Using Self-Organizing Maps to Analyze the World '95 Data Set

By Anne Bone

Outline

- Problem description
- What is SOM?
- World95 data set
- What I did
- Results
- Review

Problem Description

- There is a data set provided with the statistical software package SPSS called the World95 data set. The self-organizing map (SOM) algorithm was used to see how it would cluster the countries based on the data in the data set.

What SOM?

- The SOM algorithm was first described by Tuevo Kohonen.
- The self-organizing map (SOM) is an artificial neural network algorithm which can be used to cluster multiple variable data sets.
- Self-organizing maps are a way of displaying more complex data in two dimensional hexagonal or rectangular grids.

How SOM Works

- The algorithm first randomizes the nodes of a map of the desired size and shape.
- Then each input of the training data is put in the node it most closely fits, using Euclidean distance, and the surrounding nodes are trained to resemble that piece of data.
- The training is done twice, the second time the surrounding nodes are tuned more finely. Then the map is created.

How SOM Works cont.

- The algorithm requires several different files to run. They are randinit, which does the randomizing, vcal, which does the training, vsom, which does the mapping, and visual.

World95 Data Set

- The World95 data set contains twenty five statistics on 109 countries from 1995.
- The data set attributes include literacy rates, birth and death rates, GDP, population statistics, aids statistics, life expectancy information.
- Also included are fertility rates, what climate and region each country is in, what the major religion and the largest crop grown in each country is, and the average daily caloric intake of citizens in each country.

What I Did

- First, I decided to leave out the three data columns with the most missing data, Daily Caloric Intake, Male Literacy Rate, and Female Literacy Rate .
- Then, I also decided not to use a couple of catagorical columns, major religion and crop, because SOM is mathematically based.
- In addition, I did not use the aids data columns when running SOM.

What I Did

- First I ran through SOM the raw data set, with the hypothesis that SOM would cluster the data based mainly on population values. This is because SOM is sensitive to scalar differences in attributes because it uses Euclidean distance.
- Next, I rescaled the data so it was all between 0 and 1, while maintaining the integrity of the attibutes distributions. Then I ran SOM a second time. I did not really have any idea what the map would look like, but I was particularly curious how different it would be from the first one.

Results

As you can see from the resulting map, the results of running the raw data was as hypothesized.

The results of the second map were very interesting. For the most part, the clustering was very different, although a few countries again mapped to the same nodes.

Review

- Problem description
- What is SOM?
- World95 data set
- What I did
- Results
- Review

Questions

Are there any questions?

References

- Wikipedia, http://en.wikipedia.org/wiki/SOM
- Codebook for World95 data set. Retrieved on 12/9/2010 from http://people.ku.edu/~schrodt/pols706/world95.codebook.html