Study of Self-Organizing Maps (SOM) Algorithm

Author: Anshul Rao

Summary:

A self-organizing map (SOM) is an unsupervised machine learning technique used to produce a low-dimensional (typically two-dimensional) representation of a higher dimensional data set while preserving the topological structure of the data. For example, a data set with *p* variables measured in *n* observations could be represented as clusters of observations with similar values for the variables.[1]

As part of this project, I will learn about the SOM algorithm and convey the understanding using illustrative examples. I will also apply the algorithm on a textual dataset to demonstrate how it works and discuss its latest applications and usages.

Proposed Plan:

1. Explanation of how the algorithm works with the help of illustrative examples. This will also cover information about the advantages and disadvantages of the algorithm, highlighting in which circumstances it might be a good idea to use it.
2. Creation of a demo project in Python that uses this algorithm to create clusters in a textual dataset. I plan to use the data I downloaded from StockTwits (a social media platform designed for sharing ideas between investors, traders, and entrepreneurs) using a Python software I built. The data consists of tweets from different users for various stocks. For text clustering, I will start with cleaning/preprocessing the data which will include removing stopwords, stemming/lemmatization, etc. After this I will transform the textual data to numerical vectors using TF-IDF and then proceed with clustering. There are two SOM Python libraries that I wish to explore for this project, somoclu[4] and sklearn_som[5].
3. Discussion on some recent applications and usages of the algorithm. For example, in one academic paper a deterministic SOM was proposed to eliminate the randomness of the standard self-organizing map.[2] In another paper, to tackle catastrophic forgetting of neural networks a memoryless method that combines standard supervised neural networks with self-organizing maps was proposed. The role of the self-organizing map was to adaptively cluster the inputs into appropriate task contexts – without explicit labels – and allocate network resources accordingly.[3]

Future Scope:

Based on how the project progresses and the results I obtain, I would like to move in the direction of how SOM can be enhanced to perform computationally inexpensive and efficient clustering of textual data.

References:

1. Wikipedia: Self-organizing map, https://en.wikipedia.org/wiki/Self-organizing_map
2. A Deterministic Self-Organizing Map Approach and its Application on Satellite Data based Cloud Type Classification, https://arxiv.org/pdf/1808.08315.pdf
3. Continual Learning with Self-Organizing Maps, https://arxiv.org/pdf/1904.09330.pdf
4. somoclu: https://somoclu.readthedocs.io/en/stable/
5. sklearn-som: https://pypi.org/project/sklearn-som/
6. Research of fast SOM clustering for text information, http://140.127.22.205/100-2/ai/papers/SOM/a3.pdf
7. Improved SOM Algorithm-HDSOM Applied in Text Clustering, https://ieeexplore.ieee.org/document/567083