A Similarity Measure for Patient Sequences
A Case Study on Predicting Anxiety/Depression for College Students using Case-Based Reasoning

Jinghe Zhang

May 2, 2015
Outline

- Introduction
- Related Work
- Methodology
- Metric Evaluation
- Experiments & Results
- Conclusions & Future Work
- References
Some Facts about Mental Health in US...

- 18.6% of adults (42.5 million) suffer from mental illness, such as depression, bipolar disorders, etc.
Some Facts about Mental Health in US...

• 18.6% of adults (42.5 million) suffer from mental illness, such as depression, bipolar disorders, etc.

• 4 of the 10 leading causes of disability are mental disorders: major depression, bipolar disorders, schizophrenia, and obsessive-compulsive disorder.
Some Facts about Mental Health in US...

- 18.6% of adults (42.5 million) suffer from mental illness, such as depression, bipolar disorders, etc.
- 4 of the 10 leading causes of disability are mental disorders: major depression, bipolar disorders, schizophrenia, and obsessive-compulsive disorder.
Some Facts about Mental Health in US...

- Anxiety disorders are the most common mental illness, affecting 40 million adults. It is highly treatable, yet only one-third of those suffering receive treatment.

- Depression is a condition in which a person feels discouraged, sad, hopeless, unmotivated, or disinterested in life in general. Major depression involves at least five of these symptoms for a two-week period and it is the leading cause of disability for ages 15 to 44.3.

- Nearly one-half of those diagnosed with depression are also diagnosed with an anxiety disorder.

- Women are 60% more likely than men to experience an anxiety disorder over their lifetime and nearly twice as many women (12.0 percent) as men (6.6 percent) are affected by a depressive disorder each year.

- Anxiety/depressive disorders develop from a complex set of risk factors, including genetics, brain chemistry, personality, and life events.
Some Facts about Mental Health Among College Students...

- College Students responding to the Spring 2014 American College Health Association-National College Health Assessment reported feeling things were hopeless (46%), felt overwhelming anxiety (54%) and more than 80% reported feeling overwhelmed by all they had to do (86%).

- This subpopulation is facing significant levels of mental health problems.
Some Facts about Mental Health Among College Students...

- College Students responding to the Spring 2014 American College Health Association-National College Health Assessment reported feeling things were hopeless (46%), felt overwhelming anxiety (54%) and more than 80% reported feeling overwhelmed by all they had to do (86%).

- This subpopulation is facing significant levels of mental health problems.

- Psychiatric disorders are frequently unrecognized in primary care settings, posing physical, emotional, economic, and social burdens to patients and others.

- Early identification and treatment is helpful.
Related Work

Case-based Reasoning:

- To solve a new problem based on the solutions of similar past problems
- A recognized method for decision making in medical area; however, not successful in medicine as in other applications
- Medical data are especially complex to define a meaningful similarity metric on them.

A patient profile:

- patient → document
- diagnoses/treatment/etc. → terms/features
Related Work

Case-based Reasoning:

- To solve a new problem based on the solutions of similar past problems

Medical data are especially complex to define a meaningful similarity metric on them.

A patient profile:

- patient → document
- diagnoses/treatment/etc. → terms/features
Related Work

Case-based Reasoning:

- To solve a new problem based on the solutions of similar past problems
- A recognized method for decision making in medical area; however, not successful in medicine as in other applications
- Medical data are especially complex to define a meaningful similarity metric on them.
Related Work

**Case-based Reasoning:**

- To solve a new problem based on the solutions of similar past problems
- A recognized method for decision making in medical area; however, not successful in medicine as in other applications
- Medical data are especially complex to define a meaningful similarity metric on them.

**A patient profile:**

- **patient** → **document**
- **diagnoses/treatment/etc.** → **terms/features**
Methodology

Research Objectives:

- To develop a meaningful similarity metric for patient sequences
- To predict anxiety/depression according to case-based reasoning using the similarity metric

Two-layer Similarity Metric:

- Layer 1 (visit-level similarity): similarity between visits (itemsets) from two distinct sequences $x = (x_1, ..., x_N)$ and $y = (y_1, ..., y_M)$
- Layer 2 (sequence-level similarity): overall similarity between $x$ and $y$ according to visit alignment achieved in Layer 1
Methodology

Research Objectives:

- To develop a meaningful similarity metric for patient sequences
- To predict anxiety/depression according to case-based reasoning using the similarity metric

Two-layer Similarity Metric:

- Layer 1 (visit-level similarity): similarity between visits (itemsets) from two distinct sequences $x = (x_1, ..., x_N)$ and $y = (y_1, ..., y_M)$
- Layer 2 (sequence-level similarity): overall similarity between $x$ and $y$ according to visit alignment achieved in Layer 1
Visit-level similarity: Jaccard similarity

\[ J(x_i, y_j) = \frac{|x_i \cap y_j|}{|x_i \cup y_j|} \]  

where \( x_i \) is the \( i \)th visit in sequence \( x \) and \( y_j \) is the \( j \)th visit in sequence \( y \).
Visit-level similarity: Jaccard similarity

\[ J(x_i, y_j) = \frac{|x_i \cap y_j|}{|x_i \cup y_j|} \]  \hspace{1cm} (3.1)

where \( x_i \) is the \( i \)th visit in sequence \( x \) and \( y_j \) is the \( j \)th visit in sequence \( y \).

Sequence-level similarity:
Visit-level similarity: Jaccard similarity

\[ J(x_i, y_j) = \frac{|x_i \cap y_j|}{|x_i \cup y_j|} \]  

(3.1)

where \( x_i \) is the \( i \)th visit in sequence \( x \) and \( y_j \) is the \( j \)th visit in sequence \( y \).

Sequence-level similarity:

1. Visit alignment (to align visit \( i \) with visit \( j \) based on their similarity)
Visit-level similarity: Jaccard similarity

\[ J(x_i, y_j) = \frac{|x_i \cap y_j|}{|x_i \cup y_j|} \quad (3.1) \]

where \( x_i \) is the \( i \)th visit in sequence \( x \) and \( y_j \) is the \( j \)th visit in sequence \( y \).

Sequence-level similarity:

1. Visit alignment (to align visit \( i \) with visit \( j \) based on their similarity)
2. Penalize visit gap in aligned visit pairs and cross-alignment
Methodology (cont’d)

Visit-level similarity: Jaccard similarity

\[ J(x_i, y_j) = \frac{|x_i \cap y_j|}{|x_i \cup y_j|} \]  

(3.1)

where \( x_i \) is the \( i \)th visit in sequence \( x \) and \( y_j \) is the \( j \)th visit in sequence \( y \).

Sequence-level similarity:

1. Visit alignment (to align visit \( i \) with visit \( j \) based on their similarity)
2. Penalize visit gap in aligned visit pairs and cross-alignment

Patient A  
- Visit 1  
- Visit 2

Patient B  
- Visit 1  
- Visit 2

(a) Alignment gap  
(b) cross-alignment

3. Compute overall similarity
Methodology (cont’d)

Sequence-level similarity:

1. Visit alignment (to align visit \( i \) with visit \( j \) based on their similarity)
Sequence-level similarity:

1. Visit alignment (to align visit $i$ with visit $j$ based on their similarity)

Munkres algorithm:

$$
\begin{array}{c|cccc}
 & y_1 & y_2 & y_3 & y_4 \\
\hline
x_1 & 1.0 & 0.5 & 0.3 & 0.6 \\
x_2 & 0.1 & 0 & 0.2 & 0.5 \\
x_3 & 0.3 & 0.5 & 0.8 & 0.7 \\
\end{array}
$$

Table 1: An Example of the Similarity Matrix of Visits in Two Sequences
Methodology (cont’d)

Sequence-level similarity:

1. Visit alignment (to align visit $i$ with visit $j$ based on their similarity)
   Munkres algorithm:

   Table 1: An Example of the Similarity Matrix of Visits in Two Sequences

<table>
<thead>
<tr>
<th></th>
<th>$y_1$</th>
<th>$y_2$</th>
<th>$y_3$</th>
<th>$y_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>1.0</td>
<td>0.5</td>
<td>0.3</td>
<td>0.6</td>
</tr>
<tr>
<td>$x_2$</td>
<td>0.1</td>
<td>0</td>
<td>0.2</td>
<td>0.5</td>
</tr>
<tr>
<td>$x_3$</td>
<td>0.3</td>
<td>0.5</td>
<td>0.8</td>
<td>0.7</td>
</tr>
</tbody>
</table>

2. Penalize visit gap in aligned visit pairs and cross-alignment
Methodology (cont’d)

Sequence-level similarity:

1. Visit alignment (to align visit $i$ with visit $j$ based on their similarity)

   Munkres algorithm:

   
<table>
<thead>
<tr>
<th></th>
<th>$y_1$</th>
<th>$y_2$</th>
<th>$y_3$</th>
<th>$y_4$</th>
</tr>
</thead>
</table>
   $x_1$ | 1.0 | 0.5 | 0.3 | 0.6 |
   $x_2$ | 0.1 | 0.0 | 0.2 | 0.5 |
   $x_3$ | 0.3 | 0.5 | 0.8 | 0.7 |

   Table 1: An Example of the Similarity Matrix of Visits in Two Sequences

2. Penalize visit gap in aligned visit pairs and cross-alignment

3. Compute overall similarity

   \[
   \text{sim}(x_i, y_j) = \sum_{(x_i, y_j) \in U} f(x_i, y_j) J(x_i, y_j) \tag{3.2}
   \]

   where $U$ is the set of aligned visit pairs and $f(x_i, y_j)$ is the penalty function.
Challenges in patient similarity evaluation:

- Human judgement is expensive
- Human judgement is inconsistent

Solution: evaluation by the performance of its applications - similarity-based classification

Model A: KNN based on majority voting
Model B: Weighted KNN method
Model C: Nearest centroid classifier

Represent the set of training observations in class $i$ by its centroid $\mu_i$

$$\mu_i = \arg \max_{\mu \in X} \sum_{t \in X_i} \text{sim}(t, \mu)$$ (4.1)

Assign the new observation $x$ the label of the class $i$ whose centroid $\mu_i$ is closest to the observation

$$\hat{y} = \arg \max_{i = 1, 2, ..., N} \text{sim}(x, \mu_i)$$ (4.2)
Challenges in patient similarity evaluation:

- Human judgement is expensive
- Human judgement is inconsistent

Solution: evaluation by the performance of its applications - similarity-based classification

Model A: KNN based on majority voting
Model B: Weighted KNN method
Model C: Nearest centroid classifier

Represent the set of training observations in class $i$ by its centroid $\mu_i$:

$$\mu_i = \arg\max_{\mu \in \mathcal{X}_i} \sum_{t \in \mathcal{X}_i} \text{sim}(t, \mu) \quad (4.1)$$

Assign the new observation $x$ the label of the class $i$ whose centroid $\mu_i$ is closest to the observation:

$$\hat{y} = \arg\max_{i = 1, 2, \ldots, N} \text{sim}(x, \mu_i) \quad (4.2)$$
• Challenges in patient similarity evaluation:
  • Human judgement is expensive
  • Human judgement is inconsistent
• Solution: evaluation by the performance of its applications - similarity-based classification
Metric Evaluation

- Challenges in patient similarity evaluation:
  - Human judgement is expensive
  - Human judgement is inconsistent

- Solution: evaluation by the performance of its applications - similarity-based classification
  - Model A: KNN based on majority voting
  - Model B: Weighted KNN method
  - Model C: Nearest centroid classifier

- Represent the set of training observations in class $i$ by its centroid $\mu_i$

  \[
  \mu_i = \arg\max_{\mu \in X_i} \sum_{t \in X_i} \text{sim}(t, \mu) \quad (4.1)
  \]

- Assign the new observation $x$ the label of the class $i$ whose centroid $\mu_i$ is closest to the observation

  \[
  \hat{y} = \arg \max_{i=1,2,\ldots,N} \text{sim}(x, \mu_i) \quad (4.2)
  \]
• Data Description:
  • College Health Surveillance Network (CHSN) is the first national database of college student’s health data collected from 23 student health centers.
  • 3000 patients from each class (patients diagnosed with anxiety/depression and patients without any mental disorder)
  • Features are disease clusters according to ICD-9 codes (9th revision of the International Statistical Classification of Diseases and Related Health Problems)
  • 80 features are selected based on their information gain
Data Description:
- College Health Surveillance Network (CHSN) is the first national database of college student’s health data collected from 23 student health centers.
- 3000 patients from each class (patients diagnosed with anxiety/depression and patients without any mental disorder)
- Features are disease clusters according to ICD-9 codes (9th revision of the International Statistical Classification of Diseases and Related Health Problems)
- 80 features are selected based on their information gain

Benchmark models:
- KNN models with ”bag-of-words” (BOW) similarity
- Linear SVM model with BOW features
Experiments & Results

- Experiments:
  - Model A: KNN based on majority voting
  - Model B: Weighted KNN method
  - Model C: Nearest centroid classifier

Table 2: Average Precision, Recall, and F1 Score of Classification Models

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOW Similarity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>0.691</td>
<td>0.693</td>
<td>0.536</td>
</tr>
<tr>
<td>Recall</td>
<td>0.731</td>
<td>0.732</td>
<td>0.511</td>
</tr>
<tr>
<td>F1</td>
<td>0.711</td>
<td>0.712</td>
<td>0.522</td>
</tr>
<tr>
<td>Sequence Similarity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>0.498</td>
<td>0.514</td>
<td>0.441</td>
</tr>
<tr>
<td>Recall</td>
<td>0.591</td>
<td>0.543</td>
<td>0.545</td>
</tr>
<tr>
<td>F1</td>
<td>0.539</td>
<td>0.527</td>
<td>0.485</td>
</tr>
</tbody>
</table>

*A is KNN with majority voting; B is weighted KNN by similarity; C is the nearest centroid classifier.
Experiments & Results

- Experiments:
  - Model A: KNN based on majority voting
  - Model B: Weighted KNN method
  - Model C: Nearest centroid classifier
- 10-fold cross validation
- K is tuned to achieve optimal performance.

Table 2: Average Precision, Recall, and F1 Score of Classification Models

<table>
<thead>
<tr>
<th></th>
<th>Sequence Similarity</th>
<th>BOW Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Precision</td>
<td>0.691</td>
<td>0.693</td>
</tr>
<tr>
<td>Recall</td>
<td>0.731</td>
<td>0.732</td>
</tr>
<tr>
<td>F1</td>
<td>0.711</td>
<td>0.712</td>
</tr>
</tbody>
</table>

*A is KNN with majority voting; B is weighted KNN by similarity; C is the nearest centroid classifier.*
Conclusions & Future Work

- Proposed a similarity metric for patient sequences to enable case-based reasoning in medical decision making.

- The proposed metric is evaluated by its application in classifications and the performance is better than using BOW similarity.

- The KNN with majority voting and weighting outperform the nearest centroid method in these experiments.

- The current similarity metric is optimized in a greedy manner and other optimization methods, such as dynamic programming, will be explored to achieve global optimum.

- More features will be included in the patient similarity computation, such as medication and procedures.

- Further validation and application of this proposed similarity metric will be conducted.
References


Thank you!