Healthcare Data Prediction System using Collaborative Filtering - Machine Learning Technique

Heerah D[1], Dr. G T Raju[2]

[1] *Computer Science and Engineering, RNS Institute of Technology, Channasandra, Bangalore, India*

[2] *Professor, Head, Computer Science and Engineering, RNS Institute of Technology, Bangalore.*

Abstract-Healthcare data is very rich and it includes a record of services received, conditions of those services, and clinical outcomes or information concerning those services. With the development of incredible machine learning systems, it is presently possible to get more noteworthy bits of knowledge from the accessible information. Machine learning is when a computer has been taught to recognise patterns by providing it with data and an algorithm to help understand that data. The prediction system designed is to suggest medicines/drugs prescribed for the particular disease by training the patient's history medical data using collaborative filtering technique. The predictive results also includes the doctor suggestions based on the user rating information and the system also allows the user to query for drugs that satisfy a set of conditions based on side effects and symptoms.

Keywords—Collaborative Filtering (CL), Support Vector Machine (SVM), Machine Learning Algorithms, Prediction system

I. INTRODUCTION

Today electronic health record is growing fast and is becoming the most powerful tool in the medical toolkit. All the information has to be stored in the cloud because the size of the electronic file containing the complete patient record is estimated to be as much as six terabytes since nowadays the advanced patient monitoring systems collect the data every second that forms huge amount of dataset. A data file that large has to undergo technical processing in order to get confined outcome for the huge data collected which enables the practice of quality healthcare. A doctor alone would not be able to process this huge medical information of a patient. It will require high-powered computing, using insights from machine learning – a type of artificial intelligence that enables computers to find hidden insights without being programmed. Algorithms will interrogate with the vast data sets and surface recommended treatment plans tailored to individuals. There are two things required for the successful application of machine learning in healthcare is the intelligent algorithms and the rich data sets.

Machine learning programs take the entire knowledge that a physician has beside his experience in treating patients. Well, the idea behind artificial intelligence in the pharma industry is not to substitute doctors but to upgrade their medical expertise. Moreover, from the entire information related to diseases and its medication, the doctors have a generous amount of data available to them. When a person is affected by an infection or a disease, generally the patient will consult doctor and the doctor will prescribe some tablets. Generally the patient goes with doctor's opinion and takes any drug which is prescribed by the doctor. Mostly people will not think to get the second opinion. Example for diseases like fever, cold, headache people do not need to take second opinion. But for some disease like hormonal deficiency, chronic disorders, and other serious diseases, it is always advisable to take a second opinion from a well known physician. The type of tablets for a particular disease would follow the same pattern for any individual. The medicinal records of individual patient are recorded and these archive records are used as trained data to predict the drugs for the patient who is suffering from the same disease.

Medical data is growing huge that each and every activity of the patient is being recorded and this patient history can be used for the further prediction of the drugs for the patients whose is suffering from the similar disease. The Patient's medical information, their symptoms, disease diagnosis records, prescribed medicines and other related details are being recorded. The attributes study is processed in order to group the patients based on their illness, symptoms, diagnosis information and drugs prescribed. When a new patient is added to the database, the patient's medical history is analyzed and that patient is categorized to the already existing groups. These attributes will be compared with the new patient's symptoms whose drug has to be suggested by identifying the disease suffered by patient that patient. Training data sets are created to identify the disease by using supervised machine learning with the help of Support Vector Machine algorithm. Once disease is diagnosed, the list of medicine available for that disease is analyzed. Later the patient's profile in which these drugs are prescribed is analyzed. Best drug for the particular illness is identified and the drug is arranged based on top priority whose feedback has less side effects. Side effects of the drugs are also kept in track in order to suggest best results for the user.

II. LITERATURE SURVEY

Machine learning algorithms are being used nowadays in recommender systems for providing better recommendations. Collaborative filtering is used in recommender systems which consider user data when processing information for the recommendation. Chronic diseases such as heart disease, breast cancer, and diabetes are on the rise and there is a growing need for a system that helps diagnose these diseases at an early stage. Machine learning and neural networks have recently gained traction as reliable and robust prediction systems that help provide a deeper insight into the statistical data.Using data from the past, it is possible to design a machine learning system that helps doctors diagnose diseases with a higher level of accuracy [9]. The proposed system makes use of collaborative filtering for suggesting symptoms to the patient and machine learning algorithms like support vector machine for disease prediction.In order to give users an interactive and easy to use interface, the web platform can be used. Early diagnosis and identification of diseases play a vital role in the field of medicine. With the emergence of powerful machine learning techniques, it is now possible to derive greater insights from the available data. The proposed system allows users to enter symptoms and uses machine learning techniques to recommend similar symptoms. Another machine learning technique for classification is discussed which is used predict the possibility of having a disease. The results demonstrate the effectiveness of different machine learning techniques on the given data [2].

Essentially, rather than just presenting a directory of family doctors in alphabetical order for patients to choose from, the new primary care plan will empower patients to take an active role in selecting their own family doctors. The matchmaking process requires the healthcare network to learn about the healthcare preferences of each patient and to generate personalized recommendations accordingly. However, the healthcare network still needs to address several concerns before fully implementing and deploying the matchmaking mechanism as one of the key features in its digital health service. As patients have different levels of engagement with the healthcare network, information availability varies significantly across individuals. Firstly, a majority of patients have never before consulted with family doctors, since primary care is not mandatory before accessing other specialized services. It would be challenging for the network to learn about their preferences without data about past interactions. Secondly, patients who have had previous consultations with family doctors but want to change from their current doctor may be interested in knowing about the preferences of other patients who have visited the same family doctor. For example, patients often find a conflicting schedule with their current family doctor and may benefit from knowing about family doctors that have been visited by other similar patients. Finally, specific groups of patients, such as those with chronic illnesses, require special care and may benefit significantly from personalized primary care [5].

III. ARCHITECTURE

The drug prediction system is designed using the latest Machine learning algorithm called the collaborative filtering which is used in various field of marketing to predict the user preferences. The prediction system recommends the drugs and best doctors for the patients using the previous patient history data. The datasets are trained using classification algorithm known as support vector machine and collaborative filtering technique to analyse the disease suffered by the group of patients having the similar illness symptoms and clinical information. These trained data are used to further predict the disease for the individual patient whose attributes are compared with the similar features of the latter dataset and the corresponding drugs suggestion is analyzed for that particular disease.

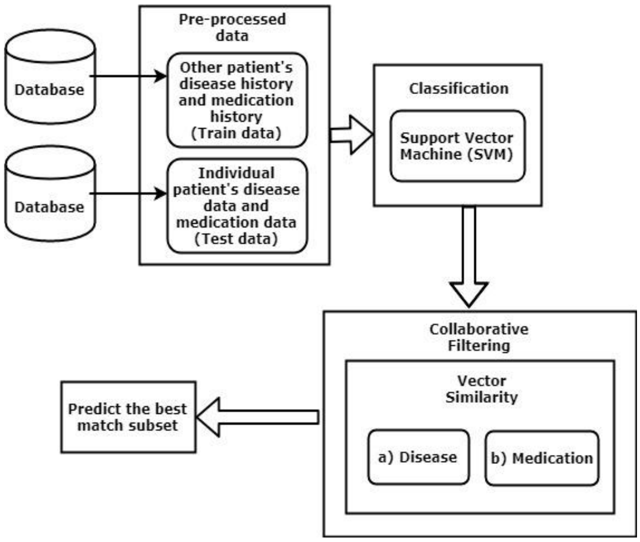

Pre-processed data
Database
Other patient's disease history and medication history (Train data)
Database
Individual patient's disease data and medication data (Test data)
Classification
Support Vector Machine (SVM)
Collaborative Filtering
Vector Similarity
a) Disease
b) Medication
Predict the best match subset

Figure 1 – Architecture diagram

Input Data

The dataset are collected from the medical cloud data service provider in order to train the data for predicting the target value. Initially the cleaning process is carried out by removing out-of-scope episodes and deriving consistent patient and doctor IDs across all hospitals. Along with the patient and doctor information the diagnosis details are also collected which is the important source for this prediction system. Diagnosis details include the patient's medical history with attributes like the diseases with corresponding symptoms and drugs prescribed for that particular disease.

Prediction Process

Pre-processed trained data from the previous step is undergone a classification phase where each data is assigned a numerical value using the Support Vector Machine which is computationally effective binary classification algorithm. The goal of SVM is to find an optimal separating hyperplane which maximizes the margin for training dataset. SVM model gets trained quickly as compared to artificial neural networks and has pretty good accuracy for small and medium-sized datasets. SVM algorithm is supported with different kernels for finding similarity between two feature vectors.

Straightforward calculations are utilized by recommendation systems with an expectation to give the most exact and pertinent things to the client with assistance of filtering information that is helpful from a substantial pool of data base. They find designs in the information present in the dataset by learning the preferences of user and show results that co-identifies with their requirements. Collaborative filtering is then used to channel the patient records which have a place just with this class. The user enters a symptom and patient records of the user class having this symptom are utilized to discover different symptoms shown by them. These related symptoms are recommended to the user. The user at that point may choose at least one recommended symptoms dependent on which progressively related symptoms are proposed. This helps the user to search for symptoms endured by patients with comparable conditions. Further drugs are suggested for the particular disease that is analyzed with the help symptoms.

Result Data

Predicted outcome of best match subset can be referred either by the doctors or by the patients for a second opinion on what exactly the doctor has suggested to the patient. Prediction process not only suggests the drug analysis for the user, it also further suggests the best doctor details for the particular disease diagnosis using the patient feedback. Also the resultant drug list is prioritized with fewer side effects causing medicines at the top followed by relative one.

IV. WORK FLOW

In the healthcare domain, applications of recommender systems include assisting the decision-making process in the disease analysis, identifying the symptoms, predicting the appropriate drug for the particular disease, supporting patients to find preventative healthcare help by suggesting doctors based on user feedback, and providing personalized healthcare guidance. With the help of machine learning algorithms the prediction analysis has made a benchmark in many fields. In accordance with the drug prediction to the user based on the patient's medical history, collaborative filtering technique is being used. To classify the attributes of the medical datasets like symptoms of the patient, disease analysis, drugs prescribed, side effects feedback by the patients, and other related attributes are categorized and the individual patients with similar attributes are predicted with the appropriate drug for a particular disease.

a. Support Vector Machine

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N—the number of features) that distinctly classifies the data points. To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplane depends upon the number of features. If the number of input features is 2, then the hyperplane is just a line. If the number of input features is 3, then the hyperplane becomes a two-dimensional plane. It becomes difficult to imagine when the number of features exceeds 3. Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. By using these support vectors, the margin of the classifier can be maximized.

b. Collaborative Filtering

Collaborative filtering (CF) is one of the most popular techniques for building recommender systems. It is a method of making automatic predictions about the requirement of a user by collecting preferences or related information from many users (collaborating). Collective filtering technology guesses that the user's preferences won't change over time. There are two main types of CF algorithms: memory user-based collaborative filtering and item-based collaborative filtering. Very often, prediction accuracy can be improved by combining them into a single model. The difference is that user-based collaborative filtering tries to find the collections of similar users and recommend the historical resources of the target user's similar users, whereas the item-based collaborative filtering tries to find the collections of similar resources and recommend the resources which are similar to the target user's resource history. The CF technology can find out the new resources which are similar to the target user's resource history. It means that CF technology can detect the potential resource that the target user may be interested in.

User-based CF can be implemented using the formulas that shows how to calculate relationship r_(d,m), the prediction about how doctor suggests the patient with the particular drug 'm' for the disease'd'. By aggregating the over all drugs 'm' prescribed for the patients with similar diseases'd' (the set of patients with similar disease is marked with S; the similarity function is marked with sim). The more similar a disease is the more influence the drug analysis has to be with the overall prediction. The value of w is the weighting factor used to scale the sum down to a single drug prediction analysis.

$$r_{d,m} = w \sum_{d' \in S} sim(d,d') r_{d',m} \dots\dots\dots\dots\dots\dots..(1)$$

$$w = \frac{1}{\sum_{d' \in S} |sim(d,d')|} \dots\dots\dots\dots\dots\dots\dots.(2)$$

c. Vector Similarity

Because both user-based collaborating filtering and item-based collaborating filtering need to measure the similarity, the three most popular similarity measure methods are as follows. The user-item score matrix R is an m × n matrix, which means there are m users and n items. Ru, c means that the user u gives the item c a score Ru, c.

In Cosine similarity, the user's scores of the items are considered as an n-dimensional vector, and the cosine angle between the users' score vectors represents the similarity between the users. The cosine similarity formula is as follows:

$$\mathrm{sim}(\mathrm{u_i}, \mathrm{u_j}) = \frac{\sum_{c=1}^{n} R_{i,c} \cdot R_{j,c}}{\sqrt{\sum_{c=1}^{n} R_{i,c}^2}\sqrt{\sum_{c=1}^{n} R_{j,c}^2}} \dots\dots\dots..\dots\dots\dots\dots.(3)$$

In Modified cosine similarity, the cosine similarity has a shortcoming that the user may not score the item, which will result in errors in the results. To make up for this, modified cosine similarity is proposed. The user ui and uj have scored two item sets Ii and Ij, respectively. The items in both Ii and Ij form a set Iij. The modified cosine similarity formula is as follows:

$$\mathrm{sim}(\mathrm{u_i}, \mathrm{u_j}) = \frac{\sum_{c \in I_{ij}} (R_{i,c} - \bar{R}_i)(R_{j,c} - \bar{R}_j)}{\sqrt{\sum_{c \in I_i} (R_{i,c} - \bar{R}_i)^2}\sqrt{\sum_{c \in I_j} (R_{j,c} - \bar{R}_j)^2}} \dots\dots..(4)$$

In Pearson correlation coefficient, the set Iij contains items that both ui and uj have scored. The Pearson's correlation coefficient formula is as follows:

$$\mathrm{sim}(\mathrm{u_i}, \mathrm{u_j}) = \frac{\sum_{c \in I_{ij}} (R_{i,c} - \bar{R}_i)(R_{j,c} - \bar{R}_j)}{\sqrt{\sum_{c \in I_{ij}} (R_{i,c} - \bar{R}_i)^2}\sqrt{\sum_{c \in I_{ij}} (R_{j,c} - \bar{R}_j)^2}} \dots\dots\dots(5)$$

For the prediction system of suggesting drugs and doctors with respect to the disease analyzed by the symptoms detail given by the user, cosine similarity technique is used in order predict the similar drug used by the patients having similar disease and symptoms. K nearest neighbour query technique is used to find two outcomes. First, use the similarity measure methods to measure the similarities between target patient and other patient. Second, select k patient with the highest k similarity degrees to be the neighbors of the target patient. Finally, form a neighbour set with the k selected patient in non-incremental order of the similarity degrees. The final outcome would the list of drug prediction for a particular disease of the target patient.

V. IMPLEMENTATION AND RESULTS

Step 1 - The medications and illness data will be stored in the Database of the Hospital environment. In order to build this system manually and not in real time, the datasets are taken from the kaggle data resource center. For the real time application, the healthcare data can be accessed from the cloud where the data is getting updataed every second. The datasets collected are huge in number where the preprocessing and data handling is done with the help of pandas python package. The data is accessed from the file and the requested job is redirected back to the prediction system where the results are stored again for the future use.

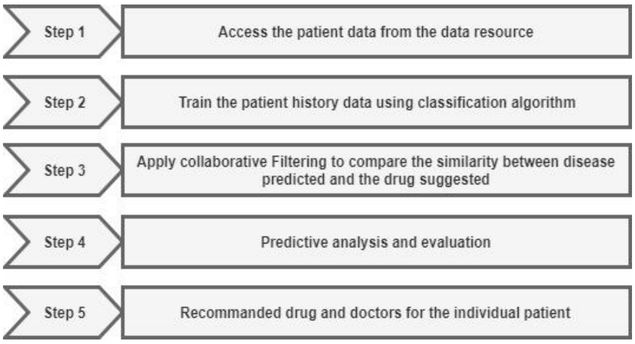

Step 1
Access the patient data from the data resource
Step 2
Train the patient history data using classification algorithm
Step 3
Apply collaborative Filtering to compare the similarity between disease predicted and the drug suggested
Step 4
Predictive analysis and evaluation
Step 5
Recommanded drug and doctors for the individual patient

Figure 2 – Flow Diagram

Step 2 – Patient history data accessed from the database is pre-processed and trained to form several attributes category. These attribute category are assigned with a numerical vector value that is used to compare and find the similarity of test data for the recommendation of medical drug for the predicted disease.

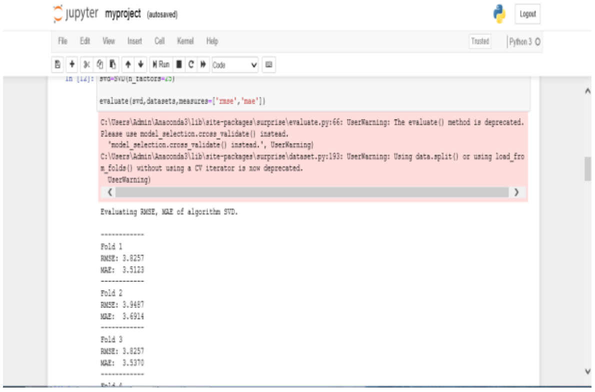
Step 3 – Collaborating filter technique is applied with the user-based collaborative filtering to find the collections of similar users and recommend the historical resources of the target user's similar users. As the comparison with the number of train data increases the accuracy of the prediction is high.

Step 4 - The prediction system allows a user to query for drugs that satisfy a set of conditions based on drug properties, such as drug indications, side effects, and drug interactions, and also takes into account patient profiles.

Step 5 - Automatic suggestion of the alternative drugs are also included. Predicted result is further examined by the doctors for to decide the dosage and other details.

The implementation is processed on Jupyter platform using python. The important packages used to implement machine learning algorithm is Surprise toolkit which is a Python scikit building and analyzing recommender systems. This toolkit provides various ready-to-use prediction algorithms which includes baseline algorithms, neighborhood methods, matrix factorization-based such as SVD (Singular value decomposition), PMF (Probabilistic matrix factorization), SVD++, NMF (Non-Negative Matrix factorization) etc. Also, various similarity measures (cosine, MSD, pearson) are built-in.

Few snippets of results are as follows:



jupyter myproject (autosaved)
Logout
File Edit View Insert Cell Kernel Help
Trusted
Python 3
Run
Code

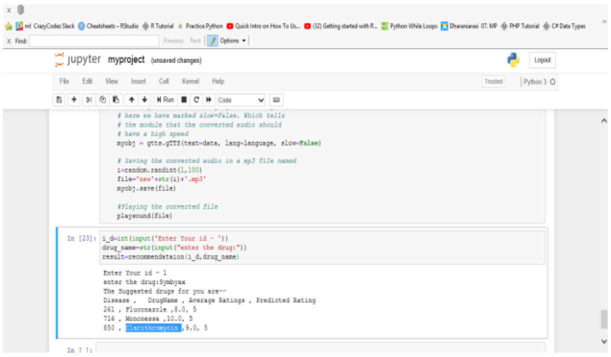
```
evaluate(svd,datasets,measures=['rmse','mae'])
```

```
C:\Users\Admin\Anaconda3\lib\site-packages\surprise\evaluate.py:66: UserWarning: The evaluate() method is deprecated.
Please use model_selection.cross_validate() instead.
  'model_selection.cross_validate() instead.', UserWarning)
C:\Users\Admin\Anaconda3\lib\site-packages\surprise\dataset.py:193: UserWarning: Using data.split() or using load_fro
m_folds() without using a CV iterator is now deprecated.
  UserWarning)
```

```
Evaluating RMSE, MAE of algorithm SVD.

------------
Fold 1
RMSE: 3.8257
MAE:  3.5123
------------
Fold 2
RMSE: 3.9487
MAE:  3.6914
------------
Fold 3
RMSE: 3.8257
MAE:  3.5370
------------
```

Figure – 3 Snippet 1



```
    # here we have marked slow=False. Which tells
    # the module that the converted audio should
    # have a high speed
    myobj = gtts.gTTS(text=data, lang=language, slow=False)

    # Saving the converted audio in a mp3 file named
    i=random.randint(1,100)
    file='new'+str(i)+'.mp3'
    myobj.save(file)

    #Playing the converted file
    playsound(file)
```

```
In [23]: i_d=int(input('Enter Your id - '))
         drug_name=str(input("enter the drug:"))
         result=recommendataion(i_d,drug_name)
```

```
Enter Your id - 1
enter the drug:Symbyax
The Suggested drugs for you are--
Disease ,   DrugName , Average Ratings , Predicted Rating
261 , Fluconazole ,8.0, 5
716 , Mononessa ,10.0, 5
850 , Clarithromycin ,9.0, 5
```

```
In [ ]:
```

Figure – 4 Snippet 2

VI. CONCLUSION

Healthcare Data Prediction System using Collaborative Filtering a Machine Learning Technique is the application where the patient's electronic health records are maintained and managed. The health records are used to make the machine learn the meaning of various medical attributes in order to give suggestion to the user based on their requirement. This prediction system group the patients with respect to the illness they are suffering from through the disease symptoms attribute further prescribed drugs and doctor details are taken as the target attribute. Therefore when a user request for the information, new user data are compared with the trained medical data of the prediction system and desired results are provided to the user.

Further the system can be enhanced by applying this prediction technique to analyse the medical data for complicated disease diagnosis. Machine can be trained with the knowledge of the medical data which is used to predict the disease and the corresponding diagnosis. Advanced medical data like X-ray images, mammograms, MRIs, ECG data etc. can be trained in order predict the disease that the patient is suffering from.

ACKNOWLEDGEMENT

I would like to thank Dr. G. T. Raju (Dept of CSE, RNSIT, Bangalore) for his valuable suggestions and comments that helped improving this works, this support is greatly appreciated.

REFERENCES

[1] Zheyun Zhong ; Yinsheng Li, "A Recommender System for Healthcare Based on Human-Centric Modeling", 2016 IEEE 13th International Conference on e-Business Engineering (ICEBE)

[2] Akshay Kamath, Amogh Parab, Neeraj Kerkar, "Symptom Recommendation using Collaborative Filtering and Disease Prediction using Support Vector Machine", International Journal of Computer Applications (0975 – 8887) Volume 179 – No.41, May 2018

[3] Ivens Portugal, Paulo Alencar, Donald Cowan, "Requirements Engineering for General Recommender Systems", Natural Sciences and Engineering Research Council of Canada (NSERC), the Ontario Research Fund of the Ontario Ministry of Research and Innovation, SAP, and the Centre for Community Mapping (COMAP)

[4] Sathish.S, Usha Nandhini, "Disease predictive, best drug: big data implementation of drug query with disease prediction, side effects & feedback analysis", Global Journal of Pure and Applied Mathematics. ISSN 0973-1768 Volume 13, Number 6 (2017), pp. 2579-2587

[5] Qiwei Han , Mengxin Ji, Inigo Mart, Inego de Rituerto de Troya, Manas Gaur, Leid Zejnilovic, "A Hybrid Recommender System for Patient-Doctor Matchmaking in Primary Care", Universidade Nova de Lisboa, School of Business and Economics, Aug 2018

[6] Dipanwita Dasgupta, Nitesh V. Chawla, "MedCare: Leveraging Medication Similarity for Disease Prediction", NSF grants IIS-1447795 and BCS-1229450.

[7] S.Lavanya, G.Lavanya, J.Divyabharathi, "Remote Prescription And I- Home healthcare based on IoT", IEEE International Conference on Innovations in Green Energy and Healthcare Technologies(ICIGEHT'17) Emre Sezgin, Sevgi Ozkan, "A Systematic Literature Review on Health Recommender Systems", The 4th IEEE International Conference on E-Health and Bioengineering - EHB 2013

[8] Grigore T. Popa University of Medicine and Pharmacy, Iași, Romania, November 21-23, 2013

[9] Arezou Koohi, ECE department, George Mason University, Fairfax, VA, USA, "Prediction of drug-target interactions using popular Collaborative Filtering methods", 2013 IEEE International Workshop on Genomic Signal Processing and Statistics

[10] Martin Wiesner and Daniel Pfeifer "Health Recommender Systems: Concepts, Requirements, Technical Basics and

Challenges”, Int J Environ Res Public Health. 2014 Mar; 11(3): 2580–2607. Published online 2014 Mar 3.