Escuela Superior de Cómputo
Instituto Politécnico Nacional
Ingeniería en Sistemas Computacionales

Evolutionary Computing

Practice 8: Self Organizing Maps

Professor: Jorge Luis Rosas Trigueros
Student: Ayala Segoviano Donaldo Horacio
Creation date: November 2 2021
Delivery date: November 12 2021

1 Introduction

1.1 Artificial Neural Networks

A **neural network** is a massively parallel distributed processor made up of simple processing units that has a natural propensity for storing experiential knowledge and making it available for use. It resembles the brain in two respects:

1. Knowledge is acquired by the network from its environment through a learning process.

2. Interneuron connection strengths, known as synaptic weights, are used to store the acquired knowledge.

Neural Networks models can be classified into three categories:

- **Feedforward networks:** transform sets of input signals into sets of output signals. The desired input-output transformation is determined by external, supervised adjustments of the system parameters.

- **Feedback networks:** the input information defines the initial activity state of a feedback system, and after state transitions the asymptotic final state is identified as the outcome of the computation.

- **Competitive, Unsupervised or Self-organizing network:** neighboring cells in a neural network compete in their activities by means of mutual lateral interactions, and develop adaptively into specific detectors of different signal patterns.

1.2 Self Organizing Maps

The **Self-Organizing Map (SOM)** is a computational method for the visualization and analysis of high-dimensional data, especially experimentally acquired information. The Self-Organizing Map defines an ordered mapping, a kind of projection from a set of given data items onto a regular, usually two-dimensional grid. SOM's uses an unsupervised learning process.

A model m_i is associated with each grid node. These models are computed by the SOM algorithm. A data item will be mapped into the node whose model is most similar to the data item, i.e., has the smallest distance from the data item in some metric.

When the models are computed by the SOM algorithm, they are more similar at the nearby nodes than between nodes located farther away from each other on the grid. In this way the set of the models can be regarded to constitute a similarity graph and structured 'skeleton' of the distribution of the given data items.

1.3 World Bank Database: World development indicators

World Development Indicators (WDI) is the primary World Bank collection of development indicators, compiled from officially recognized international sources. It presents the most current and accurate global development data available, and includes national, regional and global estimates.

For this practice, series of data from the 2015 year in the World Bank databases will be used to create a self organizing map out of some variables of some countries.

2 Material and equipment

Following are the hardware and software used during the realization of this practice. ***Google Colaboratory*** was the tool used for the development of this practice and the next list shows the specifications of the hardware and software provided by Google Colab.

- **Hardware:**
  - **CPU:** Intel(R) Xeon(R) CPU @ 2.30GHz
  - **Memory:** 12GB
  - **Disk:** 108GB
- **Software:**
  - **Platform used:** *Google colaboratory* was used for this practice
  - **Programming language:** Python 3.7.11

3 Development

3.1 Data extraction and formatting

The first part of this practice corresponds to data extraction from the World Bank's world development indicators databases. This is going to be documented since a process of selection and formatting of data had to be done so that it could be provided to the SOM for training.

As mentioned, the information used belongs to the world development indicators World Bank's databases and from the year 2015. The only issue with this database is that not all of the countries listed have available all the information, some countries may have some information that is missing for another. One solution for this could be to give default values for the missing fields of information, however, this could cause the SOM not to be too reliable since the algorithm needs all the information and it would change the structure of the map if we set some defaults or alter the information.

Instead, to solve the problem of missing information for some countries, only variable and countries which had available information were selected. To do this, first, all the data available was extracted from the year 2015. Then, around 55 countries of interest were selected as well as around 16 variables that could be critical indicators of world development. Once selected, the complete data was filtered so that it only extracted the variables selected from the countries selected.

Once the information about these countries was selected, it was analysed how many indicators were missing, and which countries were missing information, all this was done with the help of a python script. Once identified, the variables that were missing in most of the countries were discarded, and also, countries that were missing many variables were also discarded.

After the data selection and data filtering and selection, 53 countries were selected along with 12 indicators per country, all 53 countries had full information about the 12 variables so that the SOM could be trained and tested with complete information.

3.1.1 Implementation and testing

As mentioned in the introduction, a self organizing map will be built using influence from neighbors on a grid so that similar nodes can be grouped closer. The following describes the algorithm to build a Self Organizing Map.

1. Randomize the map's nodes' weight vectors, $s = 0$.
2. Traverse each input vector in the input data set $D(t)$
   (a) Traverse each node in the map
   (b) Use the *Euclidean distance* formula to find the similarity between the input vector and the map's node's weight vector.
   (c) Track the node that produces the smallest distance (this node, at position u, is the best matching unit, BMU).
   (d) Update the nodes in the neighborhood of the BMU (including the BMU itself) by pulling them closer to the input vector, for each neighbor at position v: $W(s+1) = W(s) + T(u, v, s) * a(s) * (D(t) - W(s))$
3. Increase s and repeat from step 2 until termination criterion is met.

- u is the position of the BMU.
- v is the position of a neighbor to the BMU.
- $W(s)$ is the weight vector in position v.
- $T(u, v, s)$ is a function that controls the effect of $D(t)$ over W in v.
- $a(s)$ is a function that usually fades as s increases.

The code will be implemented in Python programming language to generate the self organizing map for countries considering variables from the World Bank's World development indicators database. The implementation of the self organizing map can be found by clicking **here**.

Now, the information extracted from the database is going to be provided to the SOM to train it. From the 53 countries, only 45 were used to train the SOM, and at the end, the best matching unit for each of the 53 countries will be calculated and represented in the map.

For the first run, 500 iterations were used, 45 information vectors were provided to the SOM and the SOM has dimensions of 10 rows and 10 columns so that the 53 countries can be distributed in the map, the SOM was trained with 300 iterations and a learning rate of 0.005.

Country	Abbreviation	Country	Abbreviation
Afghanistan	AFG	Jamaica	JAM
Argentina	ARG	Japan	JPN
Australia	AUS	Malaysia	MYS
Belgium	BEL	Netherlands	NLD
Bolivia	BOL	Nicaragua	NIC
Brazil	BRA	Nigeria	NGA
Canada	CAN	Norway	NOR
Chile	CHL	Pakistan	PAK
China	CHN	Panama	PAN
Colombia	COL	Paraguay	PRY
Costa	Rica	Peru	PER
Croatia	HRV	Poland	POL
Cuba	CUB	Portugal	PRT
Denmark	DNK	Qatar	QAT
Ecuador	ECU	Romania	ROU
El Salvador	SLV	Russian Federation	RUS
Finland	FIN	South Africa	ZAF
France	FRA	Spain	ESP
Germany	DEU	Switzerland	CHE
Greece	GRC	Thailand	THA
Guatemala	GTM	Turkey	TUR
Iceland	ISL	Ukraine	UKR
India	IND	United Kingdom	GBR
Iraq	IRQ	United States	USA
Israel	ISR	Uruguay	URY
Italy	ITA	Mexico	MEX
Vietnam	VNM	-	-

Table 1: Countries in the SOM along with the name abbreviation.

Table 1 shows the 53 countries considered for the construction of the self organizing

map along with their abbreviations, which are going to be used to represent the country in the map's positions.

The following are the variables considered for the construction of the SOM. These were chosen because those were indicators available for all the countries selected and because in a personal opinion, these are important characteristics in a country and they can help to group similar countries:

1. Individuals using the Internet (% of population)
2. Population density (people per sq. km of land area)
3. GNI (constant 2015 US$)
4. Suicide mortality rate (per 100,000 population)
5. Armed forces personnel, total
6. CO2 emissions (kg per 2015 US$ of GDP)
7. Population, male
8. Population, female
9. Scientific and technical journal articles
10. Current health expenditure (% of GDP)
11. Population growth (annual %)
12. GNI per capita (constant 2015 US$)

Given this information, the algorithm will be run to see how the map organizes the information and groups the different countries. The data was normalized, i.e. for each variable, the maximum value was found, then, al the values corresponding to the same variable, were divided by this maximum value. Normalization was done because when providing non-normalized data, strange results were obtained. The information fed to the SOM can be found in a CSV format by clicking **here**.

-	IRQ	AFG	-	ZAF	UKR	RUS	-	-	CHN,IND
NGA,PAK	-	-	BOL	-	-	-	-	-	-
-	GTM	NIC	-	-	-	MYS	-	-	-
-	-	-	PER	-	ECU	-	-	JPN	-
SLV	JAM,VNM	TUR	PAN,PRY	-	CRI	-	GBR	-	BEL,NLD
-	THA	COL,MEX	-	ARG	CHL	-	DEU,USA	-	-
CUB	-	-	BRA	-	URY	FRA	-	-	ISR
-	-	ITA	-	-	-	-	DNK	-	QAT
-	PRT	-	ESP	-	CAN	AUS,ISL	-	CHE	-
HRV,POL,ROU	-	GRC	-	FIN	-	-	NOR	-	-

Table 2: Resulting SOM for the provided data.

Table 2 shows the self organizing map obtained after the first run of the algorithm. One can analyse the grouping of the countries to see that there are some similarities with the world's situation.

One of the clearest observations is that countries like Deutschland and the USA are classified into the same position int the map, and one can draw a diagonal with the countries of Finland, Canada, France, Deutschland and the USA, the United Kingdom, Japan and China and India, to make a separation between the countries with the highest development (to the right and down of the diagonal) and the countries with a mid to low development level (to the left and above the diagonal).

In the middle area of the map, many Latin American countries with can be spotted, such as Colombia, México, Panama, Paraguay, Peru, Argentina, Brazil, and so on, these could be associated with mid level development. And in the leftmost upper and lower corners, appear countries that could be associated with a low development, because there are countries such as Pakistan, Afghanistan, Guatemala, El Salvador, Cuba, Thailand, Vietnam, among a couple others.

Another observation, is that looks like the countries are grouped according to geographical locations, since countries in Latin America are grouped very closely, also, China and India, which are neighboring countries, are placed together in the map, the same repeats with Russia and Ukraine.

There are different characteristics by which one could associate the grouping of the countries, the main one is the development, since those are the variables that were considered for this SOM. So, it can be concluded that the SOM is working and that the results are congruent.

4 Conclusions

During the development of this practice, the concept of Self Organizing Map was comprehended as well as the its construction and utilities. Also, neural networks were explored when investigating about the SOM, and it was comprehended why a SOM is considered a neural network and also, the different techniques to model a neural network.

It was concluded that Self Organizing Maps can help to identify patterns in big chunks of information, and also, they can help to visualize data in a more comprehensive way. It was also concluded that SOM's are some kind of clustering, since they can separate data with similar characteristics into groups, and also, they can help to make predictions about new elements that are not in the map.

The only disadvantage that was identified about the a SOM was that the complexity for building has a big cost, since for each node, it is necessary to iterate over all of the elements in the grid several times, leading to a big cost in computational time.

5 Bibliography

- **Kohonen T., Honkela T. (2007) Kohonen network. Scholarpedia.**
- **Kohonen T. (1990). The self organizing Map. Proceedings of the IEEE.**
- **Kaykin Simon. (1999) Neural Networks and Learning Machines. Pearson Prentice Hall. USA.**
- **Rosas Jorge. (2021) Self Organizing Maps Lecture. México.**