Anne Bone
12/15/2010
Math 3220

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

Problem Description:

There is a data set provided with the statistical software package SPSS called the World95 data set. This data set includes twenty five attributes for 109 countries from the year 1995. The self-organizing map (SOM) is an artificial neural network algorithm which can be used to cluster multiple variable data sets. This algorithm was used to see how it would cluster the countries based on the data in the data set. The algorithm was run twice, once with the raw data, and once with the rescaled data to see how, if at all, the maps were different.

Analysis Techniques:

The SOM algorithm was first described by Tuevo Kohonen (Self-Organizing Map). Self-organizing maps are a way of displaying more complex data in two dimensional hexagonal or rectangular grids. The places, or nodes, inside the grid are assigned values using a training set and then the data is mapped to the nodes by calculating Euclidean distance of the data and placing it in the node it most closely fits. 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, 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. 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. The easiest way to run the algorithm is to use DOS batch files (Aleshunas, John).

The World95 data set contains twenty five statistics on 109 countries from 1995. The statistics in the data set include the names of the countries, the populations in thousands, the population densities in population divided by square kilometer, the percentage of the population living in urban areas, the largest religion group in each country, the life expectancy rates of males and females, the literacy rates for males, females, and the whole country, and the annual percent growth rate (Codebook for World95 data set). Also included are birth rates and death rates measured by annual rates per 1000 people and the birth to death ratios, as well as infant mortality rates measured by the number of deaths per 1000 live births (Codebook for World95 data set). Other statistics in the set are the GDP per capita, the region of each country, the average daily caloric intake of the people, the number of reported aids cases and the number of aids cases per 1000 people (Codebook for World95 data set). The remaining attributes are the logarithms of GDP, population, and aids rate, and the fertility rates, climate of the countries, and the largest crop grown in each country (Codebook for World95 data set).

Since there were some many data attributes, it was decided to eliminate some to simplify the experiment. A majority of the attributes had few, up to three, or no missing values, so the data was not particularly noisy. Three columns, Daily Caloric Intake, Male Literacy Rate, and Female Literacy Rate, contained only 75, 85, and 85 values respectively, so it was decided to not use them in running SOM because it would be hard

to find the missing values. The most popular religion column was eliminated because it was non-numeric data. The largest crop column was eliminated because there were so many different codes it was practically useless (Codebook for World95 data set). The three columns of aids statistics were also pruned from the data set. This resulted in a data set with 17 attributes being run through SOM.

When running SOM the first time, it was hypothesized that the map would cluster the countries based on population, and perhaps some other variables such as GDP which had the largest scales, because SOM uses Euclidean distance and thus these variable would be have higher weights and influence the map more. In order to see all the countries on the map, the labels that showed up the first time were removed from the labeling set and vcal and SOM mapper were then run again. This was repeated several times. The resulting map from running SOM is on the following page.