Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 163 (2019) 550–559

Procedia Computer Science

www.elsevier.com/locate/procedia

16th International Learning & Technology Conference 2019

Principal Component Analysis and Self-Organizing Map Clustering for Student Browsing Behaviour Analysis

Nor Bahiah Ahmad[a]*, Umi Farhana Alias[a], Nadirah Mohamad[a], Norazah Yusof[b]

[a]*School of Computing, Faculty of Engineering,Universiti Teknologi Malaysia, Johor Bahru, Johor, 81310, Malaysia.*
[b]*Faculty of Computing and Information Technology, King Abdulaziz University, Rabigh, 21911, Saudi Arabia.*

Abstract

Using e-learning as the learning platform provides flexibility to student to learn anywhere without limitation of time and space. Moodle, currently become one of the leading learning management system that deliver e-learning easily by providing customized tool for educators to deploy learning materials in various forms. The students behavior is important to be identified in order for educators to improve the teaching approach and to enhance the students performances. However, the flexibility in the learning environment cause student behaviour more challenging to be identified. In Moodle, the student's interactions and activities while learning online are captured in the log files. The data stored in the log files contain meaningful information such as the student behavior, their preferences and their knowledge level. In this study, the raw dataset captured in the log file undergo pre-processing phase such as data cleaning, transformation and selection to prepare the dataset for the analysis purposes. Principal Component Analysis (PCA) that is known for improvement of accuracy for unsupervised learning technique is used to identify the most significant features of students attributes from the log file. Course view, notes, exercises, examples and assignments are the features selected using PCA to be feed into Self-Organizing Map (SOM) in order to cluster students based on the behavior. The result shows that the technique is able to cluster the student browsing behavior using the total number of browsing frequency in 14 weeks study session, which form two cluster groups; high browsing behavior and low browsing behavior. The quality of SOM clustering technique is validated through the internal clustering evaluation and the results show that SOM produces better quality of cluster groups compared to k-mean and Partition Around Medoids (PAM).

Keywords: Student behaviour; Log file; Clustering; Self-Organizing Map (SOM); Principal Component Analysis (PCA)

* Corresponding author.
E-mail address: bahiah@utm.my

1877-0509
10.1016/j.procs.2019.12.137

1. Introduction

The enrichment of technologies has influenced the education environment such as learning management system (LMS) to manage the learning material in education institution [1]. Moodle is the most leading LMS adopted due its flexibility to deliver learning materials and to create the effective online courses [2,3]. As the LMS has large database storage, all the student's actions are captured through the communication between students and computer [4]. The changes in learning environment makes student behaviour become more challenging to be identified due to the freedom given while accessing the learning materials. The behaviours that the students made, produced a lot of data while using the learning system [5]. However, the data collected could not directly show the student behaviour without any further analysis.

Student behaviour is known as the knowledge of observable actions for student in any learning domains and as the input of construction for student's model [6]. The result from student behavior are important for identifying the knowledge level, motivation and goals of the learners [7], and also important in constructing and improving the learning content [8,9]. Therefore, student behaviour plays an important element in teaching and learning process. The changes in learning environment makes the student behaviour becomes more challenging to be identified due to the freedom given while accessing the learning material. Moreover, there is a lot of data produced while the students using the learning system [5]. Despite using observation and questionnaire [10], student's action is still captured through the communication between students and computer through keyboard and mouse [4].

E-learning supported by Moodle platform has the function to record all interactions occur in the learning system. Compared to traditional learning, the learning materials in Moodle have variety types such as text, audio, images, animation, video and online discussion for learning process [11]. Therefore, there are lot of data from all activities that were recorded in the log file. Although most e-learning uses the concept of sharing the same materials (one-size-fit-all), the behaviour may different. However, the data collected could not directly shows the student behaviour without any further analysis process because the function in Moodle can only capture the data. Therefore, data mining technique is used to analyse the data recorded for many purposes in learning improvement including identifying the student behaviour while using the learning system. Data mining is the process of knowledge discovery in database (KDD) from large data repositories [12], while Educational Data Mining (EDM) is specifically for education data where some knowledge is extracted from log and data file [13]. Table 1 shows the analysis of behaviour in learning management system done by previous researchers that contribute to learning improvement. Previous researchers approved that the patterns produced from the behavior significantly different among each other and help in many aspects such as dissimilarity between the student behavior with their performance [14], customize learning content [9] and others. Thus, the findings show that the student behaviour is important towards the improvement of learning quality.

Table 1. Behaviour analysis in learning management.

Researchers	Purpose	Technique	Indicator	Behaviour Pattern
[15]	Groups similar behaviour with different levels of achievement for student adaptation suggestion.	EM and k-means algorithm	Time spent in theoretical contents, forums and tasks	Task oriented groups and non-task-oriented groups
[16]	Analysis using lecture video to suggest the course design towards student ability and satisfaction	Clustering analysis	Streaming videos, access content and final grades	Adaptive viewer, self-regulating viewer, and infrequent viewer
[17]	Analysis of slow learner that enhance their learning capabilities and help to construct the teaching methodology	EM and k-means algorithm	image, sound, emotions and notes	RSWG, RWSG, RWGS, GRSW, and SGWR (Note: Representation of learning sequence: R-Read, S-Speaking, W-Write, G-Grammar).
[8]	Analysis towards the relationship between course usage and performance of students and affected grade.	Hierarchical agglomerative clustering	CourseID, SessionID and URL	High activity and low activity

Clustering is one of the most adopted process to analyse the student behaviour in learning environment system [18]. Before the analysis process, the data in log file need to undergo some pre-processing process such as data cleaning, transformation and selection. Feature selection is essential step for the success of data mining process by reducing the dimensionality of the data [19]. Principal Component Analysis (PCA) by Jolliffe, 1989 [20] is implemented as a method to reduce the attributes by making a new set of attributes called principal component (PC) and this technique becomes widely applied due to the suitability for unsupervised feature extraction [21]. Self-Organizing Map (SOM) is one of the clustering techniques that work well in other areas such as statistical method, industrial analyses, biomedical analyses and financial application [22]. SOM introduced by Teuvo Kohonen, 1998 [23] is a chosen technique in data mining that can be implemented from small to large size of data [24] and have the ability in solving the others technique limitation such as in implementation with simplicity, better execution speed and shorter training process using unsupervised learning [25]. Therefore, the ability and the strength of SOM encourage this study to analyse the student behaviour and find the significant patterns from e-learning system. Next subtopic discusses about the methodology conducted for this experiment.

2. Methodology

In this research area, there are no standard for data being used in analysing student behaviour in LMS. Hence, data collected for this research are from the undergraduate students Faculty of Computing, Universiti Teknologi Malaysia (UTM). The log file is extracted automatically from the Data Structure and Algorithm subject, semester 1, session 2014/2015. The information in the log file are subject code and section information, date and time of user access the system, user information, IP address and activities of student while accessed e-learning system.

2.1. Pre-processing Phase

The log file captures all the interactions that occur while the system is being used. Therefore, data pre-processing is needed to prepare the data before analysis process. The following subsection discusses about process involve in pre-processing phases.

2.1.1. Data Cleaning

Data cleaning is the process to remove any noise and irrelevant data to obtain the clean data. The raw dataset contains data from all the users that using the e-learning system. Therefore, the cleaning process removes the interaction of educator and teaching assistance (TA) because this experiment aims to analyse the behaviour of students only while using the e-learning system. Data from other users were excluded. The data from the most accessed modules by students which are assignment module, course module and resource module are included which means other than those mentioned are being removed. Initially, there are 10027 records of data in log file. After the cleaning process, there are only 6051 records of data left in the log file from week 1 until week 14 for formal teaching and learning class. Therefore, the remaining features are listed in Table 2. Each of the action from the modules were given the variable name as the identification for the analysis process. Next, the data is prepared by filtering the action based on students' information and weeks for the whole datasets.

Table 2. Features in log file.

Module		Action	Variable name	Explanation
Assignment		Submit	asgsub	Submit the works as requested.
		View	asg	View and download the information about the assignment
Course		View	courseV	Contains the information of the learning environment
Resource	Exercise	View	exercise	Allow to view and download the information about the exercise
		Submit	exersub	Allow student to submit the works as request
	Example	View	example	Allow to view and download the information about the example
	Quiz	View	quiz	Allow to view and download the information about the quiz
		Submit	submitquiz	Allow student to submit the works as request
	Test Information	View	test	Allow to view and download the information regarding the test
	Forum	View	forum	View the discussion occur among the class member and teachers
	Notes	View	notes	Allow to view and download learning material

2.1.2. Data Transformation

Data transformation transforms the data into appropriate format for the mining process. The mining process is required to discover some valuable information from the datasets. The datasets may contain the value of irregular pattern. Before data transformation occur, the data is normalized to scale the attribute data into certain range [26]. This process is important to prevent any attribute value that might overpower to one another. In this experiment, min-max normalization is used as linear transformation because of the suitability of the datasets using the continuous values as mentioned by [27] and the advantages in preserving the relationship of the data [28]. The following equation describes the process of min-max normalization that scales the numeric variable in the range [0, 1].

$$A = \frac{B - min_x}{max_x - min_x}(1 - 0) + 0 \qquad (1)$$

Where

A	=	Value of data after normalization
B	=	Value of current data in the column
min_x	=	Minimum value of each attribute
min_x	=	Maximum value of each attribute
1	=	Maximum value of range [0,1]
0	=	Minimum value of range [0,1]

The data is normalized into new sets of data with the range value 0 to 1. Next, the dataset is transformed into the format that fit the algorithm in machine learning. Fig. 1 shows the example of dataset that has been transformed in ASCII format. The name of the features is under attributes and the data contain the values for each attribute.

```
@attribute courseV numeric
@attribute notes numeric
@attribute exercise numeric
@attribute example numeric
@attribute asg numeric

@data
0.0833,0.1000,0.1111,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000
0.1111,0.1000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000
0.1389,0.3333,0.2222,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000
0.5278,0.4000,0.1111,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000
```

Fig. 1. Example of dataset in ASCII format.

2.1.3. Data Selection

Currently, there are 11 features in the dataset. However, not all the features are significant to analyse the student browsing behaviour in e-learning system. Hence, to select the most significant features from 11 attributes, Principal Component Analysis (PCA) by Jolliffe, 1989 [29] is implemented by selecting the features through new set of attributes called principal component (PC). PCA is chosen because of suitability of techniques for unsupervised learning process and improve the accuracy for unsupervised learning techniques [30,31]. The detailed process of PCA are describe in the following subtopics.

2.1.3.1. Prepare dataset

The features involved in PCA are courseV, notes, exercise, exersub, example, quiz, submitquiz, asg, asgsub, test and forum. The column represents the features and row represent each data for the features. All the data prepared without the class label.

2.1.3.2. Calculate the mean vector

The mean vector is calculated for features in each column for the whole dataset. In this process, the features also known as variables. Next, the mean is subtracted to identify the center of origin dataset across the dimension in data matrix.

2.1.3.3. Calculate the covariance matrix

Next, covariance matrix is computed for the whole dataset. The current matrix is transposed into 11 x 11 rows and column. The covariance also represents the correlation between the variables. The covariance matrix is in diagonal matrix shows the correlation for each variable.

2.1.3.4. Calculate eigenvectors and eigenvalues

Compute the eigenvectors from the correlation value in diagonal matrix and the corresponding of eigenvalue in the scale length value of 1.00. Eigenvalue is one criterion by Kaiser, 1960 [32] that is common criterion to identify principal components. The value of eigenvector and eigenvalues is correspondent to each other and come in pairs. The eigenvalue with highest eigenvalue indicates more meaningful component [33]. The eigenvectors are sorted by decreasing the eigenvalue using the ranker. The ranker gives the components in order of significance.

2.1.3.5. Sorting the eigenvectors and formed the feature vector

Next step is eigenvector from the correlation matrix which is sorted by eigenvalue from the highest to lowest values. The feature vector is built from the component based on the ranking.

2.1.3.6. Present new dataset

The new dataset contains data in highest group and lowest group. The new dataset has less dimensionality from the original data. In this research the less significant variable is ignored because the top ranked data give more valuable result for the clustering. The highest value in eigenvalue is presented in the top row. Table 3 list the example set of Principal Component based on the ranked of eigenvalue. In the first ranked, the component which is extracted as the most correlated variables become the factor for the variable been retained and interpreted [34]. Therefore, courseV, notes, exercise, example and asg are the most significant variable in this dataset.

Table 3. Set of Principal Component.

Percentage (%)	Ranked	Cumulative variance
0.7829	1	0.533courseV+0.417notes+0.366exercise+0.361example+0.336asg
0.6403	2	-0.497exersub+0.45quiz-0.446forum+0.433example-0.267exercise
0.506	3	0.498asgsub+0.479asg-0.435exersub-0.298forum-0.264example

2.2. SOM Clustering

Next, the data from pre-processing process is prepared for clustering implementation. five features from PCA become the attributes in analyzing student's browsing behavior in e-learning using SOM algorithm. Self-Organizing Map from Teuvo Kohonen, 1998 [35] clustering algorithm is implemented in the cluster analysis that apply competitive learning algorithm and enable the neurons to compete each other in the network. The process in SOM clustering involves two main processes which are initialization and training. The detail explanations are described in the following subtopic.

2.2.1. Self-Organizing Map Initialization

The initialization process starts with the specification of number of map units, map size and lattice. The map units are identified using the following equations by Vesanto, 2000 [36]:

$$munits = 5 * \sqrt{number\ of\ training\ samples} \quad (2)$$

Where,
$munits$ = The size of map units for dataset.
$number\ of\ training\ samples$ = The size of dataset.

Hence, using the map units, the map size is determined for this experiment which is 12x7. Two-dimensional lattice grid is used to project the array nodes in rectangular lattice grid. The weight of input vector is initialized randomly in the SOM model. Hence the size of input nodes is 266 numbers of data with five input of space dimension.

2.2.2. Self-Organizing Map Training

The steps in training process involve the input vector from the initialization process. The neuron is competed in each iteration to find the winning neurons called Best Matching Units (BMU). The distance for BMU is calculated using the Euclidean distance. Along the process, the neighbourhood radius decreases and is updated using the most flexible neighbourhood function which is Gaussian neighbourhood [37]. The training process is done in two cycles. The parameter involved in this step are learning rate, training length and neighbourhood radius. The training length is at least 10 times the number of map units. For the first cycle, the rough training starts with the biggest number of learning rate and neighbourhood radius. The second cycle which is fine-tuning used the small number of learning rate and neighbourhood radius. The initial neighbourhood radius is identified using the following equation:

$$\frac{maximum(map\ size)}{4} \quad (3)$$

Where,
$map\ size$ = The map size of dataset

Meanwhile, the final neighbourhood radius is set to 1. The value of learning rate is in the range 0 to 1. Table 4 shows the best learning rate that are chosen from the training process. The value of parameters for both training set are also affected by the value of quantization error (QE) and topological error (TE). The value of QE is computed through the average distance of vector and centroids that they represented [38]. QE is one of the standard evaluation SOM that are calculated after finishing each epoch [38] that show how properly the network is trained. On the other hand, TE represent the proportion value of first and second BMU [40]. The error occurs when the proportion value is not adjacent on the map lattice [38].

The learning rates give low value of QE and TE at the value 1.0. Among the value, the best starting value for learning rate is 1.0 at the 1000 rough training length and 2,000 fine-tune training length. Therefore, the learning rate is decided based on the quality of SOM through the small value of QE and TE [41]. Among the value of QE and TE, the small average of QE is 0.115 and the TE value is 0.015. The neighbourhood radius in rough training is decreased and continued the last value in fine tune training.

Table 4. Result of training process.

Rough Training (RT)		Fine-tune Training (FT)		QE	TE
Learning rate	Length RT	Learning time	Length FT		
1.0	100	0.01	200	0.127	0.018
0.8	200		400	0.122	0.012
1.0	400		800	0.118	0.013
1.0	600		1200	0.117	0.013
0.8	800		1600	0.116	0.016
1.0	1000		2000	0.115	0.015

Validation in clustering is very challenging task to perform because producing a good clustering is very subjective towards the cluster evaluation [42,43]. The clustering validation is used to evaluate the intra-cluster similarity and inter-cluster differences among the cluster. A good clustering result contain minimum intra-cluster distance while the inter-cluster is in maximum distance. Besides, different indexes give different range for a good clustering algorithm technique used. In this experiment, the internal clustering validation process is the most suitable technique which relies on the data itself. Some well-known measure for internal criteria is used for the validation process such as Silhouette index [44], Davies-Bouldin index [45] and Dunn index [46]. The internal criteria are measured using the average distance of similarity and dissimilarity of the cluster. The measurement involves the distance within the cluster member and distance between cluster groups [47]. Further discussion is described in the next section.

4. Result and discussion

The result from SOM clustering process produces two cluster behavior groups which are high browsing behavior and low browsing behavior. The dataset used for clustering consist of 266 instances and 188 instances are assigned in low browsing behavior cluster. The rest of the instances are grouped as high browsing behavior group. The centroid is cluster centre that carries the values for average distances within cluster members. Thus, the centroids are used to elaborate the characteristics of each cluster. Based on the clustering result, low browsing group shows the characteristic that contain the low value in the most of centroid value. On the other hand, the high browsing group leads the value in each attribute.

Although the high browsing group consists of high value in cluster centroid, the number of instances in this group is lower compared to low browsing group. The cluster centroid shows the similarity and dissimilarity characteristics among the cluster member and the attributes. Fig. 2 shows the example of scatter for courseV in x-axis toward the example in y-axis. The scatter plot in blue colour that represent the low browsing group also shows the behaviour in the low scale of the example while the course is browsed. High browsing group maintain the behaviour when visits the e-learning and browses the example. Next is some discussion of student behaviour toward the weeks of teaching and learning process.

Fig. 2. Example of scatter plot towards course view and example.

In the beginning of the semester, the learning materials prepared for the students are only the notes. Fig. 3 visualize the comparison of browsing behaviour in each group for week 1. The low browsing group shows more frequency in browsing the course and notes compare to high browsing group. However, students in high browsing group have curiosity in browsing the other learning material outside the week 1.

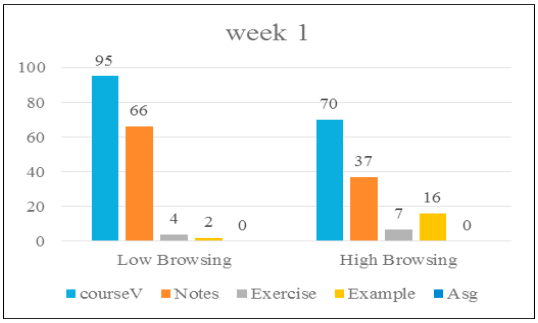


Fig. 3. Week 1 cluster group behaviour.

For week 3, the learning material prepared by educator is same as in week 2. The comparison of browsing behaviour in each group for week 3 is visualized in Fig. 4. The range for browsing hits in week 3 are three times higher than the previous weeks. Surprisingly, the high browsing group actively visit and browse the learning materials prepared by the educator. Moreover, the assignment activities also have been browsed by this group. Compared to low browsing group, the behaviour shows by this group are falling behind for all the learning activities. Besides, this group also browse only the learning material that are specifically prepared by the educator in week 3. Student shows different behaviour in one semester. The result given by clustering process can give an idea to educator in preparing the learning material based on the behaviour of the student in each week. Thus, it can encourage the student to browse the course and use the learning system more frequently.

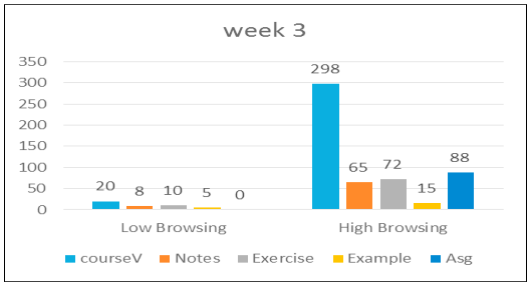
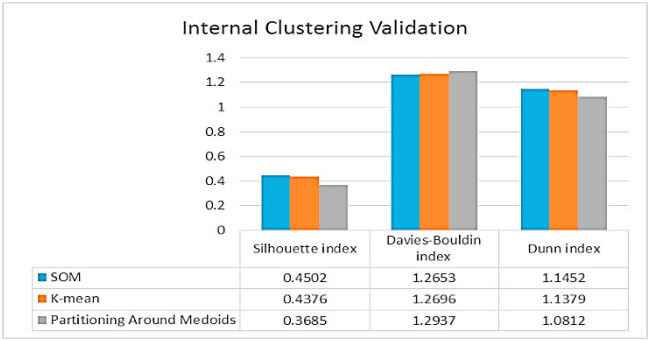


Fig. 4. Week 3 cluster group behaviour.

The cluster group is validated using the internal evaluation because of suitability of this technique for the data without class label. Dunn index and Davies-Bouldin index validate both for intra-cluster similarity and inter-cluster dissimilarity. Otherwise, the Silhouette index validates the distance within the cluster distances. Fig. 5 shows the comparison of internal validation with different clustering algorithm.



	Silhouette index	Davies-Bouldin index	Dunn index
SOM	0.4502	1.2653	1.1452
K-mean	0.4376	1.2696	1.1379
Partitioning Around Medoids	0.3685	1.2937	1.0812

Fig. 5. Comparison of internal evaluation with different clustering algorithm.

The clustering algorithms are compared with the same type of algorithm which are partitioning group. The lowest value in Davies-bouldin index indicates the algorithm used is the best algorithm to form a good cluster structure produced. While in Dunn index and Silhouette, the higher value indicates the algorithm have better result compared to others. The result shows that SOM clustering algorithm give better result compared to others algorithm with higher value of Dunn and Silhouette index, lower value for Davies-Bouldin index. Thus, it is proven that SOM clustering algorithm is the best algorithm to produce a good cluster structure.

5. Conclusion

This paper explores on how the interaction in Moodle log file can be used to analyse the students' browsing behavior using Principal Component Analysis (PCA) and Self-Organizing Map (SOM) clustering technique. The browsing behavior being analyzed is from e-learning for one semester. Firstly, the significant features for student's browsing behaviour have been identified using PCA which are course view, notes, exercise, example and assignment. The workflow for student browsing behaviour analysis have been designed. The pre-processing process were stressed especially in feature selection part. The student's browsing behaviour were analysed using SOM clustering technique. The analysis process produced two cluster groups that known as high browsing behaviour group and low browsing behaviour group. Hence, the behaviour in each group can give an idea to educators to prepare the learning content for each week. The result also proved that SOM clustering is a good clustering technique compared to others. Finally, for future works, this research would like to explore more on how the relation between student's browsing behavior with student's learning style and knowledge level for future adaptation in adaptive e-learning environment framework is and explore more on SOM clustering using GPULib for big data of educational data.

Acknowledgements

This paper is supported by Fundamental Research Grant Scheme (FRGS), vot. no. 4F496 and GUP Tier 1, vot no. 17H75. The authors would like to express their deepest gratitude to Research Management Centre, Universiti Teknologi Malaysia and MOHE for the support in R&D and Soft Computing Research Group for the inspiration in making this study a success.

References

[1] De Bra, P., Smits, D., Van Der Sluijs, K., Cristea, A. I., Foss, J., Glahn, C., and Steiner, C. M. (2013). "GRAPPLE: Learning management systems meet adaptive learning environments". *Intelligent and adaptive educational-learning systems, Springer*, Berlin Heidelberg, 133-160.

[2] Nash, Susan Smith. (2018). Moodle Course Design Best Practices: Design and develop outstanding Moodle learning experiences. Packt Publishing Ltd.

[3] Bulaeva, Manna N., Vaganova, Olga I., Koldina, Margarita I., Lapshova, Anna V., Anna V., and Khizhnyi, Anna V. (2018). "Preparation of Bachelors of Professional Training Using MOODLE". In *The Impact of Information on Modern Humans*, (**622**): 406-411.

[4] Wenger, E. (2014). Artificial intelligence and tutoring systems: computational and cognitive approaches to the communication of knowledge. Morgan Kaufmann.

[5] Gulati, P. and Sharma, A. (2012). "Educational data mining for improving educational quality". *International Journal of Computer Science and Information Technology & Security*, **2(3)**: 648–650.

[6] Sison, R., & Shimura, M. (1998). "Student modeling and machine learning". *International Journal of Artificial Intelligence in Education (IJAIED)*, (9): 128-158.

[7] Popescu, E. (2009). "Diagnosing students' learning style in an educational hypermedia system". Cognitive and emotional processes in Web-based education: Integrating human factors and personalization, advances in Web-based learning book series, *IGI Global*, 187-208.

[8] Kaur, M. (2013). "Cluster analysis of behavior of e-learners". *International Journal of Soft Computing and Engineering (IJSCE)*, **3(2):** 344–346.

[9] El Haddioui, I., & Khaldi, M. (2012). "Learning style and behavior analysis: A study on the learning management system". Manhali. *International Journal of Computer Applications*, **56(4)**: 9–15.

[10] Sullivan, A. M., Johnson, B., Owens, L., & Conway, R. (2014). "Punish them or engage them?: Teachers' views of unproductive student behaviours in the classroom". *Australian Journal of Teacher Education (Online)*, **39(6)**: 43.

[11] Moodle, 2014. https://moodle.org/. Retrieved on November 4, 2014.

[12] Maimon, O. and Rokach, L. (2010). Data mining and knowledge discovery Handbook, 1–15.

[13] Romero, C., Ventura, S. and García, E. (2008). "Data mining in course management systems: Moodle case study and tutorial". *Computers and Education*, **51(1)**: 368–384.
[14] Obadi, G., Drázdilová, P., Martinovic, J., Slaninová, K., & Snásel, V. (2010). "Using spectral clustering for finding students' patterns of behavior in social networks". *DATESO*, 118-130.
[15] Cerezo, R., Sánchez-Santillán, M., Paule-Ruiz, M. P., & Núñez, J. C. (2016). "Students' LMS interaction patterns and their relationship with achievement: a case study in higher education". *Computers & Education*, **(96)**: 42-54.
[16] Kuo, Y.-Y., J. Luo, and J. Brielmaier, (2015). "Investigating Students' Use of Lecture Videos in Online Courses: A Case Study for Understanding Learning Behaviors via Data Mining". *International Conference on Web-Based Learning*. **9412**: 231-237.
[17] Mohammad, T.Z. and Mahmoud, A.M. (2014). "Clustering of slow learners behavior for discovery of optimal patterns of learning". *International Journal of Advanced Computer Science and Applications*, **5(11)**: 102–109.
[18] Dráždilová, P. and Obadi, G. (2010). "Computational intelligence methods for data analysis and mining of eLearning activities". *Computational Intelligent for Technology Enhanced Learning,* **(273)**:195–224.
[19] Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010). "Understanding of internal clustering validation measures". In 2010 *IEEE International Conference on Data Mining*, 911-916.
[20] Jolliffe, I.T. (1989). "Principal component analysis: a beginner's guide-introduction and application". *Weather*, **45(10)**: 375–382.
[21] Farahat, A. K., Ghodsi, A., & Kamel, M. S. (2013). "Efficient greedy feature selection for unsupervised learning". *Knowledge and Information Systems*, **35(2)**: 285-310.
[22] Kohonen, T. (2014). MATLAB implementations and applications of the self-organizing map.
[23] Kohonen, T. (1998). "The self-organizing map". *Neurocomputing*, **78(9)**: 1464–1480.
[24] Vesanto, J., Himberg, J., Alhoniemi, E., & Parhankangas, J. (2000). SOM toolbox for Matlab 5. Helsinki University of Technology, Finland.
[25] Cabada, R.Z., Barrón Estrada, M.L. and Reyes García, C.A. (2011). "EDUCA: A web 2.0 authoring tool for developing adaptive and intelligent tutoring systems using a Kohonen network". *Expert Systems with Applications*, **38(8)**: 9522–9529.
[26] Visalakshi, Karthikeyani N., and K. Thangavel. (2009). "Impact of normalization in distributed k-means clustering". *International Journal of Soft Computing*, **4(4)**: 168–72.
[27] Han, J. and Kamber, M. (2006). "Chapter 2: data preprocessing". In *Data Mining: Concepts and Techniques*. San Francisco: Morgan Kaufmann, 67–71.
[28] Jayalakshmi, T., and A. Santhakumaran. (2011). "Statistical normalization and back propagationfor classification". International Journal of Computer Theory and Engineering, **3(1):** 89.
[29] Jolliffe, I.T. (1989). "Principal component analysis: a beginner's guide-introduction and application". *Weather*, **45(10)**: 375–382.
[30] Maletic, J.I. and Marcus, A. (2005). "Data cleansing data mining and knowledge discovery handbook", Springer Science-Business Media.
[31] Malhi, A. and Gao, R.X. (2004). "PCA-based feature selection scheme for machine defect classification. IEEE Transaction on Instrumentation and Measurement", **53(6)**: 1517–1525.
[32] Kaiser, H.F. (1960). "The application of electronic computers to factor analysis". *Educational and psychological measurement*.
[33] Todorov, Hristo., Fournier, David., and Gerber, Susanne. (2018). "Principal components analysis: theory and application to gene expression data analysis". Genomics and Computational Biology, **4(2)**: e100041-e100041.
[34] O'Rourke, N., & Hatcher, L. (2013). "A step-by-step approach to using SAS for factor analysis and structural equation modeling" (2nd Ed.). Cary, NC: SAS Press.
[35] Kohonen, T. (1998). "The self-organizing map". *Neurocomputing*, **78(9)**: 1464–1480.
[36] Vesanto, J., Himberg, J., Alhoniemi, E., & Parhankangas, J. (2000). SOM toolbox for Matlab 5. Helsinki University of Technology, Finland.
[37] Westerlund, M.L. (2005). "Classification with Kohonen Self-Organizing Maps". *Soft Computing*, Haskoli Islands.
[38] Pölzlbauer, Georg. (2004). "Survey and comparison of quality measures for self-organizing maps". 67–82.
[39] Martinovi˘, J. (2013). "Effective clustering algorithm for high-dimensional sparse data based on SOM". 131–147.
[40] Chattopadhyay, Manojit, Pranab K. Dan, and Sitanath Mazumdar. (2012). "Application of visual clustering properties of self-organizing map in machine-part cell formation". *Applied Soft Computing Journal*. **12(2)**: 600–610.
[41] Kenekayoro, P., Buckley, K. and Thelwall, M. (2014). "Clustering research group website homepages". *Scientometrics*, 2023–2039.
[42] Deborah, L.J., Baskaran, R. and Kannan, A. (2010). "A survey on internal validity measure for cluster validation". *International Journal of Computer Science & Engineering Survey (IJCSES)*, **1(2)**: 85–102.
[43] Rendón, E., Abundez, I., Arizmendi, A., & Quiroz, E. (2011). "Internal versus external cluster validation indexes". *International Journal of computers and communications*, **5(1)**: 27-34.
[44] Rousseeuw and J, P. (1987). "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". *Journal of Computational and Applied Mathematics*, (**20**): 53–65.
[45] Davies, D. L., & Bouldin, D. W. (1979). "A cluster separation measure". *IEEE transactions on pattern analysis and machine intelligence*, (**2**): 224-227.
[46] Dunn† and C, J. (1974). "Well-separated clusters and optimal fuzzy partitions". *Journal of Cybernetics*, **4(1)**: 95–104.
[47] Desgraupes, Bernard. (2013). "Clustering indices". *University of Paris Ouest-Lab Modal'X*, **1:34**.