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...

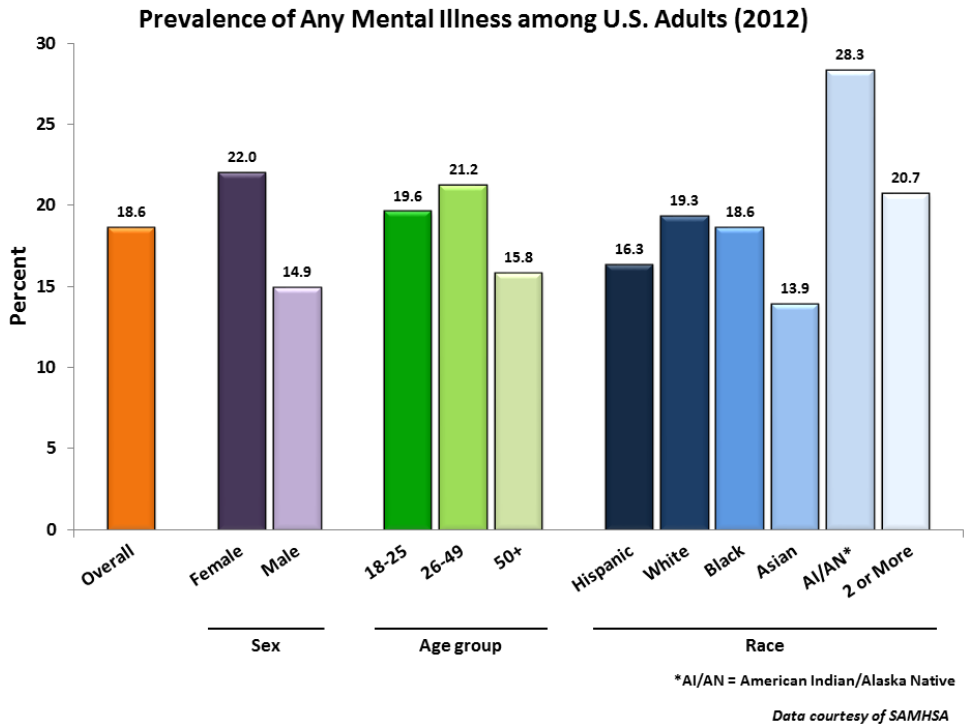
- 18.6% of adults (42.5 million) suffer from mental illness, such as depression, bipolar disorders, etc.

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.

Introduction

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.

Introduction

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:

Related Work

Case-based Reasoning:

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

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.

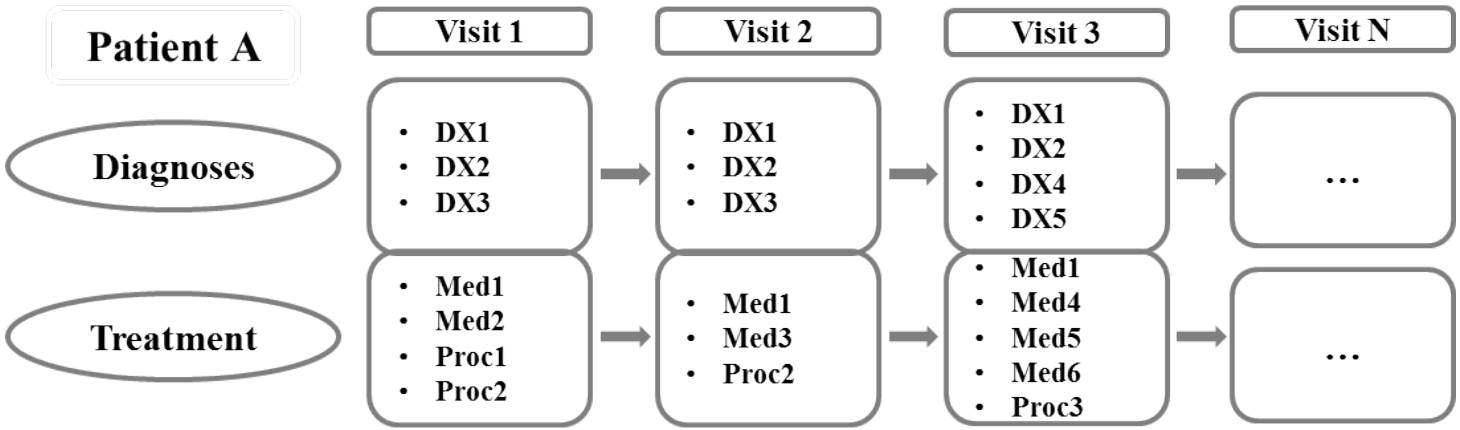
Related Work

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

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 (cont'd)

Visit-level similarity: Jaccard similarity

$$J(x_i, y_j) = \frac{|x_i \cap y_j|}{|x_i \cup y_j|} \tag{3.1}$$

where x_i is the ith visit in sequence x and y_j is the jth visit in sequence y.

Methodology (cont'd)

Visit-level similarity: Jaccard similarity

$$J(x_i, y_j) = \frac{|x_i \cap y_j|}{|x_i \cup y_j|} \tag{3.1}$$

where x_i is the ith visit in sequence x and y_j is the jth visit in sequence y.

Sequence-level similarity:

Methodology (cont'd)

Visit-level similarity: Jaccard similarity

$$J(x_i, y_j) = \frac{|x_i \cap y_j|}{|x_i \cup y_j|} \tag{3.1}$$

where x_i is the ith visit in sequence x and y_j is the jth visit in sequence y.

Sequence-level similarity:

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

Methodology (cont'd)

Visit-level similarity: Jaccard similarity

$$J(x_i, y_j) = \frac{|x_i \cap y_j|}{|x_i \cup y_j|} \tag{3.1}$$

where x_i is the ith visit in sequence x and y_j is the jth 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

(a) Alignment gap (b) cross-alignment

Methodology (cont'd)

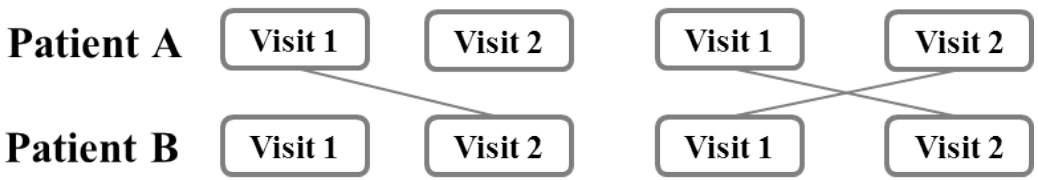
Visit-level similarity: Jaccard similarity

$$J(x_i, y_j) = \frac{|x_i \cap y_j|}{|x_i \cup y_j|} \tag{3.1}$$

where x_i is the ith visit in sequence x and y_j is the jth 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



(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)

Methodology (cont'd)

Sequence-level similarity:

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

	y_1	y_2	y_3	y_4
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

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

Methodology (cont'd)

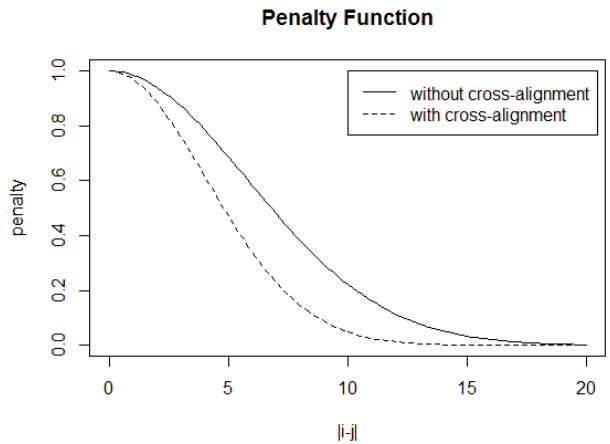
Sequence-level similarity:

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

	y_1	y_2	y_3	y_4
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

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

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:

	y_1	y_2	y_3	y_4
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

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

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

where U is the set of aligned visit pairs and $f(x_i, y_j)$ is the penalty function.

Metric Evaluation

- Challenges in patient similarity evaluation:

Metric Evaluation

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

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

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 μ_i

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

    - Assign the new observation x the label of the class i whose centroid μ_i is closest to the observation

$$\hat{y} = \arg \max_{i=1,2,\ldots,N} sim(x, \mu_i) \tag{4.2}$$

Metric Evaluation

- 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

Metric Evaluation

- 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

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

	Sequence Similarity			BOW Similarity			
	A	B	C	A	B	C	Linear SVM
Precision	0.691	0.693	0.536	0.498	0.514	0.441	0.691
Recall	0.731	0.732	0.511	0.591	0.543	0.545	0.667
F1	0.711	0.712	0.522	0.539	0.527	0.485	0.679

*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

Facts & Statistics | Anxiety and Depression Association of America. URL http://www.adaa.org/about-adaa/press-room/facts-statistics.

Mental Disorders in America - Mental Health Center: Medical Information on Mental Illness. URL http://www.medicinenet.com/script/main/art.asp?articlekey=21466.

Prevalence, severity, and unmet need for treatment of mental disorders in the world health organization world mental health surveys. *JAMA*, 291(21):2581, June 2004.

Any mental illness (AMI) among adults. NIH national institute of mental health, 2015. URL http://www.nimh.nih.gov/.

Luca Cazzanti and Maya R. Gupta. Local Similarity Discriminant Analysis. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 137–144, New York, NY, USA, 2007. ACM.

Paul Jaccard. The distribution of the flora in the alpine zone.1. *New Phytologist*, 11(2):37–50, 1912.

Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.

Ted Pedersen, Serguei V. S. Pakhomov, Siddharth Patwardhan, and Christopher G. Chute. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3): 288–299, June 2007.

Esther Thatcher. College health surveillance network, 2015. URL http://socialnorms.org/college-health-surveillance-network/.

Zhengzheng Xing, Jian Pei, and Eamonn Keogh. A Brief Survey on Sequence Classification. *SIGKDD Explor. Newsl.*, 12(1):40–48, November 2010.

Thank you!