information in clinical narratives”, JAMIA symposium proceedings, 2010, 549-554.

[8] Sharique Hasan, George T Duncan, Daniel B Neill, RemaPadman, “Towards a collaborative filtering approach to medication reconciliation”, JAMIA symposium proceedings, 2008, pages 288-292.

[9] Olga Patterson, John F Hurdle, “Document clustering of clinical narratives: a systematic study of clinical sublanguages”, JAMIA symposium proceedings, 2011.

[10] Miss Shreya B.Ahire, Ms. Harmeet Kaur Khanuja, “A Personalized Framework for HealthCare Recommendation”, International Conference on Computing Communication Control and Automation 2015, 442-445.

[11] Martin Wiesner, Daniel Pfeifer, “Adapting Recommender Systems to the Requirements of Personal Health Record Systems”, ACM 978-1-4503-0030-8/10/11, November 2010.

[12] <http://recommender-systems.org/content-based-filtering/>

[13] <http://recommender-systems.org/collaborative-filtering/>

[14] UMLS Reference Manual Bethesda (MD): National Library of Medicine (US)- 2018AB Release Information <https://www.nlm.nih.gov/research/umls/>

[15] Roberts, K. and S.M. Harabagiu, “A flexible framework for deriving assertions from electronic medical records”. Journal of the American Medical Informatics Association, 2011. 18(5): p. 568-573.

[16] Roque, F.S., et al., “Using electronic patient records to discover disease correlations and stratify patient cohorts”. PLoS computational biology, 2011. 7(8): p. e1002141.