International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 10 Issue 12, December -2020,
ISSN(O): 2249-3905, ISSN(P) : 2349-6525 | Impact Factor: 7.196| Thomson Reuters ID: L-5236-2015

CLUSTERING LARGE DATA SETS VISUALISING THE CLUSTERS USING SELF ORGANIZATION MAP

Manoj Kumar, Assistant Professor, Department of Computer Science, Government College, Narnaul, District Mahendergarh, Haryana.

Email: manojrathee.professor@gmail.com

Parveen Gorya, Assistant Professor, Department of Computer Science, Government College, Narnaul, District Mahendergarh, Haryana.

Email: parveen05cse34@gmail.com

ABSTRACT

Some schools may find it challenging to categorize children, especially when they have a large number of new students every autumn. In such cases, it's important to reconsider the project. One potential solution for schools is to use a decision support system that can automatically create groups of students based on their names. Self-organizing maps (SOMs) are a type of unsupervised learning that utilize artificial neurons to classify data. In this research project, the focus will be on classifying incoming high school students based on their previous academic achievements. The students' final scores will be determined by considering their report books and performance in earlier national tests. Currently, the fields of biological science, social science, and language studies each have their own separate group of professionals. The aim of this research project is to develop a decision support system that can assist schools in classifying incoming high school students based on their academic achievements. This system will use a self-organizing map algorithm, which is a type of artificial intelligence that can automatically group students based on their performance and other relevant information.

Keywords: *Clustering Large, Clusters Using, Visualizing, Classification.*

INTRODUCTION

At the institutions that provide the conventional secondary education in Indonesia, students who are applying to schools in Indonesia, as well as the schools themselves, have access to a broad selection of academic tracks from which to pick. Ability grouping, also known as achievement grouping, is a method that is used in most schools to organize students into classes based on their previous academic performance. Ability grouping is a strategy of arranging students in the classroom in such a way that those students who have achieved greater levels of academic performance are grouped together, while those students who have achieved lesser levels of success are grouped together. Report cards have the potential to be useful for both

intelligent and non-intelligent selection processes. The standard procedure includes a teacher picking a small number of kids from one class who have shown exceptional academic achievement and pairing them with students from other classes who have demonstrated academic achievement that is comparable to their own. The implementation of this clustering strategy has as its primary objectives the improvement of student performance, the simplification of the instructional process, and the provision of instructors with a greater degree of control over the quantity of information acquired by their pupils. The level of success that students have in their academic pursuits may be indicative of their potential in a variety of other aspects of their lives and careers. As a consequence of this, pupils are often sorted into a few distinct groups according to the results they achieved in the aforementioned areas. This task to categorize children may prove to be too much for certain schools, particularly those that welcome a significant number of new pupils each autumn. If this is the case, the project should be rethought. It is feasible that the teaching staff at a school might benefit from adopting a decision support system that has the capability of automatically creating groups of students based on a list of their names.

Self-organizing maps, often known as SOMs, are a kind of learning known as unsupervised learning. SOMs use a network of artificial neurons to reduce the dimensionality of the input. Dimensionality reduction is a typical application of this idea, and it often involves the usage of Kohonen Maps or networks. SOM employs somewhat different from that of traditional ANNs. In contrast to this, artificial neural networks (ANN) use a kind of learning that is driven by competition to improve over time and eliminate errors. The SOM makes use of a neighborhood function in order to make an attempt toward maintaining the underlying topological structure of the input space. The theoretical underpinnings of artificial intelligence are rapidly finding usage in applications that take place in the actual world. Categorization, grouping, and associating are practices that have been around for quite some time, and they continue to be a driving force behind the development of both supervised and unsupervised learning systems. In order to accurately forecast the marks that students would get at the conclusion of the semester, we used a neural network model and coupled Linear Regression and Support Vector Regression. As a result, we were able to forecast the performance of the students in the second semester by looking at their input grades from the first. An application that allows for unsupervised learning of SOMs has been created. By offering a more in-depth description of the students' cognitive structural models, this program aims to achieve its goal of more accurately classifying pupils for the purposes of a scholarship system. The students' reasons for wanting to learn mathematics were investigated by putting them into study groups with others who had similar interests and goals.

The level of academic achievement that the children have shown in the past will be taken into consideration when classifying them. The selected grades were determined using the outcomes of the national examinations as well as the rapport books from the prior study.

Self-Organizing Maps

Self-Organizing Maps, often known as SOMs, have been in use ever since 1996, when Teuvo Kohonen thought of them for the first time. SOM helps enhance the data visualization process by translating the data from a high-dimensional to a low-dimensional domain. Self-organizing neural networks are used in order to accomplish this goal. SOM is an option to consider when working with unsupervised learning data since it presupposes that the network a result of its application. This is correct due to the fact that the weight vector is a matrix. The winning group of units is the one whose weights are most similar to the input vector pattern (typically as assessed by the square of the lowest absolute value of the Euclidean distance). This similarity is a measure of how closely the weights align with the vector pattern.

The make little adjustments to their weights. Target neurons in SOM networks are dispersed in a two-dimensional area whose shape may be altered, as opposed to being lined up as they are in other ANN models, where they are arranged in a linear fashion. This is a departure from the way target neurons are configured in the majority of conventional ANN models. These characteristics differentiate this ANN model from others that came before it. When a large number of winners are surrounded by neurons of varied shapes, the weights that are produced are likely to be very distinct from one another. The SOM is responsible for rebalancing not just the connection weights of the winning neurons and their neighbors, but also the weights of the individual neurons.

The typical designs that are used to implement SOM may take on a wide variety of various shapes, and each of these forms has a unique set of characteristics as well as benefits. Two of the distinguishing characteristics of these systems are the shape of the map and the techniques used to estimate closeness. A technique for moving Every neuron in the output layer stands in for a different set of neurons in the input layer. In a manner that is similar to that of Principal Component Analysis (PCA), SOM is often used in order to lessen the overall number of dimensions. This is as a result of the fact that SOM is capable of performing the same tasks. However, the reduction in dimensionality might be construed as a reduction in the number of clusters; for this reason, SOM is listed among the clustering techniques. As a consequence of this, SOM is able to bring the total number of clusters to a lower value.

1. First things first, you'll need to decide on the general size and shape of the map. At the beginning of the process, the inputs are more significant than the outputs.

2. A random number generator whose output is scaled to the size of the map should be used to choose some samples at random from the data.

3. Third, with the help of the map, locate the node that is geographically closest to the input.

4. Perform an update to the weight, which will redistribute the neighbors of the closest node depending on the weight of each neighbor.

5. Continue to repeat step cycles have been completed), and then go on to step 3.

OBJECTIVES

1. To Study Self -Organized Map (SOM) method for clustering the DJIA and NASDAQ100 portfolios for determination of non-linear correlations between stocks.

2. To Study Self-organizing maps, often known as SOMs, are a kind of learning known as unsupervised learning

RESEARCH METHODOLOGY

As our primary source of information, we focus on the results of previous examinations in the subject areas of mathematics, physics, English, Indonesian, and the social sciences. Two of the five pieces are called the connection component and the testing results portion, respectively. The functioning of the system as well as its flowchart are both broken out in great depth. When building a functional version of the SOM algorithm, the programming language Python was selected as the one to utilize.

The numbers in the interval were normalized once they were input into the system. Assigning initial random weights, deciding on the size of the map, and setting a maximum limit on the number of potential iterations are all components of the technique known as "initialization." SOMs may be implemented in a large variety of different methods, some of are a plethora of additional possibilities. The research for this study was carried out using a grid that was 10 squares wide and 10 squares deep. A single data point with certain random weights is stored in each map cell. The size of the data point are 1 by 9. There was a limit of 5,000 potential iterations placed on the total number of repetitions that might take place.

Nearest node computation

The input that was sampled was compared to each candidate unit, and between it and the other unit. The result that was delivered to you is an array that has rows and columns, and each row and column has the map's Best Matching Unit (BMU) value.

Weight updating

The BMU node's weight is altered in compliance with the weight update rule (formula 1), which then affects the weights of the BMU node's four nearest neighbors.

$$w_j(t+1) = w_j(t) + \alpha(t)[d \quad [t] - w_j(t)]$$

$$p \qquad [t] = 1 - \left(\frac{s[t]}{m}\right)$$

$$\alpha(t) = p \qquad [t] * \alpha$$

DATA ANALYSIS

The first table presents some of the information included in the whole dataset. The information was first brought into a range from 0 to 1 for the purposes of analysis and then normalized.

The dimensions of the map were set to be 1010, and 9 values were inserted into each cell. In addition, arbitrary weights were assigned to the map's features. Table 2 displays the random weight assignments that were made to each cell in the study.

Table 1. Cell weights in the SOM map that are random

4.170220 05e-01	7.203244 93e-01	1.143748 17e-04	3.023325 73e-01	1.467558 91e-01	9.233859 48e-02	1.862602 11e-01	3.455607 27e-01	3.967674 74e-01

Following the determination of the BMU value, each iteration's weights for the four nodes that were near to one another were adjusted using formula (1). Table 3 displays the equilibrium cells' weights.

Table 2. final cell weights in the SOM map

0.371248 28	0.351578 07	0.374648 3	0.366969 35	0.361304 72	0.508267 86	0.594269 58	0.599300 84	0.746200 43

After 5000 repetitions, Assigning colors and grayscales to nodes in a SOM for the sake of visual representation requires the usage of a unified distance matrix, often known as a U-Matrix. Visually, nodes with similar relationships are grouped together and given the same color. The 10x10 built SOM has two U-matrices: a color U-matrix (3(a)) and a monochrome U-matrix (3(b).

It's not hard these towns having a population that is noticeably higher than that of the others. The outcome is compatible with the approach of separating students into three distinct departments in high schools, since this is a practice that is often followed. These departments are known as Bahasa, IPA, and IPS, which stand for language, natural science, and social science respectively. When it comes to the number of college students who decide to concentrate their studies on a particular subject area, the natural sciences are among the most popular choices.

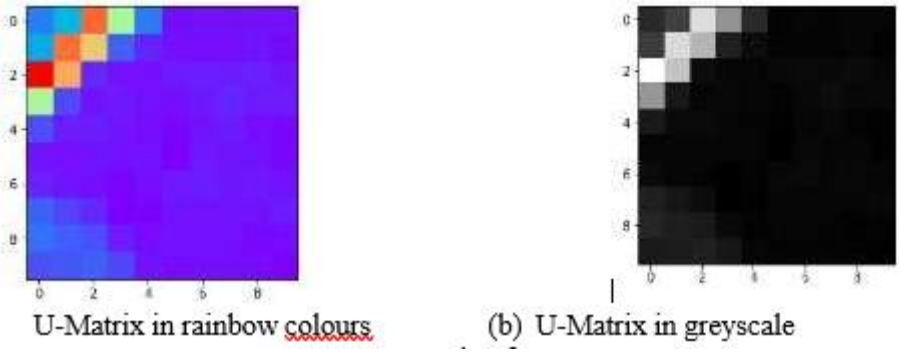

Figure 1. U-Matrix of 10x10 SOM

CONCLUSION

This study's categorization of first-year students according to their academic ability has contributed to shedding some light on the statistics of first-year students' grade point averages. There is sufficient space available to accommodate three separate clusters, which accurately reflects the traditional division between the social sciences, the natural sciences, the vast

majority of the content and provide the greatest amount of variety. The final scores of the students were determined based on the information in their rapport books as well as how well they performed on earlier national tests.

REFERENCES

1. A. Imron, "Management of School-based Students" (Manajemen Peserta Didik Berbasis Sekolah), Malang: Universitas Negeri Malang, 2012.

2. T. Kohonen, "Self-organized formation of topologically correct feature maps," Biological Cybernetics, vol. 43, no. 1, pp. 59--69, 1982.

3. I. Y. Purbasari, F. T. Anggraeny and N. Harianto, "Classification of broiler chicken eggs using support vector machine (svm) and feature selection algorithm," in International Joint Conference on Science and Technology, Nusa Dua, 2018.

4. M. Maimunah and T. Rokhman, "Classification of Declining Quality of Chicken Eggs Based on the Color of Shells Using Support Vector Machine "(Klasifikasi Penurunan Kualitas Telur Ayam Ras Berdasarkan Warna Kerabang Menggunakan Support Vector Machine)," Informatics for Educators and Professionals, vol. 3, no. 1, pp. 43-52, 2018.

5. D. Nurdiyah and I. A. Muwakhid, "Comparison of Support Vector Machine and K-Nearest Neighbor for Fertile and Infertile Egg Classification Based on Glcm Texture Analysis " (Perbandingan Support Vector Machine dan K-Nearest Neighbor Untuk Klasifikasi Telur Fertil Dan Infertil Berdasarkan Analisis Texture GLCM), Jurnal Transformatika, vol. 13, no. 2, pp. 29-34, 2016.

6. S. Lakho, A. H. Jalbani, M. S. Vighio, I. A. Memon, S. S. Soomro and S. Q. N, "Decision Support System for Hepatitis Disease Diagnosis using Bayesian Network," Journal of Computing and Mathematical Sciences, vol. 1, no. 2, pp. 11-19, 2017.

7. F. Anggraeny, I. Purbasari and E. Suryaningsih, "ReliefF Feature Selection and Bayesian Network Model for Hepatitis Diagnosis," in International Conferences on Information Technology and Business (ICITB), Bandar Lampung, 2017.

8. F. T. Anggraeny, "Prediction of Student's Academic Achievement using Artificial Neural Network (Prediksi Prestasi Akademik Mahasiswa dengan Metode Jaringan Syaraf Tiruan)," in National Seminar of Information Technology Roles in Food, Chemical, and Manufacturing Industries to Support Development (Seminar Nasional Peran Teknologi Informasi di Bidang Industri Pangan, Kimia, dan Manufaktur dalam

Menunjang Pembangunan), Universitas Pembangunan Nasional "Veteran" Jawa Timur, Surabaya, 2009.

9. S. Isljamovic and M. Suknovic, "Predicting Students' Academic Performance using Artificial Neural Network: A Case Study from Faculty of Organizational Sciences," in The Eurasia Proceedings of Educational & Social Sciences (EPESS), Konya, Turkey, 2014.

10. O. L. Usman and A. O. Adenubi, "Artificial Neural Network (ANN) Model for Predicting Students' Academic Performance," Journal of Science and Information Technology, vol. 1, no. 2, pp. 23-37, 2013.

11. E. Y. Obsie and S. A. Adem, "Prediction of Student Academic Performance using Neural Network, Linear Regression and Support Vector Regression: A Case Study," International Journal of Computer Applications, vol. 180, no. 40, pp. 39-47, 2018.

12. L. Rahmawati, A. D. Cahyani and S. S. Putro, "Utilization of SOM-IDB cluster method as an Analysis of Scholarship Acceptance Analysis (Pemanfaatan metode cluster SOM – IDB sebagai Analisa Pengelompokan Penerimaan Beasiswa)," University of Trunojoyo Madura, Bangkalan, 2013.

13. N. Hendayanti, G. Putri and M. Nurhidayati, " Accuracy of Classification of STMIK STIKOM Bali Scholarship Recipients with Hybrid Self Organizing Maps and K-Mean Algorithms (Ketepatan Klasifikasi Penerima Beasiswa STMIK STIKOM Bali dengan Hybrid Self Organizing Maps dan Algoritma K-Mean)," VARIAN Journal, vol. 2, no. 1, pp. 1-7, 2018.

14. M. Bara, N. Ahmad, M. Modu and H. Ali, "Self-organizing map clustering method for the analysis of elearning activities," in Majan International Conference (MIC), Muscat, Oman, 2018.

15. Y. Lee, "Using Self-Organizing Map and Clustering to Investigate Problem-Solving Patterns in the Massive Open Online Course: An Exploratory Study," Journal of Educational Computing Research, vol. 57, no. 2, pp. 471-490, 2019.