Oral document for ACERO presentation 2015

Anthony Chapman

April 2015

Intro

Good morning and thank you very much for coming today. I'd like to thank the organisers for giving us this great chance to get out of the office!

My name is Anthony Chapman, Hello. My background is in coming science and mathematics, very theoretical stuff.

PhD

Around 4 months ago, I was tricked into a PhD with the Applied Health department to apply some of the theories I've been working on. It involves the analysis of routinely acquired antenatal data and their outcomes or consequences.

For example, it has been statistically proven that a certain type of growth rate of a fetus leads to a higher probability of it having asthma.

Intro ctn..

But this talk isn't about any of that. Today I'm going to be talking about a method I created (with the guidance of my supervisors), for evaluating artificially complete data.

So let's get started,

The problem

What is the problem we're trying to solve? One of this very first problems I spotted when I started this PhD, was the huge amount of missing data in what I am supposed to analyse.

I'm willing to bet that most, if not all, of you who have looked at or analysed real world data, have experienced some degree of missing data. To illustrate the whole idea of artificial completeness, I have acquired a sample dataset that represents what I will have, once all the approvals have been granted.

How complete is the data? Here the vertical axes show the percent of completeness and the horizontal ones show how many fields are present per record.

You can see that all the records have at least one field in them and that there are only about 5% of records which are complete. That's not very much at all!

The problem ctn..

Let's look at this from another point of view, this graph shows you how many values are missing per field, or column. The vertical axes is the percentage of missingness and the horizontal is the column number. As you can see there is a huge amount of independent missingness.

The solution

Imputation, for those of you who don't know, is the process of replacing missing values with some values. There are many different ways to do this.

We have chosen a method called "Multiple Imputation by Chained Equations" or MICE for short. It works by replacing missing values according to what is around them.

What happens is, we have our original dataset, we apply MICE and we get an artificially complete dataset. Brilliant isn't it! Or is it?

How do we know record 327 has values blahhh ??

——-Ask who looks at data with missing values ————

Just because this method worked on one dataset, how do we know it will work on yours?

The idea

The following method, will work on any dataset. It will identify whether MICE or any other imputation method will work on your dataset.

We need a benchmark to test out method against, that means we need a dataset with just complete values in it that could represent our needs. Well, we can create this!

So here's the idea, we have an original dataset with missing data.

The idea (graphs)

Using the graphs we calculated earlier, we can find out exactly what values are missing and where. These give us the missing characteristics of our dataset

The idea with subset

We can great a subset of the original dataset with only the complete values is it, this is our bench mark. Using the missing characteristics, we create x amount of artificially missing datasets.

These datasets are now mini representations from the original dataset. They are little mini-me datasets that behave in the same manner as the original.

The idea ctn

Then we can apply MICE to all of these artificially missing datasets to create artificially complete datasets. This new complete datasets represent what would happen if you applied MICE to the original dataset, except now, we have something to compare then too.

Clustering

Before we continue, let me talk a little about clustering and cluster validation.

——— Talk about multidimensional representation of data and clustering

——- Cluster validation will tell you the characteristics of a dataset in terms of clusters

The idea complete

What have we got now then? We have a complete dataset and we have artificially complete datasets obtained from the complete dataset.

Using cluster validation techniques, we can compare datasets against each other. We can compare their characteristics and find out how close to the truth they are!

If we wanted further verification of this, we can run some sort of regression on each dataset as this will also let us know how close datasets are to each other.

The results of my experiment aren't as important as the notion that this will work in your data. My results mean nothing to