Please use approach B - 50 20 20 10 marking scheme: Data Wrangling Focus

CSC 485C Term Project 2
Anna Russo Kennedy
Matt Hemmings
Adrian Kilian
Ryan Schafer

Stress and Disease

Table of Contents

Introduction

For our term project, we were interested in looking at new methods for early detection of chronic diseases. Specifically, we wanted to examine the relationship between stress and a set of diseases not normally attributed to stress, and determine whether high stress levels play a role in the onset of these diseases. As a primary tool, we chose to use Weka to try and leverage machine learning against the problem to see if we could uncover any implicit relationships between variables.

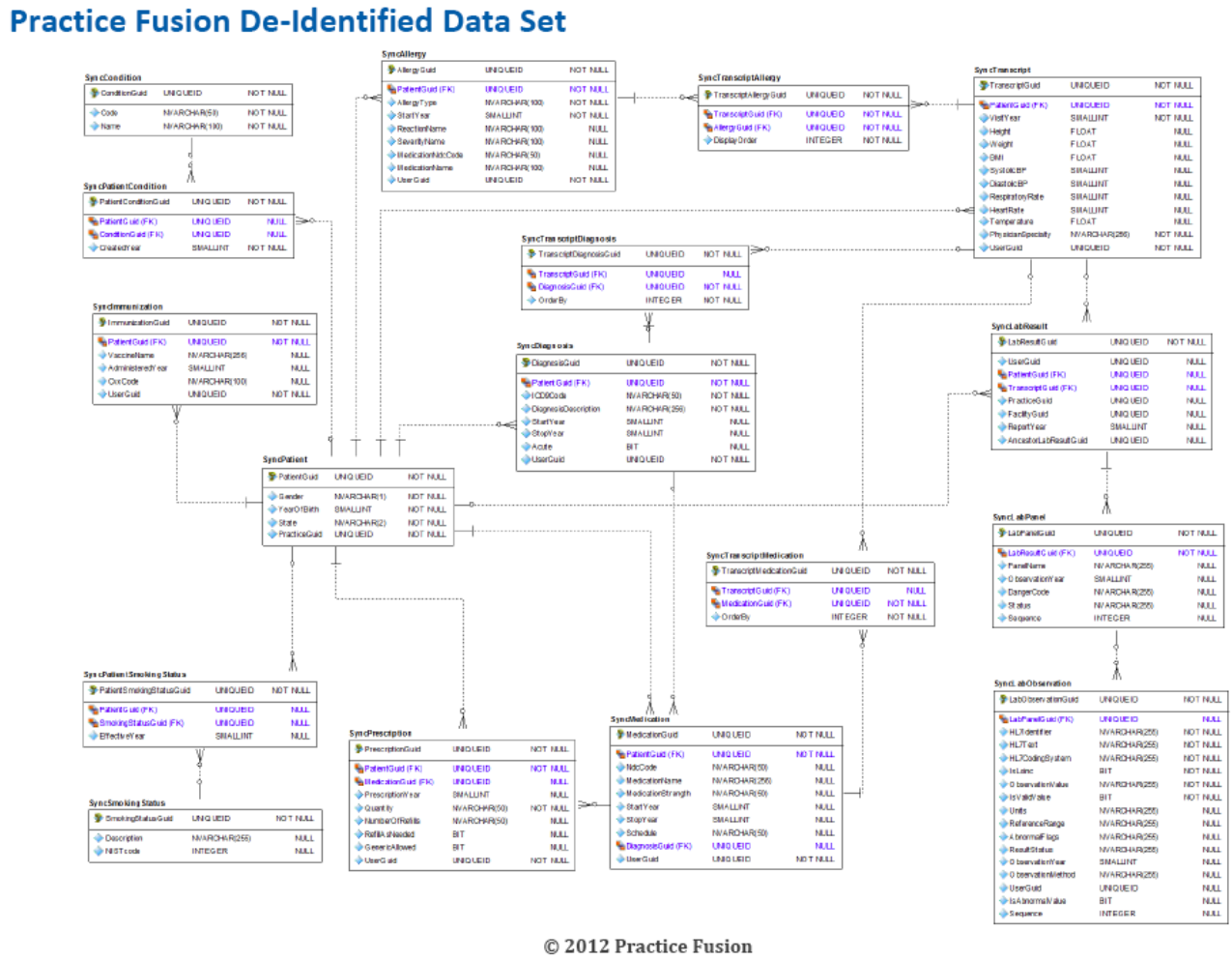
The obvious first step was to acquire a dataset that we would be able to manipulate for our purposes. After a fair amount of searching, we were finally able to find a closed Kaggle competition[1] with an available data set of anonymized patient records.

[1] http://www.kaggle.com/c/pf2012-at

At this point, two team members began work separately on the data set, both with the aim of addressing the question of the stress-disease connection. As such, we ended up with two different paths, that both ended up at the same conclusion: there was not enough data to draw any firm conclusions. In this paper, we describe the two approaches taken, what we would do differently next time, and what we feel would be the ideal set of data to approach this question.

The Data

The data set obtained from the closed Kaggle competition is a set of de-identified patient records spanning 3-4 years of doctor's visits, provided by the company Practice Fusion[2], an Electronic Health Records (EHR) service provider. The data was originally from a large relational database, and was broken down into 15 CSV files, with an accompanying Data Dictionary[3] and Data Model Diagram (see Figure 1, below). The data set was a complex collection of information on diagnoses, lab results, medications, allergies, immunizations, vital signs, and health behavior.



[2] http://www.practicefusion.com/

[3] https://www.kaggle.com/c/pf2012/download/PracticeFusionDataSetDictionary.pdf

Figure 1

Approach One

Wrangle

We began by concatenating several of the CSV files in the data set so that we could run the result through Weka to see what outcomes we could gather from it.

First, we tried combining the data from several CSV files, pulling out the values for BMI (body mass index), BP (blood pressure), whether the patient was taking medication, and if and how much they smoked per day. Since the thrust of the project was to have a correlation of a stress indicator associated with the disease, we realised we would need a scale for stress. We located the Holmes and Rahe Stress Scale[4], which evaluates life events to assign a numerical value and interval that indicates an individual's stress level. Since there was no list of these stress numbers associated with the diagnoses found in the Kaggle competition, we were forced to invent some for ourselves.

In order to create the stress numbers, the data set was run through a python script that would break each line apart and assign a random number between 0 and 450, then check which interval it fell in and, based on the interval, a probability that the disease was stress related was assigned. The higher the interval, the higher the probability that the disease diagnosed by the physician was assigned as stress related.

Analyze & Visualize

Loading this data in Weka initially led to the result that, since there was no disease name or identifier on the file, the machine learning algorithm broke down the results solely by stress value, as it had no other means to determine a result (See Figure 2, below).

[4] http://www.mindtools.com/pages/article/newTCS_82.htm

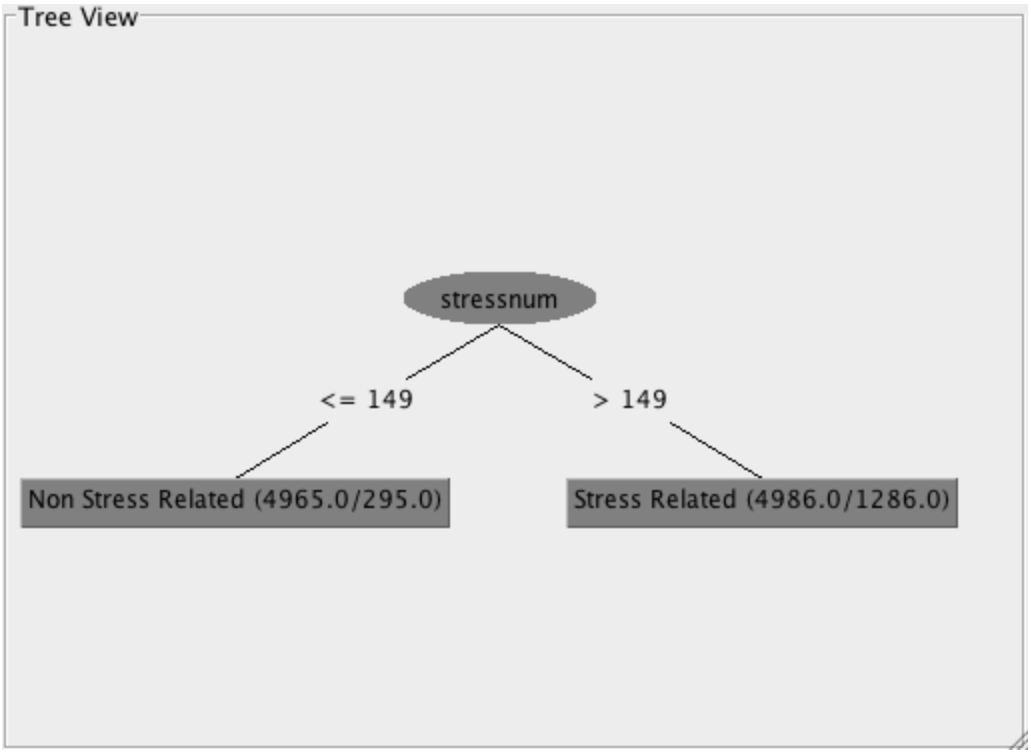


Figure 2

When we tried to associate the value of the patient ID to a disease, we discovered that the same patient would have visited a physician more than once - information found in yet another file - resulting in more than one diagnosis. When we tried to narrow focus by diagnosis ID to the unique ID assigned at the visit, we discovered that these were not correlated to a patient ID, but to a TranscriptDiagnosis ID, that was, in turn, in its own file. As we dug deeper, it became clear that the relations kept getting bigger and bigger and were still not yielding the results we had hoped for.

Approach One Conclusions

In order to fix the data problems we encountered, we would have needed a set of data in something similar to the following format:

@attribute Symptom1 REAL
@attribute Symptom2 REAL
@attribute Symptom3 REAL
.
.
.
@attribute SymptomN REAL

@attribute StressNum REAL

@attribute Disease {Disease1,Disease2,Disease3,...,DiseaseN}

This would have allowed the algorithm to diagnose whether the disease was related to the stress interval the patient was found to fall into. This data would be easiest to gather if there was a single system used to input patient data that could be queried to get data sets. This way, trends over time could be observed by patient and you could tell if, as stress worsened, their other symptoms varied as well.

Approach Two

Wrangle

For this approach we chose to examine the relationship between stress and a narrowed set of diseases including arthritis, cancer, MS, IBS and diabetes. This approach used only one CSV file from the provided data set, which contained nearly 1,000,000 records with patient IDs, diagnoses (by way of ICD-9 codes[5]) and dates of diagnosis. Each diagnosis from each doctor's visit was a separate entry in the file, so there were many entries for each patient. To filter and collate the data into a useable format for our purposes and for Weka to work with, we wrote a MapReduce job. This used our simple MapReduce framework from Assignment 2 to bucket the data by patient ID, and to indicate whether or not they had been diagnosed with any of the 5 diseases in question, as well as whether or not they had been diagnosed as having stress-related issues. This resulted in an .arff file formatted as follows:

@attribute patient_id string (unique)
@attribute cancers {0,1}
@attribute diabetes {0,1}
@attribute stress {0,1}
@attribute MS {0,1}
@attribute IBS {0,1}
@attribute arthritis {0,1}

The resulting data set was much smaller than the original - approximately 4600 records.

Analyze & Visualize

We loaded the data file into Weka and used a J48 decision tree to determine how stress factored into the various diseases, and in the hopes of uncovering some new relationships between the variables. We worked through using each of the disease variables as the predictive class variable. Figure 3 shows Weka's decision tree for predicting whether or not a patient would be diagnosed with cancer, which it was able to do with ~91% accuracy, using 10-Fold

[5] http://en.wikipedia.org/wiki/List_of_ICD-9_codes

cross-validation.

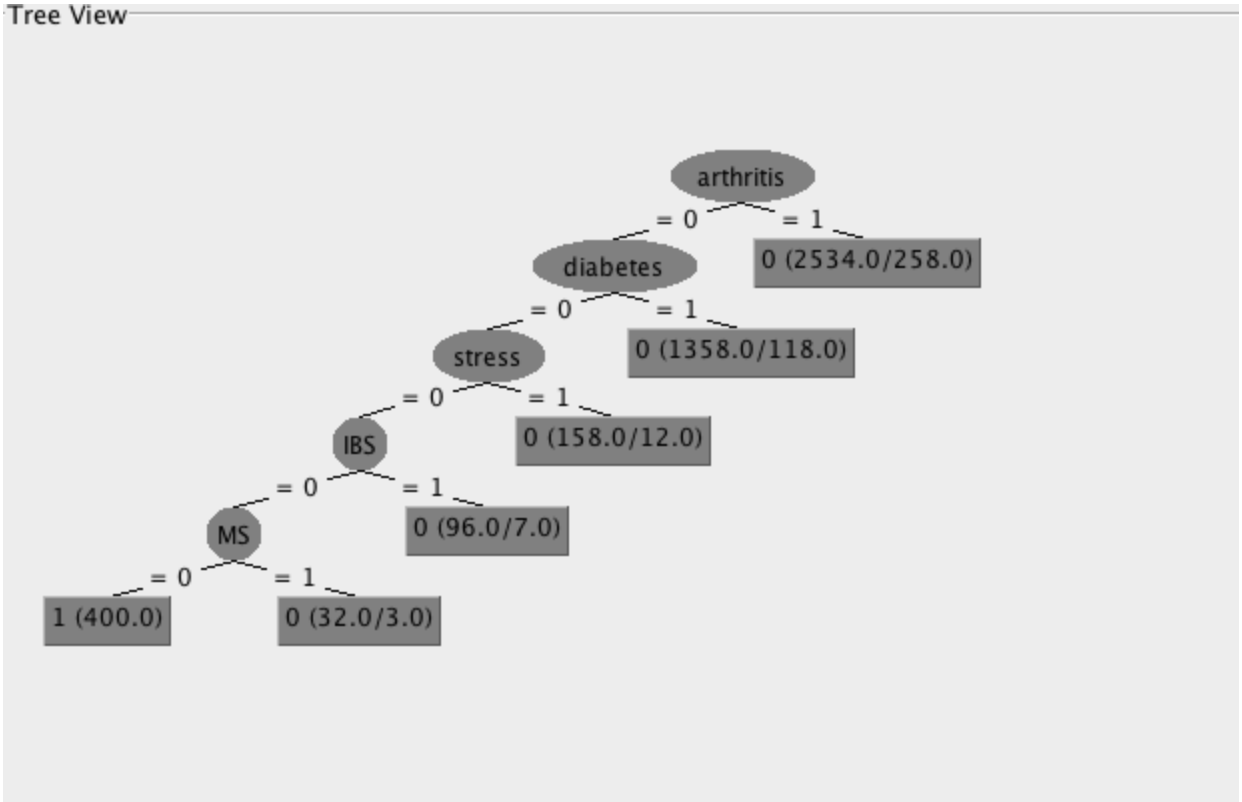


Figure 3

Approach Two Conclusions

The ideal way to approach the study of this question would be to have many years of medical records, tracking a set of patients and their symptoms for at least 20 years. Each year, each patient would complete a Holmes and Rahe Stress Scale[6]-like evaluation, and their various diagnoses would be tracked in a similar manner to the current data set. In truth, machine learning would not really be necessary to answer this question, as we would simply be analyzing the trends of the data, and whether or not a correlation can be drawn between individuals with high stress ratings and the onset of these diseases not typically associated with stress.

[6] http://www.mindtools.com/pages/article/newTCS_82.htm

Conclusions

While our examinations of this data set proved inconclusive, we believe that the question of the relation of stress and disease still remains a valid and important one. There is much that can be discovered in further studies in this area.