Applying Self-Organizing Map to Explore MPG

Using the MPG dataset with SOM

Cao Mai

12/14/2009

MATH 3210 - Data Mining Foundations
Professor Aleshunas

Executive Summary

The miles per gallon of 398 cars from various manufactures and models from 1970-1982 along with other attributes were introduced into the self-organizing map (SOM) algorithm to hopefully discover interesting clusters. The self-organizing map is an unsupervised neural network providing a mapping of high-dimensional data into a visual two-dimensional output. Essentially, the algorithm cluster similar input vectors together. After the completion of the analysis, it is recommended SOM should be used for this dataset. Applying SOM to the mpg dataset produce very well defined clusters among virtually all of the attributes. Generally, from the map, cars with more cylinders clustered in areas with lower gas mileage, higher displacement and horsepower, weigh more, accelerate quicker, and to a lesser degree from earlier years. Cars with lesser cylinders clustered oppositely to that of higher cylinder cars for virtually all attributes. The results, however, were not perfect, some misclassifications were observed.

Problem Description

This study examine the miles per gallon samples of 398 cars of various manufactures and models from 1970-1982 along with other attributes in an attempt to discover interesting clusterings. Other attributes within the sample included number of cylinders, displacement, horsepower, weight, acceleration, model year, and origin. The goal is to apply self-organizing map algorithm (SOM) developed by Teuvo Kohonen at Helsinki University of Technology to discover interesting strong clusters. Another part of the goal was to hopefully determine what influence a car’s miles per gallon.

Analysis Technique

The self-organizing map is an unsupervised neural network providing a mapping of high-dimensional data into a visual two-dimensional output. Essentially, the algorithm cluster similar input vectors together.

The process of SOM includes: Initially, the SOM algorithm generates a map and randomizes the weights for all neurons on the map so that every neuron has a different set of starting weights from every other neuron. Each neuron has one weight for every attribute in the dataset. For example, the dataset used in this paper has 6 attributes so initially each neuron on the map would have 6 randomized weights. Then, SOM is trained with the input data. An input vector is presented to the map and one of the neurons, and only one, will have weights that are the closest match to that input vector. Often, first they may not be very similar at all; nonetheless, one of the neurons will still be closest. The closest matched neuron raises a flag and "captures" the vector. This closest "capture" is commonly calculated based on Euclidean distance in SOM; it is determined by calculating the distance between the input vector and every neuron in the network, and the lower the distance between the neuron and the input vector the more likely the neuron will "capture" that vector.

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}.$$

(Wikipedia, 2009)

The captured neuron then adjusts its weights to be even closer to the actual values of the vector, but never exactly like the input vector. SOM also adjusts the neighbors of the capturing neuron, so that they are also more similar to the capturing neuron; however, they are adjusted only part way, not to become the identical to the capturing neuron. At this point the SOM is ready for another vector, and the training continues. As training continues, neighborhoods are adjusted to become more similar, and different instances are “attracted” to different neighborhoods (Pyle, 2003). A SOM can be consider as a “rubber surface” that is stretched and bent all over the input space, so as to be close to all the training points in that space.

For the mpg dataset, some preprocessing of the data was done before it was introduced to the SOM algorithm. The “origin” attribute was taken out because the description did not mention much what the attribute was about. It was also discarded because most of the instances had 1s, while 2s and 3s appeared sporadically. There were some instances (6), that had missing values for the attribute “horsepower” and those were taken out because SOM cannot handle missing data. The attribute “cylinder” was used to label because more cylinder is known to inversively affect gas mileage.

The dataset was also prepared for SOM because SOM only reads space-delimited files. The testing parameters for SOM applied to the dataset were as follows:

- First ran with an epoch of 1,000, training factor of 0.05, and a radius of 10.
- Second ran with an epoch of 10,000, training factor of 0.02, and a radius of 3

Vcal was used to apply labels to the map and SOM_mapper was used to generate a visual map. SOM_mapper was also used to generate attribute values from the map (Aleshunas, 2009).

Assumptions

- No invalid data
- SOM clusters similar instances and generates a visual map
- The algorithm preformed correctly.
- The data introduced to the algorithm is correct

Results

Applying SOM to the mpg dataset produce very well defined clusters among virtually all of the attributes. Generally, from the map, cars with more cylinders clustered in areas with lower gas mileage, higher displacement and horsepower, weigh more, accelerate quicker, and to a lesser degree from earlier years. Cars with lesser cylinders clustered oppositely to that of greater cylinder cars for virtually all attributes.

These results were not too surprising because it is widely known that engines with more cylinders will weigh more be bigger and need a longer car to accommodate for the large engine, have higher horsepower, accelerate quicker and as a result have lower gas mileage than engines with lower cylinders.

Although the map produced very well defined clusters, the results are not perfect because there are instances where six cylinder engines grouped in regions of higher gas mileage than four cylinder engines. There are also instances where six cylinder engines grouped in regions with lower gas mileage than eight cylinder engines. Acceleration behaved similarly, some six cylinder instances clustered in regions of higher values than four cylinders instances. In addition, from the map, it can be said that the acceleration of six cylinder and four cylinder cars don't significantly differ because from the map six cylinder cars accelerate only about one second faster than four cylinder cars. Eight cylinders, however, accelerate significantly faster than all others, about 2-3 seconds faster than six cylinder cars and 2.5-4 seconds faster than four cylinder cars.

Issues

- Ignore missing data
- Dataset description not compete
- SOM only uses categorical data

Appendices

Dataset downloaded from:

http://mercury.webster.edu/Aleshunas/Data%20Sets/Supplemental%20Excel%20Data%20Sets.htm

SOM programs and SOM_mapper downloaded from:

http://mercury.webster.edu/Aleshunas/Source%20Code%20and%20Executables/Source%20Code%20and%20Executables.html

Results Attached

Reference

Aleshunas, John. Source Code & Executables. “Self-Organizing Map (SOM)” Retrieved: 2 Dec 2009 <http://mercury.webster.edu/Aleshunas/Canary/Canary_Home.html>

Pyle, D. (2003). *Business Modeling and Data Mining.* San Francisco: Morgan Kaufmann.

"Self-organizing map" Wikipedia, The Free Encyclopedia. Retrieved: 2 Dec. 2009, Wikimedia Foundation, Inc.. <http://en.wikipedia.org/wiki/Self-organizing_map>