A Collaborative Filtering Approach to Assess Individual Condition Risk Based on Patients’ Social Network Data

Xiang Ji  
New Jersey Institute of Technology  
323 Martin Luther King Jr. Blvd  
Newark, NJ, 07102  
xj25@njit.edu

Soon Ae Chun  
CUNY College of Staten Island  
2800 Victory Blvd  
Staten Island, NY, 10304  
soon.chun@csi.cuny.edu

James Geller  
New Jersey Institute of Technology  
323 Martin Luther King Jr. Blvd  
Newark, NJ, 07102  
geller@njit.edu

ABSTRACT
Healthcare research has shown that conditions are correlated with each other, for example, in patients with type-2 diabetes, chronic nephropathy often results from diabetic nephropathy. This correlation is called comorbidity relationship. The comorbidity relationships are often so complex that it is difficult to comprehend them. A disease prediction model extending the collaborative filtering used in recommender systems was developed to use publicly available patients’ social network data to predict such comorbidity relationships, and to help doctors as well as uninformed patients to assess potential health risks.

Categories and Subject Descriptors
H.2.8 [Database Applications]: Data mining  
J.3 Life and Medical Sciences: Medical information systems

General Terms
Algorithms

Keywords
Recommender System, EHR, Social Computing

1. INTRODUCTION
Through lab tests, certain conditions of patients will be diagnosed, and these diagnosis results will be recorded in an electronic health record (EHR). Such a record provides a summary of an individual’s medical history and is often made accessible to the patient online. In recent years patients have begun to turn to social media, particularly patient communities, for patient-generated knowledge. PatientsLikeMe [1] has over 187,000 members and covers over 500 health conditions. The registered users in PatientsLikeMe publish their demographic information as well as their medical conditions associated with the date of first symptom and the date of first diagnosis.

Research has shown that some conditions are correlated with each other to a measureable degree ("co-morbidities") [2]. Due to similar molecules, gene structures, and patients’ life styles, the appearance of certain conditions leads to a higher likelihood for the occurrence of certain other conditions. These correlation relationships are usually complex. The objective of this paper is to use patients’ social network data to model the comorbidity relationships with a recommender system technique and to predict the most probable conditions a patient in the future. This model is intended to help doctors and patients to discover potential future conditions more quickly and to prescribe treatments at the earliest stage possible.

2. RELATED WORK
A recommender system is a tool to predict the ratings of users for unrated items based on their previous ratings or ratings from similar users. Collaborative filtering (CF) [3] is one of the most efficient recommendation techniques to date. CF recommends items based on the rating history of similar users. It searches for similar users in terms of the rating history and then predicts the ratings of a new item based on an aggregation function of similar users’ ratings on that item. The value of the aggregation function is calculated as the utility of the new item, which measures how likely it is that a user will be interested in the item.

Recommendation techniques have been previously applied to solve medical problems. Davis et al. [4] described the CARE system, which was the first system that applied recommender system to disease prediction. Hassan and Syed [5] incorporated demographics, comorbidity, lab test results, and outcome into the feature space and evaluated the performance of collaborative filtering on 4,500 patients in predicting sudden cardiac death and recurrent myocardial infarction. EHRs were utilized for providing the dataset by the above researchers to create a recommender system. In this paper, we utilize social network data available online. Compared with EHRs, the data on social networks have the advantages of open access and the lack of privacy issues.

3. METHOD
There are two scenarios of prediction, the first scenario is to predict a new user’s future conditions, and the second one is to predict every existing user’s conditions. Due to space limitations, we only discuss the prediction of condition for a new user.

Let \( I = \{\text{all users}\} \), \( J_i = \{\text{all the conditions of user } i\} \). When a new user enters data, this user is regarded as user \( 0 \), and \( J_0 = \{\text{all the conditions of user } 0\} \). The set head is the new user’s conditions that will be used to compare her/him with other patients. The set Condition-Union \( D \) is defined as \( \bigcup_{i \in I} J_i \). The set Target \( T \) is defined as \( T = D \setminus \text{head} \). The objective of the prediction model is to predict the likelihood of each condition in \( T \).
For each condition \( c \) in \( T \), the neighbors \( N_c = \{ i \mid i \in I \land c \in f_i \} \), which are all users who were also diagnosed with condition \( c \). The probability of user 0 having condition \( c \) in the future is calculated by the following formula:

\[
P_{0,c} = k \sum_{i \in N_c} w(0,i)
\]

Where \( k \) is the normalizing factor, and is defined as the total number of patients in the neighborhood, formally \( k = 1/|N_c| \). The weight \( w(0,i) \) is defined as the proportion of conditions of user \( i \) to the conditions in \( head \). More generally, the weight (similarity) of user \( i \) and user \( j \) is defined by the following formula:

\[
w(i, j) = \frac{|\{x \mid x \in head \land x \in f_j\}|}{|head|}
\]

where \( head \) is the set of conditions reported by user \( i \). In this new-user scenario, \( head \) is the set of all conditions reported by the new user. Condition \( c \)'s support \( S_c \) is defined as the ratio between the number of patients who have condition \( c \) and the total number of patients. After \( P_{0,c} \) and \( S_c \) are computed, the list of potential conditions \( PL_i \) is defined as \( \{0, c, P_{0,c}, S_c \mid c \in T \} \), where the tuple \( <0, c, P_{0,c}, S_c> \) represents that user 0 has the probability \( P_{0,c} \) of getting condition \( c \) in the future, with support \( S_c \).

An example diagnosis dataset shown in Table 1 illustrates how the above method works in predicting a new user’s potential conditions. The conditions of each patient are ordered by the patient’s diagnosis date for each condition. For example, \( P_1 \) was first diagnosed with \( C_1 \), then diagnosed with \( C_2 \) and then diagnosed with \( C_5 \) and \( C_6 \), \( I = \{P_1, P_2, P_3, P_4, P_5\} \). \( D = \cup_{i \in I} f_i = \{C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8\} \). When a new user \( P_0 \) is diagnosed with conditions of \( C_1 \) and \( C_3 \), then \( head = \{C_1, C_3\} \). Target \( T = D \) \( - \) \( head = \{C_2, C_4, C_5, C_6, C_7, C_8\} \). Consider the first condition \( C_7 \) in \( T \), \( N_c = \{P_2, P_3\} \), \( w(P_0, P_1) = 1 \), and \( w(P_0, P_3) = 0 \), then \( P_{0,c} = 0.5 \). \( S_c = 0.4 \). The tuple for condition \( C_7 \) is \( <0, C_2, 0.5, 0.4\rangle \). Similarly, the tuples for the other conditions in \( T \) are \( <0, C_4, 0.5, 0.4\rangle \), \( <0, C_6, 0.25, 0.4\rangle \), \( <0, C_6, 0.5, 0.2\rangle \), \( <0, C_7, 0.5, 0.8\rangle \), and \( <0, C_8, 0.5, 0.4\rangle \). In this example, \( C_7 \) is the most probable future condition for the new user, followed by \( C_6 \).

Table 1. An Example of Diagnosis Dataset

<table>
<thead>
<tr>
<th>Patient</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P_1 )</td>
<td>( C_1, C_2, C_3, C_4, C_7 )</td>
</tr>
<tr>
<td>( P_2 )</td>
<td>( C_2, C_3, C_4, C_5 )</td>
</tr>
<tr>
<td>( P_3 )</td>
<td>( C_5, C_6, C_8, C_7 )</td>
</tr>
<tr>
<td>( P_4 )</td>
<td>( C_7, C_8, C_6 )</td>
</tr>
<tr>
<td>( P_5 )</td>
<td>( C_6, C_7 )</td>
</tr>
</tbody>
</table>

4. Experiments

4.1 Dataset
We collected patient and diagnosis datasets by scraping patients’ public profiles in PatientsLikeMe. The patient dataset contains 17,407 patients and their demographic information. The diagnosis dataset contains 35,606 diagnoses, and each contains 6 attributes: Id, Condition, Cid, IsPrimary, SymptomDate, and Diagnosis Date. Conditions with most patients are shown in Table 2.

4.2 Results
We used two metrics: coverage and half-life decay accuracy to evaluate the prediction performances. The coverage is used to measure what percentage of conditions diagnosed for patients in the future is covered in the prediction list. The half-life decay accuracy is the ratio between the sum of likelihoods of correct conditions in the predicted ranked list and the sum of likelihoods of the correct conditions in perfectly ranked list. The results are shown in Table 3, where \( k \) is size of prediction list and head size is set to 2. Examples of predictions for each patient are in Table 4.

Table 2. The conditions with most patients

<table>
<thead>
<tr>
<th>Condition</th>
<th>Number of Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS (Multiple Sclerosis)</td>
<td>3459</td>
</tr>
<tr>
<td>Fibromyalgia</td>
<td>3164</td>
</tr>
<tr>
<td>Major Depressive Disorder</td>
<td>1624</td>
</tr>
<tr>
<td>Generalized Anxiety Disorder</td>
<td>1106</td>
</tr>
<tr>
<td>Chronic Fatigue Syndrome</td>
<td>914</td>
</tr>
</tbody>
</table>

Table 3. Prediction Results

<table>
<thead>
<tr>
<th>( k )</th>
<th>Average Accuracy</th>
<th>Average Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.267</td>
<td>0.220</td>
</tr>
<tr>
<td>10</td>
<td>0.280</td>
<td>0.298</td>
</tr>
<tr>
<td>20</td>
<td>0.299</td>
<td>0.401</td>
</tr>
<tr>
<td>50</td>
<td>0.302</td>
<td>0.517</td>
</tr>
<tr>
<td>100</td>
<td>0.302</td>
<td>0.578</td>
</tr>
</tbody>
</table>

Table 4. Example of prediction on individual patient

<table>
<thead>
<tr>
<th>Patient Id</th>
<th>Diagnosed Conditions</th>
<th>Top 2 Predicted Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>296</td>
<td>Migraine, Fibromyalgia</td>
<td>Chronic Fatigue Syndrome, Generalized Anxiety Disorder</td>
</tr>
<tr>
<td>42</td>
<td>Eating Disorder, Phobic disorder</td>
<td>Social Anxiety Disorder, PTSD</td>
</tr>
<tr>
<td>50</td>
<td>HIV, Seborrheic Dermatitis</td>
<td>Bipolar Disorder, Lactose Intolerance</td>
</tr>
</tbody>
</table>

5. Conclusions and Future Work
A novel method was developed and used for predicting online patients’ future conditions based on past conditions. Prediction performance was evaluated using coverage and half-life accuracy. Experiments show that the model is able to predict future conditions for online users with 0.3 accuracy and between 0.22 and 0.57 coverage, depending on the size the ranked list. In the future, we will consider better temporal similarity measures and more demographic features, such as gender, age, and location, etc.

6. References