Zulfaqar Int. J. Def. Sci. Eng. Tech. Vol.1 Issue 1 (2018) 50-56

ZULFAQAR International Journal of
Defence Science, Engineering & Technology

Journal homepage: www.zulfaqar.upnm.edu.my

Monitoring Students Performance using Self Organizing Map Trend Clustering

Puteri Nor Ellyza Nohuddin[a,*], **Zuraini Zainol**[b], **Aliimran Nordin**[a]

[a] *Institute of Visual Informatics, Universiti Kebangsaan Malaysia 43650 Bangi, Selangor, Malaysia*
[b] *Faculty of Science and Technology, Universiti Pertahanan Nasional Malaysia, Kem Sungai Besi 57000 Kuala Lumpur, Malaysia*

*Corresponding author: *puteri.ivi@ukm.edu.my*

ARTICLE INFO

Article history:
Received
25-05-2018
Received in revised
13-06-2018
Accepted
26-06-2018
Available online
30-06-2018

Keywords:
Data Mining,
Clustering,
Trends,
Student Performance,
SOM

e-ISSN:
Type: Article

ABSTRACT

The analysis of relation between student performance and other variables in education setting is often useful in identifying influential factors on performance. Consequently, the need for adopting an effective tool to process these big data has risen. The analysis of big data will transform passive data into useful information. Data mining is referred to an analytic process designed that discovers data patterns and relationships between datasets. In this study, clustering is used to cluster student grade datasets to generate trend line clusters. The aim of the study is to assist lecturers and academic advisors to recognize the progress of their students.

Introduction

Excellent student performance is one of indicators of course learning outcome at the university. Examinations, assignments and course projects are commonly used as course assessments to evaluate the students understanding and knowledge gain. Analyzing students' performances grow into more challenging because of the growing amount of data in educational databases. Therefore, the need of having effective tools to process these student data has risen. The term data mining is often referred to an analytic process designed that discovers data patterns and relationships between data variables. Many data mining approaches can be adapted to analyze the data such as classification, clustering and association rules mining depending on the suitability data collection and objectives of data analytical process.

This study illustrates how data from student performance can be exploited to construct a closely focused curriculum addressing students' understanding in the course syllabus. The authors proposed a monitoring student performance application tool using Self Organizing Maps (SOM) to cluster a number of student grade records. The application tool was developed to assist lecturers and teachers in general, especially in analyzing the student performance using the grades data from tests and exams. Based on the experiments, the test results have shown that the application is capable of accommodating number of data sets and produce the lines graphs. Since, monitoring student performance is a key task in teaching and learning, it is essential to evaluate the student performance to ensure students understand the course content. There are several studies on monitoring student performance using data mining technologies. Moucary et al proposed a hybrid technique to analyse and predict student performance using Neural Network and Clustering (Moucary et al, 2011). In other studies, clustering is used to group the student data in analyzing the student performance to gain deep insights (Kadiyala & Potluri, 2014; Bindiya et al, 2010).

The rest of the paper is organized as follows. Section 2 discusses the background of several topics that are related to trend analysis and data clustering using SOM. Then Section 3 provides a description of the modules of the proposed framework for SOM Trend Clustering tool. This is followed by Section 4 with a demonstration of the data clustering analysis using a set of big data. Finally, in Section 5, the paper is concluded with a brief summary and future research work.

Background and Related Works

This section discusses literature reviews on related tend clustering topics which includes trend analysis, data mining tasks and clustering technique.

i. Trend analysis

We always refer time series data as a special type of chronological data in which each record is in a temporal manner. With time, a trend analysis can be done in which to study the movement of the data. Trends can be both upward and downward, concerning increasing or decreasing performance, respectively, of an object (Ding et al, 2008; Nohuddin et al, 2018). While there is no specified minimum amount of time required for a direction to be measured as trend, the longer time series data is analyzed, the more prominent the trends.

Time series analysis can be visualized as sets of trend lines. The trend line graph of the trend line may be increased and decreased according to the evaluation of overall time series data (Nohuddin, 2010, 2017). This graph also demonstrates a very interesting pattern when multiple trends occur simultaneously. In other words, trend analysis is also a quantitative method to determine the pattern of past data that have been collected on a regular basis according to the time sequence of events. Patterns of the past can be used as a basis for predicting the student performance. In some cases, a large numbers of trend graphs are generated due to a tremendous amount of time series data records. Consequently, a technique of grouping these trends can be applied to determine types of trends exist in time series data.

ii. Data mining Tasks

Data mining is a technology to extract non trivial data patterns and knowledge from large raw datasets. It allows companies and governments to use the information you provide to reveal more than rows of data records. Discovering information from data can be in two main types: description and prediction. Therefore, data mining assists to simplify and summarize the data in a manner that decision makers, non-technical managers and users can understand and make decisions based on the generated patterns they have observed. There are four major techniques in data mining (Witten, 2016), they are:

a. Clustering

Clustering refers to a technique by which a set of records are group into a number of clusters based upon similar characteristics in the datasets. The similarity is determined by using a distance function for example Euclidean and Manhattan functions. K-Means, Hierarchical and Self Organizing Maps are among popular techniques in clustering.

b. Classification

Classification is a supervised data mining technique in which it classifies a set of records according to assigned data classes to levels of a categorical factor based upon their characteristics. A training dataset is used to train and develop a classification algorithm which can then be used to predict which category unknown cases are most likely to belong to.

c. Association Rules

Association rules are used to identify relationship or correlation between the data variables or attributes. Given a set of data transactions, association rules aim to find the rules which enable us to predict the occurrence of a specific item based on the occurrences of the other items in the transaction.

d. Prediction

Prediction refers to the development of forecasting models that can predict future data trends. It studies the current data patterns and forecast whether the pattern is going to repeat again. Regression models, classification, clustering are among methods used in prediction.

iii. Clustering and Self Organizing Maps

As mentioned in previous section, clustering is an unsupervised process of grouping similar datasets into clusters (Witten, 2016). Clustering can also reveal previously undetected relationships in a dataset. There are several common clustering techniques such as K-means, Hierarchical, Self Organizing Maps and many more. The dynamic method in clustering is how to rule the similarity between two objects, so that clusters can be discovered from datasets with prominent similarity within clusters and low similarity between clusters. Often, clustering measures similarity or dissimilarity between objects using a distance measure such as Euclidean, Manhattan and Minkowski (Witten, 2016). Thus, the aim of clustering is to ease users' interpretation and analysis on massive datasets.

One of clustering technique which is Self-Organizing Maps (SOMs) were first introduced by Kohonen (Kohonen, 1997, 2001). Essentially, SOMs may be viewed as a neural network based technique designed to reduce the number of data dimensions in some input space by projecting it onto a n × m "node map" which plots the similarities of the input data by grouping (clustering) similar data items together at nodes. SOMs have been utilized in many research areas.

The SOM Trend Clustering Framework

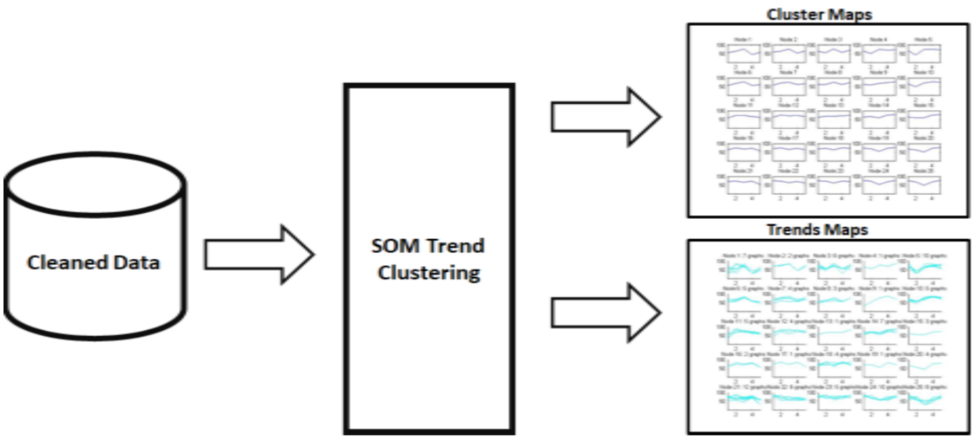

Fig. 1: Framework of SOM Trend Cluster Analysis

Fig. 1 presents the framework for Trend Clustering Technique which contains of three modules: (i) Data Pre-processing (ii) SOM Trend Clustering and (iii) SOM Output Maps: Cluster and Trend Maps.

i. Data Pre-processing

The proposed method initiates with data pre-processing. Data processing comprises the procedure of converting and cleaning the raw data into a specific data format and value before generating the trend line clusters. Each value is converted into nominal or/and numerical data type. After data preprocessing is done, it will be the input data for SOM trend clustering module.

ii. SOM Trend Clustering

In this study, SOM Trend Clustering module was developed using Matlab (Mat Amin et al, 2014; Nohuddin et al, 2018). SOM function which takes in a set of numerical data. This module is able to process a large number of data records and generate cluster trend lines on a map. The SOM function is setup as a 5 x 5 cluster maps so as it groups the data set into 25 clusters. The model applies Euclidean Function to measure the similarities of each data records and distribute the data accordingly to form trend line patterns. Each record in the training set is presented to the SOM in turn and the output nodes compete for the record. For each record, once it has been assigned to a node, the network's weightings are adjusted to reflect the new position. A feature of the network is that adjacent nodes hold similar records; the greatest dissimilarity is thus between nodes at opposite corners of the grid. It also identifies how many types of trend lines exist within the numerical data. These trend lines show what type of trend data such as increasing and decreasing lines according to the overall dataset.

In final process, the SOM Trend Clustering module produces two kinds of maps which are the cluster map and trend maps that illustrate a set of trend lines.

iii. SOM Output: Cluster and Trend Maps

The second module is SOM Output: Cluster and Trend Maps. There are two (2) kinds of maps which are the SOM Cluster Maps which illustrate the trend line types and SOM Trend Maps that illustrate a set of trend lines. It provides the details of each trend line cluster identified in SOM Trend Clustering. It also displays a label indicating numbers of data records contain in each trend line cluster. Therefore, it is possible to track which data record belongs to which trend line clusters.

Experiment and Results

This section describes the experiment using student grade data and the generated results of trend line clusters on student performance.

i. Student Grade data

Table 1. Some example of Student Grade Data.

Student ID	Test1	Test2	Midterm	Quizzes	Final
2014255	55	78	63	50	80
2014301	83	45	88	52	47
2014023	75	70	42	74	63
2014155	67	86	54	87	86
2014135	59	60	76	65	78

Student Grade data is used to evaluate the proposed SOM Trend Clustering application. The simulated data is based on typical course assessment grade. Table 1 shows some example of the student grade data. The data has six attributes: Student ID, Test1, Test2, Midterm, Quizzes, and Final.

However, only Test1, Test2, Midterm, Quizzes, and Final attributes are used for presenting the trend lines. There are about 135 records in the dataset. The value range for trend attributes is between 0-100. The data is cleaned and formatted as required.

ii. The SOM Cluster Map

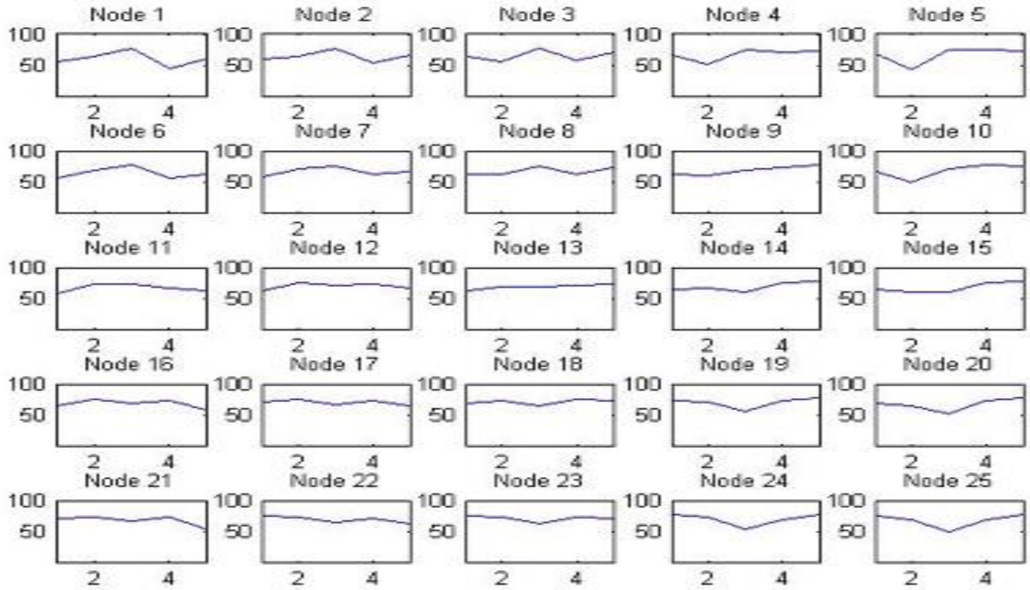

Fig 2: Student Performance SOM Cluster Maps

The first output of SOM Trend Clustering module is SOM Cluster Map. It illustrates the trend line types exists in the analyzed grade data.

Fig. 2 displays 25 cluster maps of student performance. The SOM Trend Clustering algorithm distributes 135 student grade records using SOM function and forms the 5 attribute values into trend lines.

For example, Node1, 2 and 6 represent the trend cluster for students who scored well in their Midterms however have the lowest marks in their quizzes. Node 9 and 13 hold progressing trend line of student grades. Whereas Node 25 clustered the student grades who scored well in Test1 and Final Exam.

iii. The SOM Trend Map

The second output of SOM Trend Clustering is the SOM Trend Maps. It has the same dimension as the SOM Cluster Maps, 5 x 5 maps (25 maps). These maps provide details of each cluster maps in Fig. 3 such as numbers of cluster memberships and the actual contours of trend lines.

In Fig. 3, the lecturers or management are able to analyze further from the trend maps. Node 21 is having the highest number of trend lines (12 trend lines) which represent students who scored relatively 70-80 marks in Test1, Test2, Midterm and Quizzes but did badly in their finals (scored less than 50 marks). Then, Node 5 and Node 24 are having 10 trend lines in each cluster. Node 5 collects the trend lines of students who scored less than 50 marks in Test2 but improved their grades in other assessments. Whereas Node 24 has trend lines of students who scored high marks in Test1, Test2, Quizzes and Final Exam but they attained low marks in their Midterms.

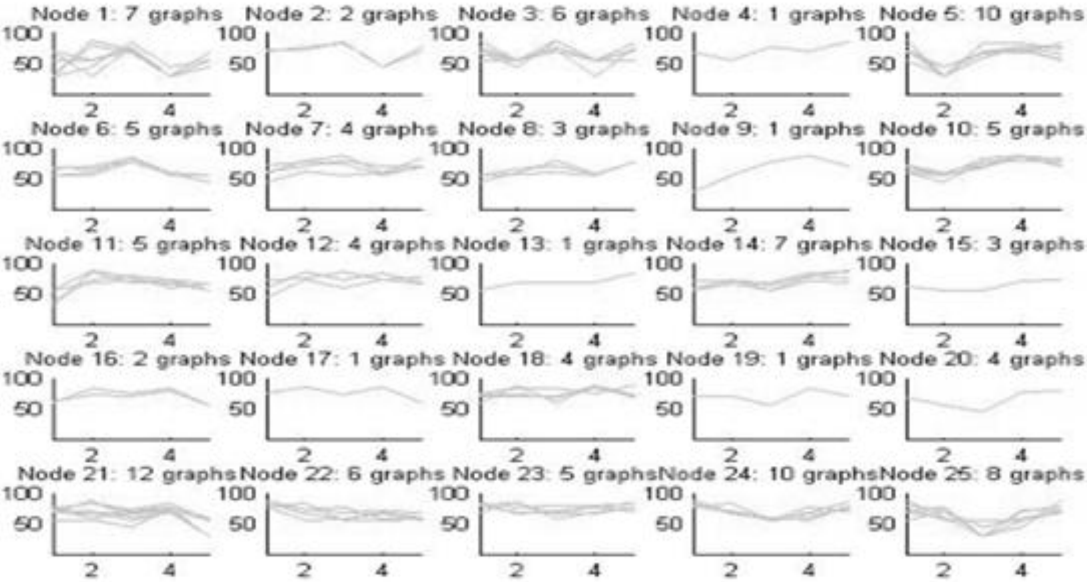

Fig. 3: Student Performance SOM Trend Maps

We can deduce that from the SOM Cluster and Trend maps, lecturers and teachers are able to investigate the trends of student performances throughout the semester. With the generated maps, lecturers and teachers are able to identify types and numbers of trend lines represent student performance based in their academic assessments.

Conclusion

The SOM Trend Clustering application was developed to identify and analyze student performance data. This application, it can assist academician and management to identify changes that transpire in student performance data as well as facilitate researchers evaluate this type of time series or trend data.

For future works, there is still a lot of improvement that can be applied on SOM Trend Clustering application. SOMs are often described as a visualisation technique. However, given a large and/or complex data set the number of items within each cluster may still be large. This was found to be the case with respect to the degree of similarity of trend lines in each cluster. One potential solution was to increase the size of the map, however this may lead to an undesirable computational overhead and was found not to resolve the situation because some map nodes remain empty. Also, the system can be enhanced further into an online monitoring system to make it easier for academician, management and researchers to investigate wherever they are located.

References

Moucary, C. E., Khair, M. and Zakhem, W. (2011) Improving Student's Performance Using Data Clustering and Neural Networks in Foreign-Language Based Higher Education. *The Research Bulletin of Jordan ACM, Vol. II* (III). pp 27.

Kadiyala, S. and C. S. Potluri, C. S. (2014) Analyzing the Student's Academic Performance by using Clustering Methods in Data Mining. *Int. J. Sci. Eng. Res., vol. 5*, no. 6, pp. 198–202.

Bindiya M. V., Jose, T., Unnikrishnan, A. and Poulose, J. K. (2010) Clustering Student Data to Characterize Performance Patterns. *Int. J. Adv. Comput. Sci. Appl.*, pp. 138–140, 2010

Ding, H., Trajcevski, G., Scheuermann, P., Wang, X. and Keogh, E. (2008) Querying and mining of time series data: experimental comparison of representations and distance measures. *Proc. VLDB Endow*. 1, 2

Rashid, R. A. A., Nohuddin, P. N. E. and Zainol, Z. (2017)Association Rule Mining Using Time Series Data for Malaysia Climate Variability Prediction. *IVIC 2017*, pp.120-130.

Nohuddin, P. N. E., Christley, R., Coenen, F. and Setzkorn, C. (2010) Trend Mining in Social Networks: A Study Using a Large Cattle Movement Database. *ICDM 2010*: 464-475.

Witten, I. H., Frank, E., Hall, M. and Pal, C. (2016) Data Mining (4th Edition) Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

Kohonen, T. (1997) Self-Organizing Maps, Springer Series in Information Sciences.

Kohonen, T. (2001) Self-Organizing Maps, Springer Series in Information Sciences.

Mat Amin, M.N., Nohuddin, P. N. E and Zainol, Z. (2014) Trend cluster analysis using self-organizing maps. Information and Communication Technologies (WICT), *2014 Fourth World Congress*, pp 80-84.

Nohuddin, P. N. E., Zainol, Z. and Amir, M. A. U. (2018) Trend Cluster Analysis of Wave Data for Renewable Energy. *Advanced Science Letters, Volume 24*, Number 2, February 2018, pp. 951-955(5)