<Self-Organizing Map and an Anorectic Data set>
(MATH 3220 -Theresa Helm)

Executive Summary

This experiment involves the use of Self-Organizing Map (SOM) and how well it can predict if a person is anorexic or not, using the given anorectic data set. SOM clusters data based on each nodes vector weight. It is used to visual high-dimensional data with a low-dimensional graph. The anorectic data set that was used consisted of 22 attributes and had 218 people surveyed. The anorectic data set is highly sanitized to help protect the identity of the human subjects.

In order to figure out if SOM is an efficient way to find out if a person is anorectic or not, the original data set was run with SOM. The results showed that the original data set is extremely noisy, which would cause the labeled map to be incorrect.

The noisy data was then taken out of the data set and run with SOM again. The results showed that without the noisy data, the new data set was able to be clustered more efficient.

There were no labels given for this data set. A labeling system was made using the diagnosis column. After adding the labeled set to the data set without noisy data, a SOM was made for the entire data set. The labels that were set for anorectic and non anorectic were clustered together. The results showed that this data set could easily misdiagnose a person

Problem Description

This experiment will see if a mathematical algorithm will support medical research. Can one predict if a person is anorectic or not based on the results from running SOM using the given anorectic data set.

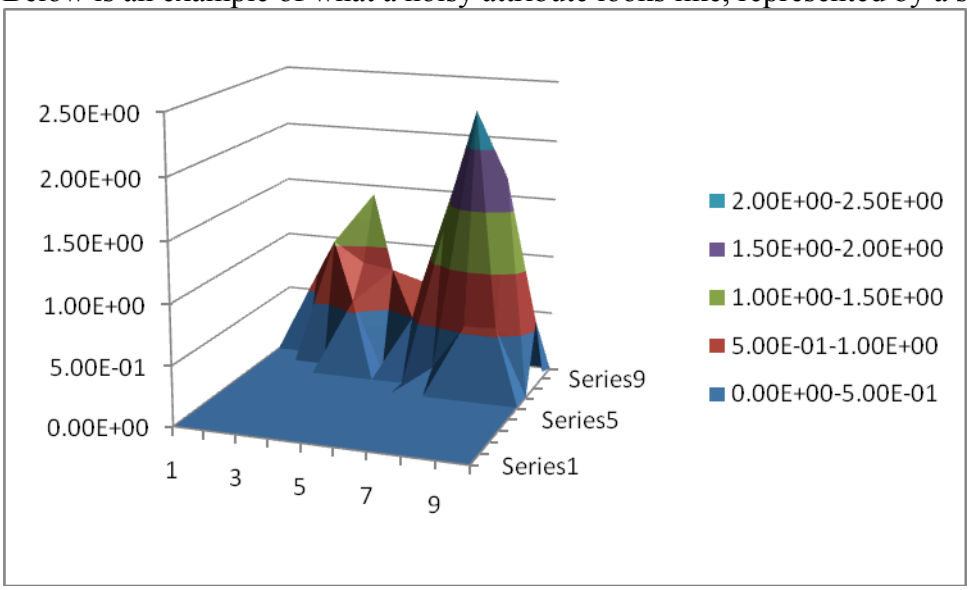
Analysis Technique

In order to find out if a mathematical algorithm supports medical research, the Self-Organizing Map was used. SOM is an artificial neural network algorithm that maps multivariate data into a two-dimensional grid. The resulting map has the property that here is a strong correlation between proximity of the map nodes and similarity of the vectors associated with theses nodes. SOM operates in two modes: training and mapping. Training builds the map using input examples. Mapping automatically classifies a new input vector. A self-organizing map consists of components called nodes or neurons. Associated with node is a weight vector of the same dimension as the input data vectors and a position in the map space. The usual arrangement of nodes is a regular spacing in a hexagonal or rectangular grid. SOM describes a mapping from a higher dimensional input space to a lower dimensional map space. In order to create the map, SOM finds a

node with the closest weight vector to the vector taken from data space and assigns the map coordinates of this node to our vector. It also uses the Euclidean Distance formula to find the vectoring neighbors (Self-Organzing Map, 2010).

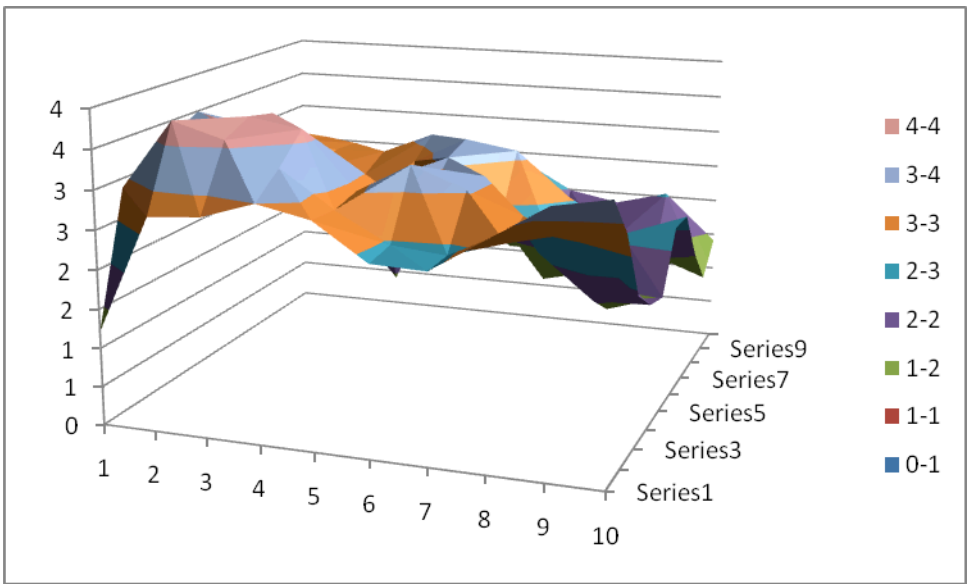
The anorectic data set that was used consisted of twenty-two attributes. It is a highly sanitized data set meaning all personally identifying information has been either removed or replaced with neutral codes so that its public use does not compromise the privacy of the actual human subjects described by the data entries. There were 219 human sources found to enter the data (*MATH 3220 Data Mining Methods)*.

The first step in this experiment was to see if the data set was noisy. The original anorectic data set was run with SOM. After the SOM was ran, a surface graph was made for each individual attribute. The results showed that the original data set is extremely noisy.
Below is an example of what a noisy attribute looks like, represented by a surface graph.



This graph represents attribute 21- diagnosis. This graph is extremly noise, because there is not a clear distrubution for the entries.

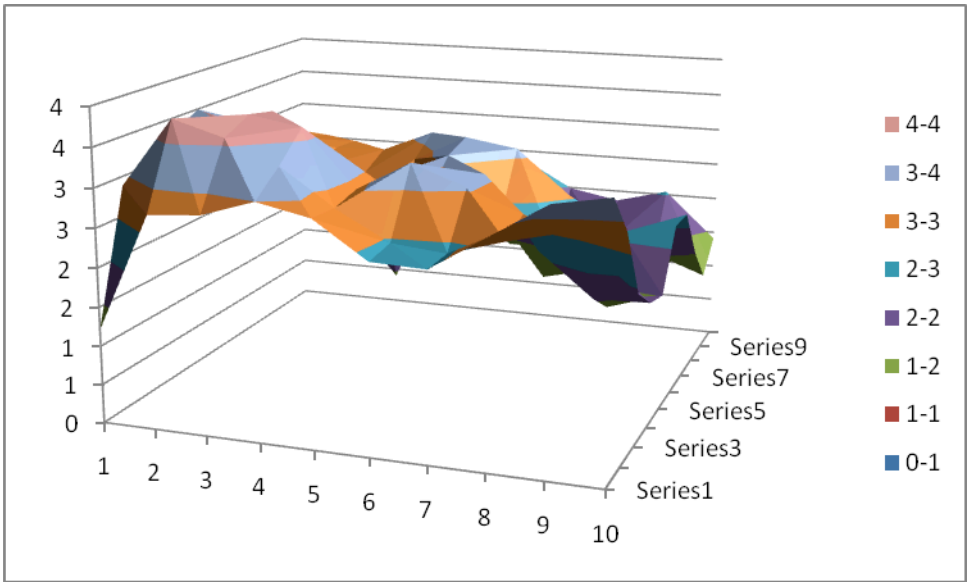
Below is a graph that had less noise, one that would be more optimal.



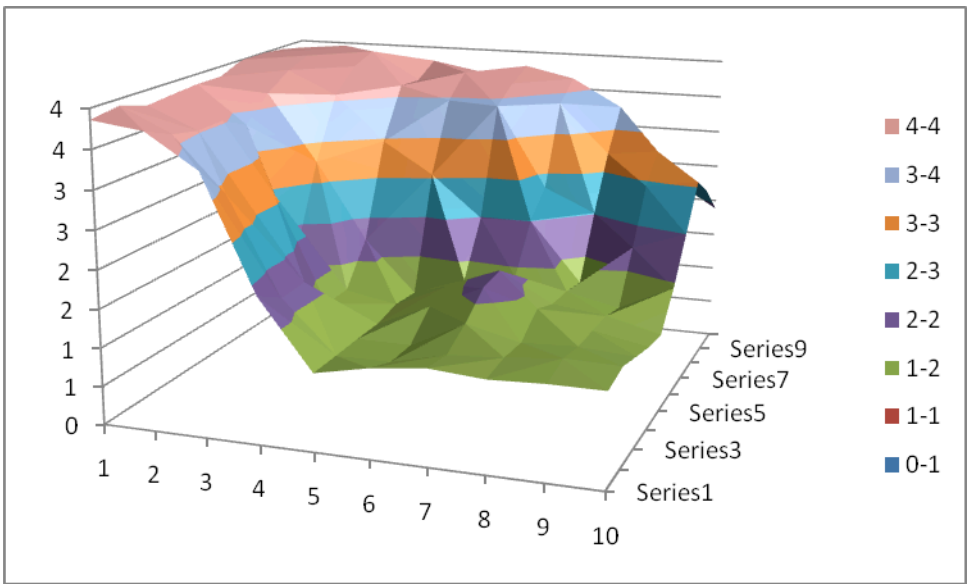
This graph represents attribute 1- Body weight. This graph clearly has a better distribution of data, but is still noisy.

After look at each attribue's surface graph, the noisy data was taken out of the data set. This would be attributes 17-22. After the noisy data was taken out, SOM was ran again, and had a huge impact on each attribute.

Below is attribute number one with 22 attributes.



Below is the same attribute, but ran with only 16 attributes.



There is more of a disbtrution of color, and the graph is more clearly seen.

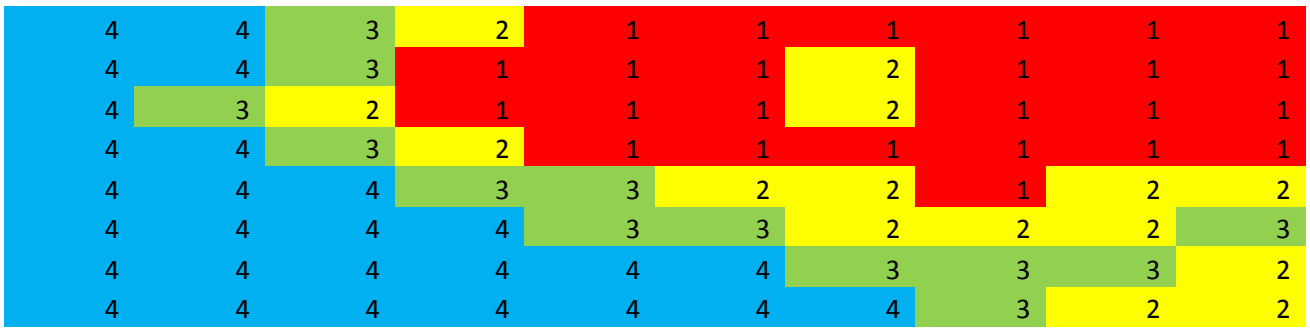
Along with making surface graphs of each attribute, a labeled map was made of each as well. A map was made with 22 attributes and with 16 attributes.

Below is attribute number one ran with 22 attributes.

1	3	3	3	3	2	2	3	3	3
3	4	4	3	3	2	2	3	3	3
3	4	4	4	3	3	3	3	3	2
3	4	4	3	3	3	3	2	2	2
3	3	3	2	1	3	2	2	1	2
3	3	3	3	3	3	3	3	2	2
3	3	3	3	3	3	2	2	2	2
3	3	3	2	2	2	1	1	1	2
3	3	2	1	2	1	1	1	1	1
2	2	2	2	2	2	2	2	2	1

Below is the same attribute, but ran with 16 attributes.

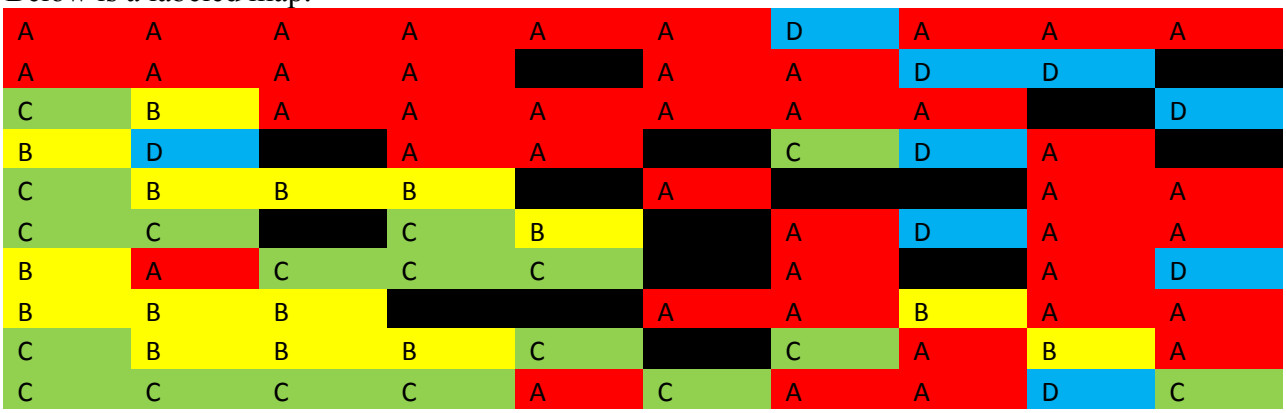
4	4	3	2	1	1	1	1	1	1
4	4	3	2	1	1	1	1	1	1



With only 16 attributes, the data was clustered better together and had more distribution.

After a labeled map was created for each attribute, a labeled map was created for the entire data set (with only 16 attributes). Since the data set was not given specific labels (i.e. anorectic or non- anorectic) labels were made for the data. Using the first diagnosis column, a labeled map was made.

Below is a labeled map.



In the map, 1 was replaced by ‘A’, 2 was replaced by ‘B’, 3 was replaced by ‘C’ and 4 was replaced by ‘D’. The black spaces are the misclassified data. Also, ‘A’ meant non-anorectic, ‘B’ and ‘C’ meant border and ‘D’ mean anorectic. It is clearly seen that ‘A’ and ‘D’ were clustered together and ‘B’ and ‘C’ were clustered together. ‘A’ and ‘D’ should not have been clustered together, but it is expected that ‘B’ and ‘C’ would be clustered together.

Assumption

An assumption that was made was that this data would not be able to cluster. There are many attributes, such as time of interview and patient diagnosis, that did not contribute to a proper way of diagnosing if someone is anorexic or not. By adding the labels, I would be able to get a more defined labeled map, but it turned out that the labeled map is not a sufficient way to cluster this data set.

Results

This is not a sufficient way to cluster this data. Using SOM does not support medical research because this anorectic data set does not have attributes that contribute to if a person is anorectic or not. This data set is also extremely noisy. After taking the attributes that were noisy out, the surface graphs changed; the surface graphs had more distribution. Also, when producing the labeled map for the entire data set, the 'A' and 'D' were clustered together, meaning that a person could easily get misdiagnosed as being anorectic or not anorectic using this data set.

Issues

A big issue that was encountered during this experiment was that there were no labels given. Usually for a data set that is entered into SOM, there are labels given. In order for SOM to cluster correctly, it needs a labeling entry. Also, a lot of the attributes that were in the anorectic data were not attributes that were found in the medical research. Attributes that they could have use would be: BMI (body mass index), age, gender, height, healthiness, and weight. This anorectic data set does have weight, but weight is different for everyone. Someone who is 5'7" is going to have a different weight than someone who is 5'0". This data set also has attributes that are hard to measure on a 1-4 scale. An attribute such as sexual behavior needs an explanation to answer, not a 1-4 scale.

Appendices

Below is the anorectic data labels that was used in this experiment.

	Name	Type	Width	Decimals	Label
1	weight	Numeric	8	0	Body Weight
2	mens	Numeric	8	0	Menstruation
3	fast	Numeric	8	0	Restriction of food intake (fasting)
4	binge	Numeric	8	0	Binge eating
5	vomit	Numeric	8	0	Vomiting
6	purge	Numeric	8	0	Purging
7	hyper	Numeric	8	0	Hyperactivity
8	fami	Numeric	8	0	Family relations
9	eman	Numeric	8	0	Emancipation from family
10	frie	Numeric	8	0	Friends
11	school	Numeric	8	0	School/employment record
12	satt	Numeric	8	0	Sexual attitude
13	sbeh	Numeric	8	0	Sexual behavior
14	mood	Numeric	8	0	Mental state (mood)
15	preo	Numeric	8	0	Preoccupation with food and weight
16	body	Numeric	8	0	Body perception
17	time	Numeric	8	0	Time of interview
18	diag	Numeric	8	0	Patient Diagnosis
19	tidi	Numeric	8	0	Time/diagnosis interaction
20	number	Numeric	8	0	Patient Number
21	diag2	Numeric	8	0	Diagnosis
22	time2	Numeric	8	0	
23					

References

MATH 3220 Data Mining Methods . (n.d.). Retrieved December 1, 2010, from John Aleshunas: http://mercury.webster.edu/aleshunas/MATH%203220/MATH%203220%20Home.htm

Self-organizing map. (2010, December 13). Retrieved December 14, 2010, from Wikipedia: http://en.wikipedia.org/wiki/Self-organizing_map