<SOM Network>

(Khoa Doan)

Executive Summary

The University of Florida wants to see if there is any pattern in its graduate students' characteristics from 8 majors in a dataset of 1100 students. The characteristics are the students' gender, salary, and graduation date. In the finding process, the characteristics are passed through a Self-Organizing Map (SOM) network. The SOM network uses the SOM algorithm that accepts the input information about each student, processes the information and maps the input to a 2-dimensional m x n grid. Inputs with similar characteristics will be mapped closely to each other on the grid. After running the whole dataset through a SOM network with a 10 x 10 grid, we have the grid as in Map 1. Based on Map 1, we have concluded that about 60% of agriculture student have the same characteristic because there's a very clear cluster that has 60% of agriculture students. Aside from this cluster, we cannot say anything on the remaining part, since the majors' mappings are very mixed together.

Problem Description

The University of Florida has collected data about its graduate students and their salaries. The dataset comprises 1100 records about the students' gender, college, salary, number of degrees and graduation date. Based on this dataset, the university wants to see if there is any cluster of students that can be served as a representative for one of the eight colleges. The finding will base on three attributes – gender, salary, and graduation date.

Analysis Technique

The chosen approach is the Self-Organizing Map (SOM) algorithm. It's chosen because we already have access to the SOM program. In addition, the algorithm produces an output that's easy to visualize and explain.

The SOM algorithm is actually a neural network which consists of two layers of nodes (as depicted in Figure 1). A node is a processing unit that accepts one or more inputs and produces an output. The first layer of nodes is the input layer, which simply receives an input vector and forwards it to the second layer called competitive layer. Each node in the input layer is connected to each node in the competitive layer by an arc that is labeled with a weight. A weight is a quantity that indicates the importance of each connection from the input nodes to the competitive nodes. At the competitive layer, each node produces an output that will compete with others to determine the final output of the network. The competitive is based on the Euclidean distance between the input and weight vectors to each competitive node. The final output will be at the node with the smallest distance. The learning process then takes place by adjusting the weights so that the best output is even better the next time the same input vector is used.

Revised: 13 December 2003

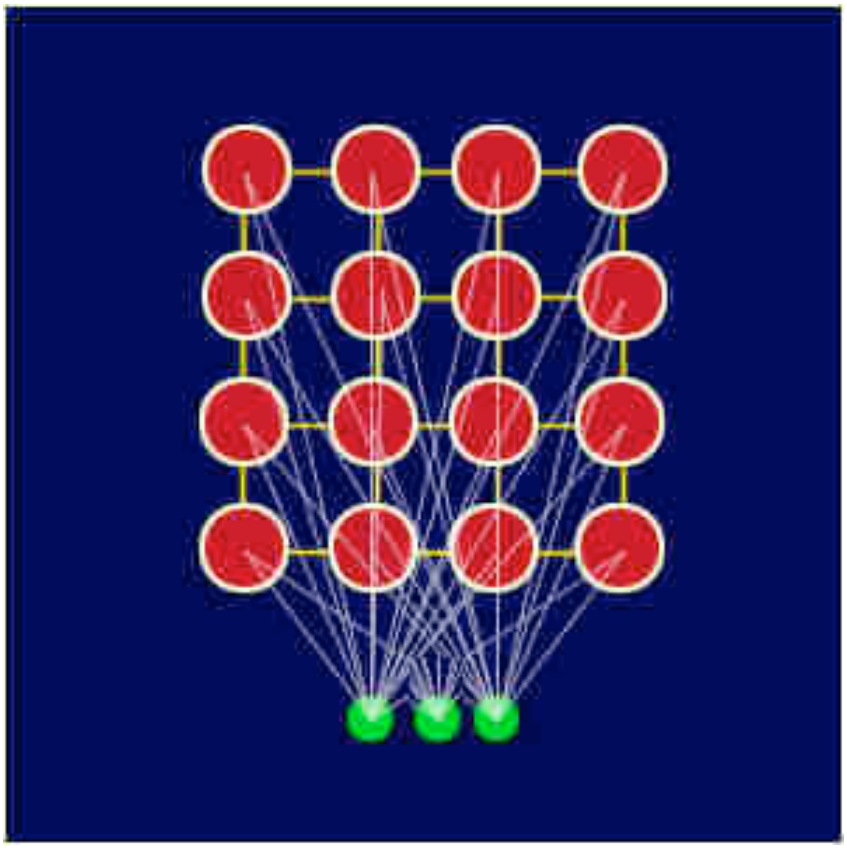

Figure 1

The SOM application used to cluster the dataset has four small programs, which are randinit.exe, vsom1.exe, vsom2.exe and vcal.exe. The randinit.exe initializes a 2-dimension grid with a user-defined size and arbitrary values of the weights from the input layer to each grid node. The vsom1.exe and vsom2.exe then performs the training in 1000 steps and the vcal.exe generates the visual mapping.
The training is performed with a grid size of 10x10 and the mapping is described in Map 1.
Map 1 shows that there's really a cluster containing 60% of Agriculture students with a confidence of above 80% (there are 4 other-major cells and 20 agriculture cells). This suggests that students of agriculture major will most likely have characteristics that can be distinguished from other majors. Aside from this cluster, the remaining map doesn't show clearly any other cluster of any major.
Map 1 also raises a question the effectiveness of the algorithm. There are 8 majors, and there are only 4 of them on the grid. One possible reason is that the other majors have very similar characteristics to the major on the map, thus are mapped to the same location and overridden by the four majors on the map. Therefore, another training is constructed with a grid size increasing to 20x20 as in Map 2. Because Map 2 still yields only 4 majors, we can conclude that the suspected reason is highly correct.