



# Application of learning analytics using clustering data Mining for Students' disposition analysis

Sanyam Bharara<sup>1</sup>  · Sai Sabitha<sup>1</sup> · Abhay Bansal<sup>1</sup>

Received: 30 April 2017 / Accepted: 3 September 2017 / Published online: 5 October 2017  
© Springer Science+Business Media, LLC 2017

**Abstract** Learning Analytics (LA) is an emerging field in which sophisticated analytic tools are used to improve learning and education. It draws from, and is closely tied to, a series of other fields of study like business intelligence, web analytics, academic analytics, educational data mining, and action analytics. The main objective of this research work is to find meaningful indicators or metrics in a learning context and to study the inter-relationships between these metrics using the concepts of Learning Analytics and Educational Data Mining, thereby, analyzing the effects of different features on student's performance using Disposition analysis. In this project, K-means clustering data mining technique is used to obtain clusters which are further mapped to find the important features of a learning context. Relationships between these features are identified to assess the student's performance.

**Keywords** Learning analytics · Educational data mining · Disposition analytics · Academic analytics · Learning management systems

## 1 Introduction

Learning analytics is the measurement, assembly, scrutiny and reporting of data about learners and their learning frameworks. LA helps for the purpose of enhancing learning and the surroundings in which it happens. It provides the stakeholders (teachers, principals, parents and students) with appropriate and faster feedback about learning

---

✉ Sanyam Bharara  
bharasanyam@gmail.com

Sai Sabitha  
saisabitha@gmail.com

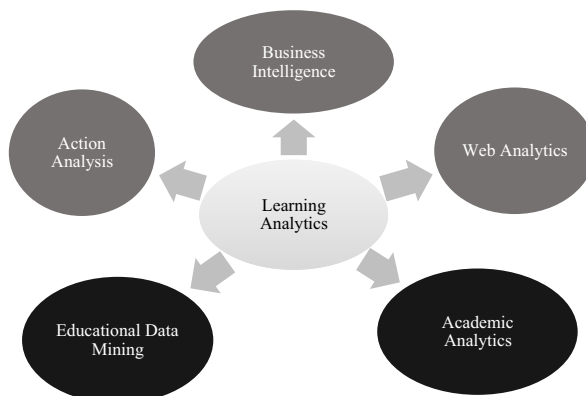
Abhay Bansal  
abhaybansal@gmail.com

<sup>1</sup> Amity University Uttar Pradesh, Noida, 201301 Uttar Pradesh, India

processes. For example, real-time feedback about learners' progress is obtained by tracing and analyzing learners' activities through actionable dashboards (Erik Duval 2011). This capacity forms the basis for reshaping educational models, and has its impact on current and future education models. This analytics helps in improving teaching techniques, learning activities and gain hidden knowledge about learners (Romero and Ventura 2010).

Administrators may use data to detect the areas that needs improvement and the learning resources that can be delivered to the learners most effectively to improve their learning (Zorrilla et al. 2005). LA is a confluence of fields like Business Intelligence, Action Analytics, Web analytics, Educational data mining and Academic analysis. In Fig. 1 various fields of Learning Analytics are depicted. The various applications of LA are discussed as follows: a) Student retention is a well-established application of learning analytics and the “Purdue Signals” software system is used to build predictive models about academic standing of learners. b) “eAdvisor” provides an insight about their progression during a particular course. c) The other learning management systems – such as “Blackboard Analytics-3” and “Desire2Learn Insights-4” helps in understanding the interactions of students (Desire2Learn 2012). d) Inspecting the usage of a learning system by students, and reviewing student achievements helps teachers to identify the possible patterns and take judgment on the future plan of the learning action (Mandinach 2012; Mandinach and Jackson 2012). e) Assessment and Feedback, also is an important application of LA. “Smart Feedback” delivers remarkable information produced on the basis of user's data about their welfares and the learning background (Hattie and Timperley 2007).

**Educational data mining** (EDM) can be explained as processes designed for the analysis of data from educational settings for better understanding of students and the environment in which they learn in. The relationship between Learning Analytics and Educational Data Mining can be explained as ‘EDM is a combination of Learning Analytics and Academic Analytics’. In EDM specifically, researchers focus on intelligent tutoring paradigms whereas LA is more intensive towards enterprise learning systems, i.e. Learning Management Systems (LMS)



**Fig. 1** Different fields of Learning Analytics

and Content Management Systems (CMS). Open Educational Resources (OER) – free digitized material for students and self-learners, are important in improving the teaching and learning processes (Sabitha et al. 2016a). These OERs can also be used for the analytics as it caters diverse needs of students and supports different approaches for learning.

Disposition Analytics are dispositions conveyed to the learning environment by the student. They help us to understand student inclination to take up a course and also helps in understanding how a student is probable to react to the course-associated intrusions. Disposition can also be termed as “Tendency to behave in a certain way”, or “A by-product of learning activity”.

The main aim of our research work is to find meaningful features in a learning context to study the inter-relationships between the found metrics using the concepts of Learning Analytics and Educational Data Mining, thereby, analyzing the effects of different features on student’s performance using Disposition analysis.

The contribution of this paper is further divided as follows: Section II sheds light on the literature survey on Learning Analytics, Educational Data Mining and Clustering data mining techniques. Section III - sheds light on experimental design and analysis.

## 1.1 Theoretical background

**Learning Analytics:** Learning analytics (LA) can be illustrated as an informative application of web analytics intended at profiling the learners, a procedure of collecting and examining details of distinct student interactions in virtual learning actions. LA consists of analytic tools used to develop learning and education.

**Educational Data Mining:** Educational data mining is evolving as a research field with a group of computational as well as psychological approaches and exploration methods for understanding how scholars learn. New computer-oriented interactive learning approaches and tools—smart tuition systems, imitations, games—have unlocked openings to gather and inspect student data, to explore patterns and fashions in those data, and to make new findings and test theories about how students learn. Data of large number of scholars can be fetched from online LMS which may comprise of many variables which can be discovered by data mining procedures for model construction.

## 1.2 Review of previous works

### 1.2.1 Learning analytics and educational data mining

Learning has been defined as a product of interaction and the learning design consists of learners, instructors, content, tutors, etc. Learning Analytics is all about – Effectiveness of the course, catering to the needs of the students, supporting the needs of learners in a better way, effectiveness of interactions, and further improvements in it (Lias and Elias 2011). As a large quantity of educational assets moved on web, a huge amount of data containing these collaborations became accessible. This is principally true because of distance education, in which a much greater proportion of communications are computer-oriented (Arora et al. 2017).

LA also helped to overcome the uncertainty of allocating resources, developed competitive advantages, and also improved quality and value of learning experience. Analytics and Big Data have an important role to perform in the future of higher education as discussed in (Siemens and Long 2011).

LA has also been used for visualization and recommendation. A dashboard for students and teachers is also provided by Information visualization techniques. Also recommendation helps in dealing with paradox of choice and helps converting abundance of resources from a problem into an asset of learning (Duval 2011). LA can also be applied to the reusable content in a personalized and authentic learning experience achieved by integrating Knowledge Objects of Knowledge Management System (KMS) and Learning Objects of Learning Management System (Sabitha et al. 2016a).

LA has been explained as an important field of Technology Enhanced Learning (TEL) in (Ferguson 2012). It discusses the relations between learning analytics, educational data mining and academic analytics. Two research societies - EDM and Learning Analytics & Knowledge (LAK) have established separately for enlarged, high-quality research into the models, approaches, technologies, tools and impact of analytics (Siemens and d Baker 2012).

“Course Signals” is a system that acts as a resolution to allow teachers, the chance to employ the influence of learner analytics to deliver instant real-time feedback to a student. It depends not only upon grades to forecast students’ performance, but also demographic features, past educational history, and scholars’ effort as measured by communication with “Blackboard Vista”- Purdue’s LMS (Arnold and Pistilli 2012).

The reference model for LA in (Chatti et al. 2012) and (Chatti et al. 2014) discusses the future research in LA based on four dimensions – data, environments, context (what?), stakeholders (who?), objectives (why?), and methods (how?).

A generic framework is designed in (Greller and Drachsler 2012) that can act as a valuable guide to setting up learning analytics services in sustenance of educational rehearsal and learner leadership, in quality assurance and cultivating teacher efficacy and effectiveness.

The relationships between LA and EDM are discussed in (Baker and Inventado 2014) and also how these two communities have grown with a common interest of exploiting big data for the benefit of education and learning.

### *1.2.2 Previous work using LA dataset*

According to the previous work done by Amrieh et al. (2016), on the Learning Analytics Dataset, the student’s behavioral features were used and analyzed to understand the effects of the student’s interaction with LMS.

The above work highlighted that a large number of students of low-level grades and middle-level grades were highly interactive with LMS as compared to students of other grade levels, as depicted in its Fig. 3 titled ‘Educational Stages Visualization’ (Page 125).

Also, according to the above work, the important subjects involving high interaction between LMS and students were listed as - IT, French, Arabic, Biology and Science, as depicted in its Fig. 4 titled ‘Educational Topics Visualization’ (Page 125).

The above predictions were done using data mining techniques like Bagging, Boosting and Random Forest methods (Amrieh et al. 2016).

The following Table 1 discusses the related works in the fields of LA and EDM and Table 2 discusses the related works in the fields of LA and EDM involving interactional features.

### 1.3 Major challenges in learning analytics

As per (Chatti et al. 2014), some of the major challenges are:

- An important challenge in LA is to collect and integrate raw data from numerous, heterogeneous sources, frequently obtainable in different layouts, to draft a useful scholastic dataset that imitates the dispersed activities of the learner. This is required due to an increasing trend of converging learning content of KMS with LMS by various Data mining techniques (Sabitha et al. 2016b). Policies and best exercises on binding big data in TEL have to be discovered and shared by the LA research community.
- The ethical and confidentiality challenges of the participants must be kept into consideration while attempting to institutionalize LA and comprise it into the day-to-day learning actions.
- The challenges in Personalized Learning Analytics (Goal-oriented Learning Analytics) are to describe the correct Goal / Indicator / Metric (GIM) triple beforehand beginning the LA exercise to get effective analytical results.
- A big challenge is to find an appropriate lifelong learner modeling. This model is a store for the assembly of learning data about a distinct learner. It would also comprise both long-term and short-term aims of the learner.
- Another challenge is Open assessment which is an all-inclusive term joining different valuation methods for identifying learning in open and networked learning surroundings.

**Table 1** Research Works in EDM and Learning Analytics

AuAuthor	Area of Research / Technique
Educational Data Mining	
Heiner et al. 2007	To test the projective accuracy of a prevailing difficulty metric using SpellBEE peer-tutoring system and to build alternate metrics that use collected data to attain an improved fit
Learning Analytics	
Lias and Elias 2011	Defining LA and how it focusses on effectiveness of educational factors.
Siemens and Long 2011	LA with Big Data in higher schooling
Ferguson 2012	LA as an area of Technology Enhanced Learning and relationship between LA, EDM and Academic Analytics
Siemens and d Baker 2012	EDM and Learning Analytics Knowledge as two separate research communities with a common interest.
Chatti et al. 2012	The four dimensions of LA (What? Who? Why? How?) are discussed using a reference model

**Table 2** Research works in LA and EDM involving Interactional features

Author	Area of Research / Technique	Features Selected
<b>Learning Analytics</b>		
Zorrilla et al. 2005	Administrative use of LA to achieve high quality outcomes	Administrative Usage Features
Hattie and Timperley 2007	Smart feedback on the basis of data about the operator's benefits and the learning context	Learning Context
Romero and Ventura 2010	Learning Analytics used to improve learning process and learning activities	Learning Process and Activities
Duval 2011	Visualization and Recommendation techniques of Learning Analytics for assets of learning	Relationship with assets of learning
Arnold and Pistilli 2012	Purdue's Course Signals is a system which uses LA to deliver real-time feedback to a scholar	Real-time feedback
Desire2Learn; 2012	Learning analytics based on the students' interaction with their LMS	Interactional Features
Greller and Drachsler 2012	Generic Framework to setup LA services in support of educational practice and learner guidance	Learner Guidance
Mardinach and Jackson 2012	Detecting patterns and Decision making by analyzing students' endeavors and use of LMS	Usage of LMS
Shahiri et al. 2015	Learning Analytics applied on educational data using classification technique	Interactional Features
Amrieh et al. 2016	Student's Performance Prediction model using Bagging, Boosting and Forest methods.	Behavioral features
<b>Educational Data Mining</b>		
Baker and Yacef 2009	Focuses on increase in prediction and reduction in frequency of relationship mining within EDM	Relationship mining features
Scheuer and McLaren 2012	Educational Data collected from students interactions with an educational system and administrative and demographic data	Interactional features

The major challenges considered in our research work are to find the significant indicators or metrics and to use mixed-method assessment methods.

#### 1.4 Different clustering techniques used in previous works

Various clustering techniques have been already used in the fields of E-learning (Sabitha et al. 2016a, b, Sabitha et al. 2017; Tian et al. 2008; Jili et al. 2009), Business (Kuhlmann et al. 2003; Cohn and Hull 2009; Chen et al. 2012; Bharara et al. 2017), Knowledge (Fayyad et al. 1996; Liao et al. 2008; Bharara et al. 2017) and Learning Analytics.

To assess the different clustering techniques used in the previous works in the fields of Learning Analytics and Educational Data Mining, various publications from journals and conferences (listed in Table 3), are referred in order to get a clear idea of how each clustering technique has been put into use to achieve different goals.

Tables 4 and 5 illustrates the research works related to Learning Analytics and Educational Data Mining, in which different types of Clustering techniques like K-means, C-means, Fuzzy K-means, K-prototypes, Fuzzy Clustering, Co-operative, PSO farthest first, and Expectation Maximization, Agglomerative, Markov Clustering, etc. are used for data mining and analysis.

These findings can then be collaborated as in Tables 6, 7, and 8 in order to find out which clustering technique is best suitable so as to reach our goal. Also, Figs. 2 and 3 shows pie chart distribution of different clustering techniques used in E-learning and EDM respectively.

## 2 Methodology

This research work introduces a students' performance model with a new category of features, which are called Interactional features. Kalboard 360 – a Learning Management System (LMS) is used to collect the educational dataset. Clustering data mining techniques are used in this model to assess the effect of student's interactional features and student's parental involvement features on student academic performance. Also, data collection and preprocessing steps are used to apprehend the nature of these kind of features.

Following are the steps of methodology adopted in this research work:

Step 1: Collection of data.

Step 2: Analyzing each and every feature and its implication, so as to better understand what all features to be selected.

**Table 3** Research papers selection process

Sources	No. of papers selected	No. of relevant papers referred
IEEE	15	10
ACM	5	2
Conferences	18	15
Others	22	5

**Table 4** Different Clustering techniques used in previous works of LA

S. No	Author	Year	Research Problem	Clustering Technique
1	Chen et al	2007	To identify learning performance assessment rules	K-means
2	Zheng et al	2007	To determine the optimal parameters and partitions for clustering algorithms	K-means, Discrete Markov Model
3	Feng et al	2008	To provide personalized e-learning environment on learner personality	K-means, Fuzzy
4	Jili et al	2009	To classify the e-learning behavior of learners	Fuzzy
5	Chu et al	2009	An adaptive learning case reference method for problem-based e-learning on maths teaching for scholars with slight disabilities	K-prototypes
6	Zhao et al	2010	Clustering Access Patterns in E-learning Environment	Fuzzy K-means
7	Chellatamilan et al	2011	Consequence of Mining educational Data to enhance Adaptation of learning in e-Learning System	K-means
8	Ghorbani and Montazer	2012	To group learners on the basis of their cognitive flairs of learning	K-means, C-means
9	Cobo et al	2012	To model learner's contribution profile in online debate forums	Agglomerative
10	Eranksi and Moudgalya	2012	Determining the impact of human features on user preferences	K-means
11	Draždilová et al	2008	To derive social-network graphs in student e-learning activities	Agglomerative
12	Antonenko et al	2012	To cluster the e-learning behavior of learners	Agglomerative
13	Kizilcece et al	2013	Scrutinizing Learner Subpopulations in MOOCs	K-means
14	Aher and Lobo	2012	Recommending best course combination to student	K-means
15	Valsamidis et al.	2012	Analyze the weblog data of Learning Management System	Markov



**Table 5** Different Clustering techniques used in research works of EDM

S. No.	Author	Year	Research Problem	Clustering Technique
16	Salazar et al.	2004	To identify variables influencing performance of undergraduate students	C-means
17	Wook et al	2009	To evaluate undergraduate performance in semester exam	PSO farthest first
18	Tie et al	2010	Teaching Method of Fundamentals Course of PC based on Cluster Analysis	K-means
19	Cobo et al	2010	A Strategy based on Time Series and Agglomerative Hierarchical Clustering	Agglomerative
20	Zheng and Jia	2011	Educational data clustering method based on better particle swarm optimize	K-means
21	Banumathi and Pethalakshmi	2012	Upgrading Indian Education by Using Data Mining	K-means
22	Parack et al	2012	Predicting Academic Trends and Patterns	K-means
23	Lahane et al	2012	Divisive approach of Clustering for Educational Data	K-means
24	Bovo et al	2013	Clustering Moodle data as a device for sketching students	Expectation Maximization
25	Govindarajan et al	2013	Continuous Clustering in Big Data Learning Analytics	Cooperative
26	Wijayanto	2015	Clustering approach on elementary school intakes and outputs qualities	Agglomerative

**Table 6** Depiction of different clustering techniques used in E-learning

Algorithm type	Technique	Published Papers (S.No. wise)	Frequency
Non-Hierarchical Algorithm	K-means	1, 2, 3, 4, 5, 6, 7, 8	8
	C-means	3	1
	Fuzzy K-means	3, 9	2
	K-prototypes	10	1
	Fuzzy Clustering	8, 11	2
Hierarchical type algorithm	Agglomerative Clustering	12, 13, 14	3
	Markov Clustering	15	1
	Discrete Markov Model (DMM)	4	1

Step 3: Preprocessing of data is done which basically includes Normalizing of some attributes and also Dimensionality reduction.

Step 4: Data cleaning is performed after the preprocessing step.

Step 5: Selection of features to define the input data, delete inappropriate data and find new features category.

Step 6: Applying k-means clustering algorithm on the selected features so as to find clusters of students with similar features.

## 2.1 Data collection

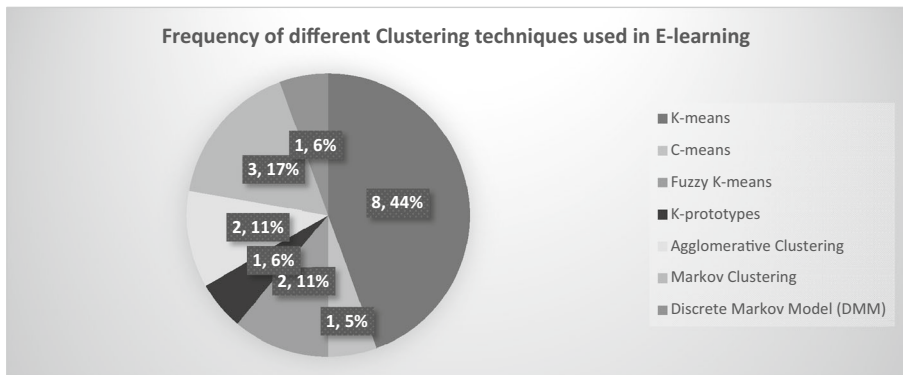
The data is collected from a dataset named “Students’ Academic Performance Dataset” from [Kaggle.com](https://www.kaggle.com) (Kaggle 2016). The original source of the dataset is - Elaf Abu Amrieh, Thair Hamtini, and Ibrahim Aljarah, The University of Jordan, Amman, Jordan, <http://www.Ibrahimaljarah.com>, [www.ju.edu.jo](http://www.ju.edu.jo). The increase of internet usage in learning has fashioned a new context known as web-based education or LMS. The Learning Management System is a digital framework that manages and simplifies online learning. The LMS is mainly used to manage learners, observe student involvement, tracking students’ progress throughout the system. The LMS also assigns and manages learning assets such as registration,

**Table 7** Depiction of different clustering techniques used in EDM

Algorithm type	Technique	Published Papers (S.No. wise)	Frequency
Non-Hierarchical Algorithm	K-means	16, 17, 18, 19,20	5
	C-means	21	1
	Co-operative	22	1
	PSO Farthest first	23	1
	Expectation Maximization[EM]	24	1
Hierarchical type algorithm	Agglomerative Clustering	25, 26	2

**Table 8** Student Features and their Description

Features Category	Feature	Feature Description
Demographical Features	Nationality	Nationality of the student (Kuwait, Lebanon, etc.)
	Gender	Gender of the student (Male/Female)
	Place of Birth	Birth place of the student (Saudi Arabia, Iran, USA, Jordan, Kuwait, Lebanon)
	Educational Stage/ School level	Student's educational stage (Primary, middle and high school levels)
Academic Features	Grade Level	Student's Grade level (G-01,02,03,04,05,06,07,08,09,10,11, 12)
	Section ID	Student's section (A,B,C)
	Semester	Student's school year semester as (First or second)
	Topic	Course subject (Math, Arabic, Science, Quran, English, IT)
	Student Absence Days	Student absence days (Above-7, Under-7)
Parental Involvement Features	Parent Answering Survey	Is Parent answering the surveys provided from school or not
	Parent School Satisfaction	The Amount of parent satisfaction from school as follow (Good, Bad)
	Parent responsible for student	Student's responsibility is of father or mum
Interactional Features	Discussion groups	Student Actions during interaction with Kalboard 360 e-learning system.
	Visited resources	
	Raised hands	
	Viewing announcements	

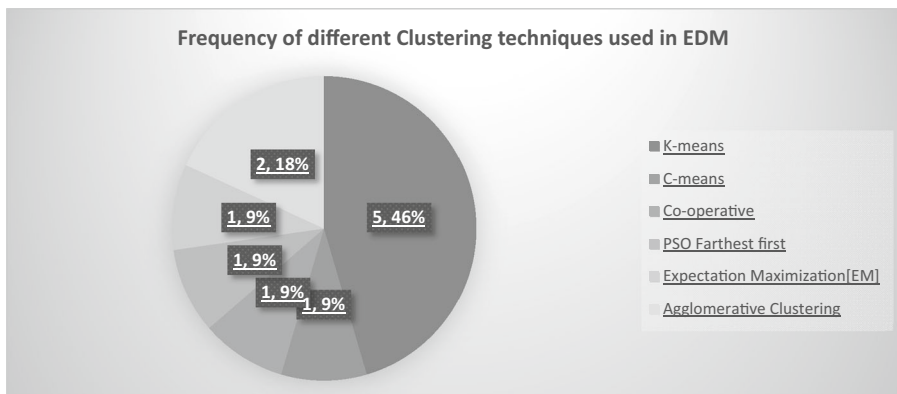


**Fig. 2** Pie chart distribution of different clustering techniques used in E-learning

classroom and the online learning delivery. In this project, we have collected an educational data set from an LMS called Kalboard 360. It is a multi-agent LMS, designed to facilitate education through the usage of leading-edge technology. In such a system users can synchronously access the educational resources using any device with Internet connectivity. In addition it also involves parents and school-management in the learning experience which makes it a truly extensive process, as it connects and engages all parties. A learner activity tracker tool, called experience API (xAPI) is used to collect the data. The xAPI is an element of the Training and Learning Architecture (TLA) that allows to observe learning progress and learners' activities like reading an artifact or viewing a training video. The xAPI helps the learning activity providers to determine the learner, activity and things that define a learning know-how. The aim of xAPI in this project is to monitor student interaction through the educational process for evaluating the features that may have an effect on student's performance.

## 2.2 Educational dataset

In this project that data set has a collection of 500 students with 16 features (Kaggle 2016). These features are categorized into three main classes: (1)



**Fig. 3** Pie chart distribution of different clustering techniques used in EDM

Demographic features such as gender and nationality, (2) Academic features such as educational Stage, grade Level and section, (3) Parental Involvement features such as parent Answering Survey, parent Responsible and Parent School Satisfaction, (4) Interactional features, such as raised hand on class, visited resources, Announcements Viewed. All of these features focus on learner and parent progress on LMS. Whereas, some other features which focus on Students' performance are Student's Grade Marks and Student's Absence Days.

Table 8 shows the dataset's features and their description and Fig. 4 shows a screenshot of the dataset considered.

### 2.3 Feature analysis

The student's performance is affected by many features like the gender variances feature. It has been approved that the aptitudes of students are different and depend on gender (Putrevu 2001). In previous works it was also found that most of female learners have a positive learning style when compared to male learners (Meit et al. 2004). Other research works state that male students have a positive opinion of e-learning compared to female students (Ong and Lai 2006). For the parent responsible feature, different studies have shown that there is a positive affiliation between the parent's tutoring and student's performance (Ermisch and Francesconi 2001) and it is mainly valid when the student is being followed up by their mother (Agus and bin Mohamed Makhbul 2002). Then school attendance feature is an important feature in scholastic success. It is also known that there is a direct relation amongst good attendance and student achievement (Rothman 2001). These researches prove the positive relation between such features: gender, parent responsible, school attendance and students' performance (DeKalb 1999). This project considers new type of features, called interactional features. These feature are linked to the learner interaction with educational system. Student interaction is one of the main research work in educational psychology field.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
	gender	Nationality	PlaceOfBirth	StageID	GradeID	SectionID	Topic	Semester	Relation	raisedhands	VisitedResources	Announcements	Discussion	ParentAnswered	Parentschool	StudentAbsenceDays	Class
1	M	KW	Kuwait	lowerleve	G-04	A	IT	F	Father	15	16	2	20	Yes	Good	Under-7	M
2	M	KW	Kuwait	lowerleve	G-04	A	IT	F	Father	20	20	3	25	Yes	Good	Under-7	M
3	M	KW	Kuwait	lowerleve	G-04	A	IT	F	Father	10	7	0	30	No	Bad	Above-7	L
4	M	KW	Kuwait	lowerleve	G-04	A	IT	F	Father	30	25	5	35	No	Bad	Above-7	L
5	M	KW	Kuwait	lowerleve	G-04	A	IT	F	Father	40	50	12	50	No	Bad	Above-7	M
6	M	KW	Kuwait	lowerleve	G-04	A	IT	F	Father	42	30	13	70	Yes	Bad	Above-7	M
7	M	KW	Kuwait	lowerleve	G-04	A	IT	F	Father	35	12	0	17	No	Bad	Above-7	L
8	M	KW	Kuwait	MiddleSci	G-07	A	Math	F	Father	35	12	0	17	No	Bad	Above-7	L
9	M	KW	Kuwait	MiddleSci	G-07	A	Math	F	Father	50	10	15	22	Yes	Good	Under-7	M
10	F	KW	Kuwait	MiddleSci	G-07	A	Math	F	Father	12	21	16	50	Yes	Good	Under-7	M
11	F	KW	Kuwait	MiddleSci	G-07	B	IT	F	Father	70	80	25	70	Yes	Good	Under-7	M
12	M	KW	Kuwait	MiddleSci	G-07	A	Math	F	Father	50	88	30	80	Yes	Good	Under-7	H
13	M	KW	Kuwait	MiddleSci	G-07	B	Math	F	Father	19	6	19	12	Yes	Good	Under-7	M
14	M	KW	Kuwait	lowerleve	G-04	A	IT	F	Father	5	1	0	11	No	Bad	Above-7	L
15	M	lebanon	lebanon	MiddleSci	G-08	A	Math	F	Father	20	14	12	19	No	Bad	Above-7	L
16	F	KW	Kuwait	MiddleSci	G-08	A	Math	F	Mum	62	70	44	60	No	Bad	Above-7	H
17	F	KW	Kuwait	MiddleSci	G-06	A	IT	F	Father	30	40	22	66	Yes	Good	Under-7	M
18	M	KW	Kuwait	MiddleSci	G-07	B	IT	F	Father	36	30	20	80	No	Bad	Above-7	M
19	M	KW	Kuwait	MiddleSci	G-07	A	Math	F	Father	55	13	35	90	No	Bad	Above-7	M
20	F	KW	Kuwait	MiddleSci	G-07	A	IT	F	Mum	69	15	36	96	Yes	Good	Under-7	M
21	M	KW	Kuwait	MiddleSci	G-07	B	IT	F	Mum	70	50	40	99	Yes	Good	Under-7	H
22	F	KW	Kuwait	MiddleSci	G-07	A	IT	F	Father	60	60	33	90	No	Bad	Above-7	M
23	F	KW	Kuwait	MiddleSci	G-07	B	IT	F	Father	10	12	4	80	No	Bad	Under-7	M

Fig. 4 Screenshot of educational dataset

Student engagement was defined as “the quality and quantity of students’ cognitive, psychological, behavioral and emotional reactions to the learning procedure and also to in-class/out-of-class educational and social activities to attain effective learning results” (Gunuc and Kuzu 2015). Student engagement comprises not only the spent time on chores but also their aspiration to take part in some activities (Stovall 2003).

## 2.4 Data preprocessing

For preprocessing of data, normalization mechanism is used by which the students’ performance is converted from numerical values into nominal values. To complete this step, the data set is split into three nominal intermissions (High Level, Medium Level and Low Level) based on student’s total grade/mark such as: values from 0 to 69 fall under Low Level interval, values from 70 to 89 fall under Middle Level interval and values from 90 to 100 fall under High Level interval. Also the Interactional features of this dataset i.e. Raised Hands, Announcements Viewed, Discussions participated and Resources visited are normalized using the 0–1 normalization formula which is given as:  $z_i = (x_i - \min(x)) / (\max(x) - \min(x))$ , where  $x = (x_1, \dots, x_n)$  and  $z_i$  is now your  $i^{\text{th}}$  normalized data. Then these interactional features are categorized into three nominal intervals (High level, Medium level and Low level) based on their count such as: values from 0 to 33 fall under Low Level interval, values from 34 to 66 fall under Middle Level interval and values from 67 to 100 fall under High Level interval. This research work uses normalization so as to scale the attributes values into a small range [0.0 to 1.0]. Figure 5 illustrates the normalization in values of various attributes.

## 2.5 Data cleaning

One of the main preprocessing tasks is the Data cleaning which when applied on this data set, removes irrelevant items and missing values. The data set of this

serial num	gender	Place of Birth	GradeID	Topic	raisedHands	VisitedResources	AnnouncementsView	Discussions	ParentAnswer	ParentschoolSatisf	StudentAbsenceDays	Relation	Grade-mark	Cluster
1	M	Kuwait	G-04	IT	0.15	0.161616162	0.00164945	0.193977551	Yes	Good	Under-7	Father	M	cluster_1
2	M	Kuwait	G-04	IT	0.2	0.202020202	0.00206341	0.244897959	Yes	Good	Under-7	Father	M	cluster_1
3	M	Kuwait	G-04	IT	0.1	0.070707071	0.000721501	0.295918367	No	Bad	Above-7	Father	M	cluster_2
4	M	Kuwait	G-04	IT	0.3	0.252525253	0.002576788	0.346938776	No	Bad	Above-7	Father	M	cluster_2
5	M	Kuwait	G-04	IT	0.4	0.505050505	0.005153577	0.5	No	Bad	Above-7	Father	M	cluster_2
6	F	Kuwait	G-04	IT	0.42	0.303030303	0.003092146	0.704081633	Yes	Bad	Above-7	Father	M	cluster_1
7	M	Kuwait	G-07	Math	0.35	0.121212121	0.001236858	0.165265096	No	Bad	Above-7	Father	L	cluster_2
8	M	Kuwait	G-07	Math	0.5	0.101010101	0.001030715	0.214285714	Yes	Good	Under-7	Father	M	cluster_1
9	F	Kuwait	G-07	Math	0.12	0.212121212	0.002164502	0.5	Yes	Good	Under-7	Father	M	cluster_1
10	F	Kuwait	G-07	IT	0.7	0.808080808	0.008245723	0.704081633	Yes	Good	Under-7	Father	M	cluster_1
11	M	Kuwait	G-07	Math	0.5	0.888888889	0.00907295	0.806122449	Yes	Good	Under-7	Father	H	cluster_0
12	M	Kuwait	G-07	Math	0.19	0.060606061	0.000618429	0.11224498	Yes	Good	Under-7	Father	M	cluster_1
13	M	Kuwait	G-04	IT	0.05	0.01010101	0.000103072	0.102040816	No	Bad	Above-7	Father	L	cluster_2
14	M	Lebanon	G-08	Math	0.2	0.141414141	0.001434001	0.183673469	No	Bad	Above-7	Father	L	cluster_2
15	F	Kuwait	G-08	Math	0.62	0.707070707	0.007215007	0.602040816	No	Bad	Above-7	Mum	H	cluster_0
16	F	Kuwait	G-06	IT	0.3	0.404040404	0.004122861	0.663265306	Yes	Good	Under-7	Father	M	cluster_1
17	M	Kuwait	G-07	IT	0.36	0.303030303	0.003092146	0.806122449	No	Bad	Above-7	Father	M	cluster_2
18	M	Kuwait	G-07	Math	0.55	0.131313131	0.001339993	0.908163265	No	Bad	Above-7	Father	M	cluster_2
19	F	Kuwait	G-07	IT	0.69	0.151515152	0.001546073	0.969387755	Yes	Good	Under-7	Mum	M	cluster_1
20	M	Kuwait	G-07	IT	0.7	0.505050505	0.005153577	1	Yes	Good	Under-7	Mum	H	cluster_0
21	F	Kuwait	G-07	IT	0.6	0.606060606	0.006184292	0.908163265	No	Bad	Above-7	Father	M	cluster_2
22	F	Kuwait	G-07	IT	0.1	0.121212121	0.001236858	0.806122449	No	Bad	Under-7	Father	M	cluster_1

Fig. 5 Screenshot of dataset values after preprocessing (Normalization and Dimensionality Reduction)

project comprises 20 missing values in several features from 500 records, thus these records with missing values are deleted from the data set, and so the data set was now of 480 records.

## 2.6 Feature selection

Feature selection is an important step in data preprocessing field. The objective of this process is to select a suitable subclass of features which can competently define the input data, decreases the dimensionality of feature space, and deletes redundant and inappropriate data. This process can help in improving the data quality, therefore the performance of the learning algorithm. According to Table 2, it can be seen that the Interactional features were considered for research in the field of Learning Analytics and Educational Data Mining. In accordance with the review, our focus will be on the Interactional features so as to understand these features and their relationships with other features like Demographic features and Academic features. The newly defined category of features which is extracted from the dataset are Parental Involvement features which mainly include following features - Parent Responsibility, Parent Answering Survey and Parent School Satisfaction. Table 9 depicts the features used in previous research works and feature considered in this research work.

## 3 Experiment and design

### 3.1 Clustering technique

The data mining approach i.e. K-means clustering is best suitable for the chosen dataset and is better than any other data mining approach because the students' interactional behavior is heterogeneous in nature so it is necessary to form tighter clusters of students with similar interactional characteristics. Also, it is proven that K-means clustering gives the best result when dataset are distinct or well separated from each other. Value of 'K' was determined by using direct method involving Average Silhouette method as shown in Fig. 6(a). The average silhouette width of the whole dataset was found out to

**Table 9** Selection of features

Features considered in existing research works	Features considered in this research work
Learning Context	Parental Involvement
Real-time feedback	Interactional features
Interactional Features	Demographic features
Learner Guidance	Academic features
Usage of LMS	
Behavioral features	
Relationship mining features	

be 0.624 at  $K = 3$ , as shown in Fig. 6(b). Using this method, the suitable value of  $K$  was found out to be 3 as the width as highest at  $K = 3$ .

The Clustering data mining technique was applied on the dataset given in Fig. 5. The attributes considered are – Parental features (Parent Responsibility, Parent Answering Survey and Parent School Satisfaction), Interactional features (Raised hands, Visited resources, Announcements Viewed, Discussions participated), Academic features (Students' Absenteeism, Students' Grade Marks, Grade level, Grade ID, Course Topic), and Demographic features (Place of birth, Gender). The clusters obtained are discussed further:

(a)

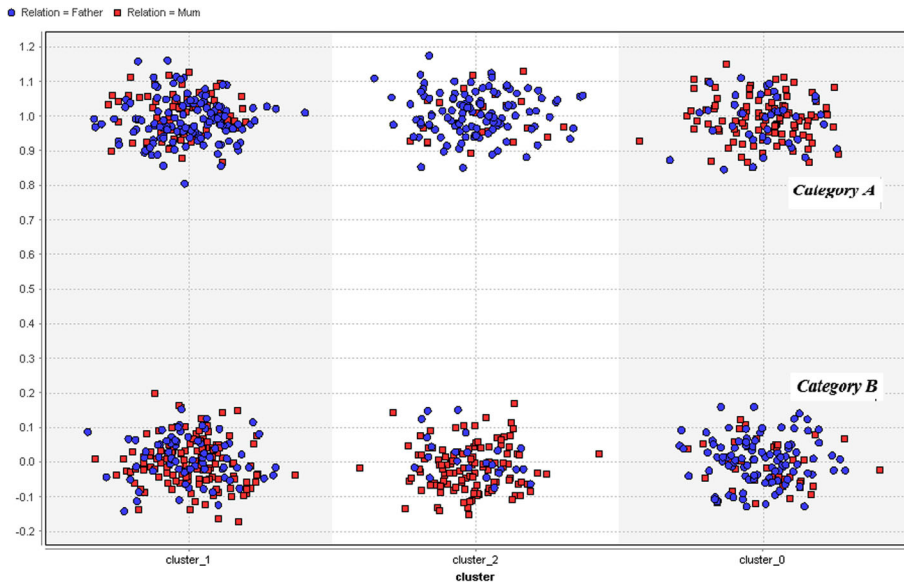
Row No.	Id	cluster	neighbour	silhouette
1	1	1	0	0.667
2	2	1	0	0.669
3	3	1	0	0.671
4	4	1	0	0.672
5	5	1	0	0.674
6	6	1	0	0.676
7	7	1	0	0.678
8	8	1	0	0.679
9	9	1	0	0.681
10	10	1	0	0.683
11	11	1	0	0.685
12	12	1	0	0.686
13	13	1	0	0.688
14	14	1	0	0.690
15	15	1	0	0.691
16	16	1	0	0.693
17	17	1	0	0.695
18	18	1	0	0.696
19	19	1	0	0.698
20	20	1	0	0.699
21	21	1	0	0.701

(b)

Criterion	Value
Average silhouette width	0.624

**Fig. 6** a applying average silhouette method on the dataset at  $K = 3$  (b) finding the maximum average silhouette width of the dataset





**Fig. 7** Clusters formation based on Parent responsibility

### 3.2 Clusters formation and analysis

The three clusters are analyzed using the Scatter Multiple plot on the basis of 5 different aspects which are: Students' Interaction, Parent Responsibility, Students' Absenteeism, Students' Grade Marks, and Parent-School Relationship.

#### 3.2.1 Cluster analysis based on parent responsibility

X-axis: Cluster number.

Y-Axis: Parent Responsibility (Father or Mother).

In Fig. 7, the upper 3 clusters form category A, which are the actual clusters with their attributes value as '1', whereas the bottom 3 clusters in category B have their attribute value as '0'. Category A consists of positive values and category B consists of negative values, thus the category B is counterproductive. Therefore, all the analysis is done taking into consideration clusters of category A.

Cluster 0 – The cluster shows that majority of it is formed for Mother being responsible for student's education.

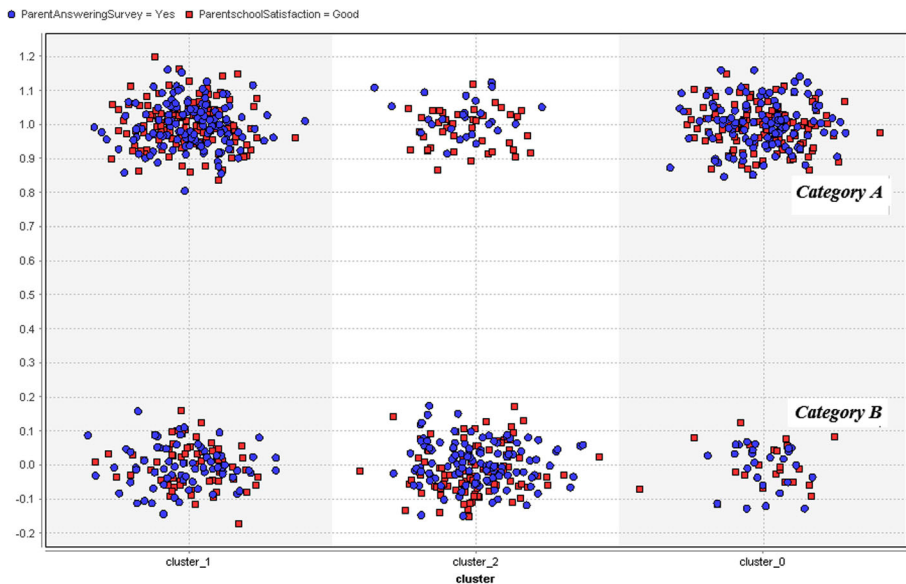
Inference: Cluster 0 is Mother-oriented in terms of student's education responsibility.

Cluster 1 - The cluster shows there is equality of cluster values for Mother as well as Father being responsible for student's education.

Inference: Cluster 1 is equal-oriented in terms of student's education responsibility.

Cluster 2 - The cluster shows that majority of it is formed for Father being responsible for student's education.

Inference: Cluster 2 is Father-oriented in terms of student's education responsibility.



**Fig. 8** Clusters formation based on Parent-school involvement

### 3.2.2 Cluster analysis based on parent-school relationship

X-axis: Cluster number.

Y-Axis: Parent Involvement Factors (Parent Answering Survey, Parent School Satisfaction).

In Fig. 8, the upper 3 clusters form category A, which are the actual clusters with their attributes value as '1' i.e. Parent Answering Survey = Yes and Parent School Satisfaction = Good, whereas the bottom 3 clusters in category B have their attribute value as '0' i.e. Parent Answering Survey = No and Parent School Satisfaction = Bad. Category A consists of positive values and category B consists of negative values, thus the category B is counterproductive. Therefore, all the analysis is done taking into consideration clusters of category A.

Cluster 0 – The cluster shows that majority of values for Parent Answering Survey are 'Yes' and also values for Parent School Satisfaction are 'Good'.

Inference: Cluster 0 reflects highly positive parental involvement.

Cluster 1 – The cluster shows that majority of values for Parent Answering Survey are 'Yes' and also values for Parent School Satisfaction are 'Good'.

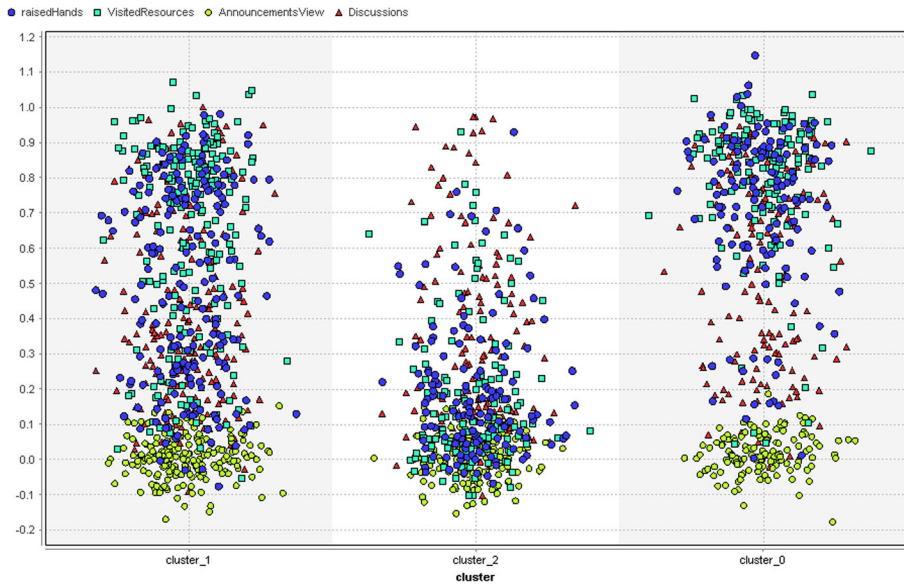
Inference: Cluster 1 reflects positive parental involvement.

Cluster 2 - The cluster shows that majority of values for Parent Answering Survey are 'No' and also values for Parent School Satisfaction are 'Bad'.

Inference: Cluster 2 reflects negative parental involvement.

### 3.2.3 Cluster analysis based on students' interaction

(Refer Fig. 9) X-axis: Cluster number.

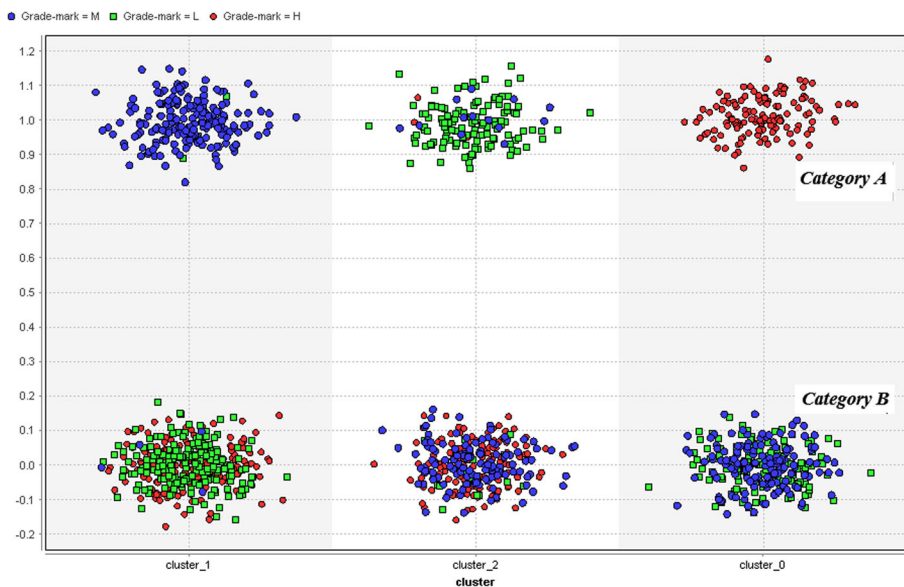


**Fig. 9** Clusters formation based on Students' Interaction

Y-axis: Interactional features (Raised Hands, Announcements viewed, Discussions, Visited Resources).

Cluster 0 - The cluster shows that majority of values for all the interactional features are increasing in density as we go from '0' (No) to '1' (Yes).

Inference: Cluster 0 reflects high interaction between students and LMS.



**Fig. 10** Clusters formation based on Students' Grade Marks

Cluster 1 - The cluster shows that majority of values for all the interactional features, the density is equally distributed as we go from '0' (No) to '1' (Yes).

Inference: Cluster 1 reflects both high as well as low interaction between students and LMS.

Cluster 2 - The cluster shows that majority of values for all the interactional features are decreasing in density as we go from '0' (No) to '1' (Yes).

Inference: Cluster 2 reflects low interaction between students and LMS.

### 3.2.4 Cluster analysis based on Student's grade marks

X-axis: Cluster number.

Y-axis: Student's Grade Marks High -  Medium -  Low - 

In Fig. 10, the upper 3 clusters form category A, which are the actual clusters with their attributes value as '1', whereas the bottom 3 clusters in category B have their attribute value as '0'. Category A consists of positive values and category B consists of negative values, thus the category B is counterproductive. Therefore, all the analysis is done taking into consideration clusters of category A.

Cluster 0 - It can be clearly seen that in this cluster, the majority of values of Students' Grade

Marks are  - High grade marks

Inference: Cluster 0 reflects high grade marks of students in this cluster.

Cluster 1 - It can be clearly seen that in this cluster, the majority of values of Students' Grade

Marks are  - Medium grade marks

Inference: Cluster 1 reflects medium grade marks of students in this cluster.

Cluster 2 - It can be clearly seen that in this cluster, the majority of values of Students' Grade

Marks are  - Low grade marks


Inference: Cluster 2 reflects low grade marks of students in this cluster.

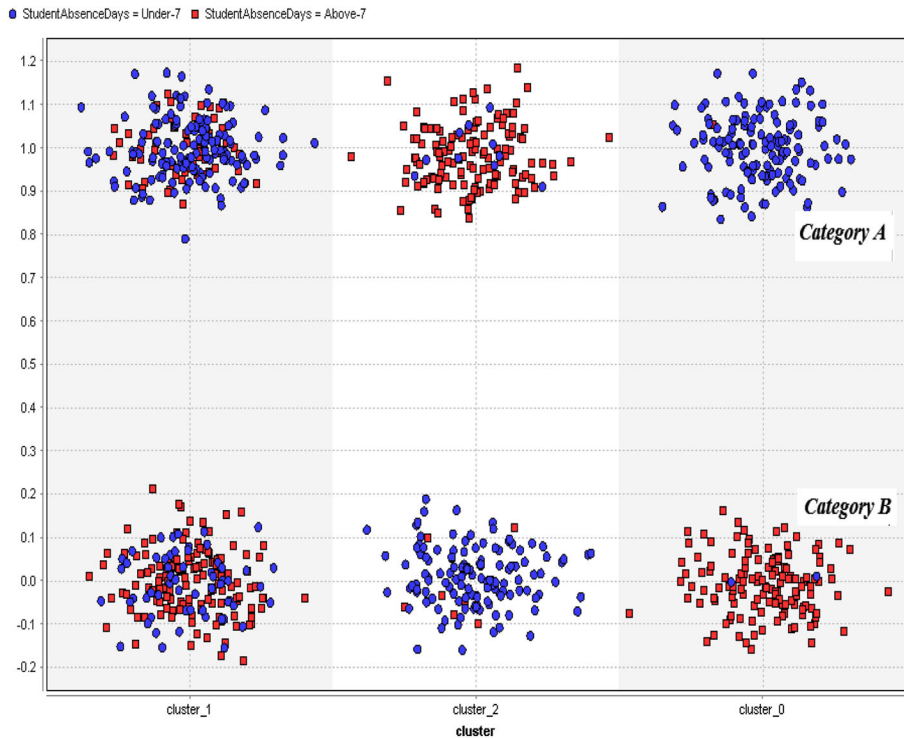
### 3.2.5 Cluster analysis based on students' absenteeism

X-axis: Cluster number.

Y-axis: Student's Absence Days  - Under 7  - Above 7

In Fig. 11, the upper 3 clusters form category A, which are the actual clusters with their attributes value as '1', whereas the bottom 3 clusters in category B have their attribute value as '0'. Category A consists of positive values and category B consists of negative values, thus the category B is counterproductive. Therefore, all the analysis is done taking into consideration clusters of category A.

Cluster 0 - It can be clearly seen that in this cluster, the majority of values of Students' Absence Days are  - i.e. no. of absence days 'Under-7'. Inference: Cluster 0 reflects low absenteeism of students in this cluster.



**Fig. 11** Clusters formation based on Students' Absenteeism

Cluster 1 - It can be clearly seen that in this cluster, the majority of values of Students' Absence Days are ● - i.e. no. of absence days 'Under-7'.

Inference: Cluster 1 reflects low absenteeism of students in this cluster.

Cluster 2 - It can be clearly seen that in this cluster, the majority of values of Students' Absence Days are ■ - i.e. no. of absence days 'Above-7'.

Inference: Cluster 2 reflects high absenteeism of students in this cluster.

### 3.3 Overall clusters description

For an overall evaluation, refer to Tables 10, 11, and 12. In Tables 10 and 11, the values of various attributes are depicted statistically and all the critical values of each and every feature are highlighted for all the three clusters separately. In Table 12, the three clusters are compared on the basis of various attributes like – Parent Responsibility, Parent Answering Survey, Parent School Satisfaction, Students' Interactions, Students' Grade Marks, and Students' Absenteeism. Both the tables are then mapped and analyzed to find out possible and valid outcomes.

### 3.4 Summary of analysis

From the above tables, Tables 10, 11 and 12, we can clearly see that disposition factors i.e. Students' Interaction (Raising hands, Discussion participation, Viewing

**Table 10** Cluster Analysis using the statistical ratios for features including Interactional Features, and Demographic features

Features	Cluster 0	Cluster 1	Cluster 2
No. of Records	141	196	143
Raised Hands			
Low	10.3%	36.7%	82.2%
Medium	18.7%	23.1%	14.3%
High	71%	40.2%	3.5%
Visited Resources			
Low	5.2%	19.3%	82.4%
Medium	14.1%	19.3%	15.5%
High	80.7%	60.4%	2.1%
Announcements Viewed			
Low	24.1%	44.8%	87.3%
Medium	41.4%	35.2%	12%
High	34.5%	20%	0.7%
Discussions			
Low	32.6%	44.3%	62.5%
Medium	23.6%	32.3%	22.8%
High	43.8%	23.4%	14.7%
Gender			
Male	53.2%	62.7%	81%
Female	46.8%	37.3%	19%
Birth Place			
Egypt	2.1%	1%	2.7%
Iran	0%	1.5%	2%
Iraq	9.9%	4.1%	0%
Jordan	36.8%	43.3%	27.2%
Kuwait	24.8%	35.2%	53%
Lebanon	7.8%	3%	1.3%
Libya	0%	0%	4%
Morocco	0.6%	1%	0.6%
Palestine	4.2%	2.2%	0%
Saudi Arabia	4.4%	2.2%	3.4%
Syria	1.4%	1.5%	0.6%
Tunis	2.1%	1.5%	2.1%
USA	4.2%	3.5%	2.1%
Venezuela	0.7%	0%	0%

Announcements, Visited Resources count) with LMS affects the students' performance and grade marks. We can also see that Parental Involvement factors affect students' performance as well.

Some of the important points to be focused upon are:

1. Kuwait men are more responsible than women for their children's education.
2. Jordan women are more responsible than men for their children's education.

**Table 11** Cluster Analysis using the statistical ratios for features including Academic features, Parental features, Students' Absenteeism and Students' Grade Marks

Features	Cluster 0	Cluster 1	Cluster 2
No. of Records	141	196	143
Grade ID			
02	29%	27.5%	36.3%
04	9.9%	8.6%	11.8%
06	9.2%	6.6%	4%
07	20.5%	20.4%	22.3%
08	23.4%	30.6%	16%
09	0%	2.1%	0.6%
10	0.7%	1.1%	0.6%
11	4.5%	2.1%	2%
12	2.8%	1%	3.4%
Topic			
Arabic	13.4%	11.7%	11.8%
Biology	11.3%	4.5%	2.8%
Chemistry	7.1%	4%	4%
English	11.3%	7.1%	10.4%
French	14.2%	14.7%	11.1%
Geology	4.2%	9.1%	0%
History	2.8%	6.1%	2%
IT	10.6%	18.8%	30%
Math	4.2%	2.5%	7%
Quran	5.6%	4%	4%
Science	11.3%	10.4%	10.4%
Spanish	3%	6.1%	5.5%
Parent Answering Survey			
Yes	80.8%	66.3%	18.2%
No	19.2%	33.6%	81.8%
Parent School Satisfaction			
Good	83.6%	66.4%	69%
Bad	16.4%	33.6%	31%
Student Absence Days			
Above-7	2.8%	27.5%	93%
Under-7	97.2%	72.5%	7%
Parent Responsible			
Father	29%	62.2%	84%
Mum	71%	37.8%	16%
Grade Marks			
High	100%	0%	89.7%
Medium	0%	2%	6.7%
Low	0%	98%	3.6%

### 3. Grade-2, 7, 8 students are more involved in using the LMS.

The above statement can be proved by the cluster-wise bar graphs in Figs. 11 and 12 which shows that in each cluster the number of students are high in grade-

**Table 12** Clusters analysis based on all selected features

Features	Cluster 0	Cluster 1	Cluster 2
Parent Responsible	Mother	Both	Father
Parent School Satisfaction	Yes	Yes	No
Parent Answering Survey	Good	Good	Bad
Students' Interaction	High	Medium	Low
Grade Marks	High	Medium	Low
Students' Absenteeism	Low	Medium	High

levels - 2,7 and 8. A similar finding has been obtained in a previous research work (Amrieh et al. 2016).

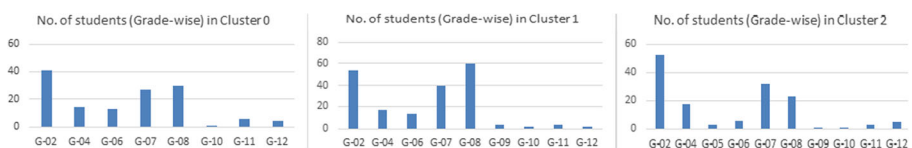
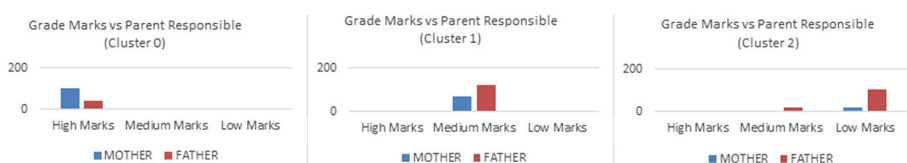
- Arabic, biology, French, science and IT are the subjects in which the students are highly interacting with LMS.

The above statement can be proved by referring to Table 11. Also, a similar finding has also been obtained in a previous research work (Amrieh et al. 2016).

- Fathers responsible for the students' education are less involved in surveys and those students are with less attendance.
- Mothers responsible for the students' education leads to excellent performance of student.

The above statement can be proved by Fig. 13 which consists of cluster-wise bar graphs depicting Mother's participation in student's education yielding high marks in cluster 0 whereas father's participation is related to low or medium marks of students as shown in cluster 1 and 2.

All these inferences are based upon the chosen dataset of 480 students. The actual trend may differ from our analysis and can be predicted with the help of another dataset with more values from numerous other Learning Management Systems and E-Learning systems.

**Fig. 12** Bar graphs representing grade-wise count of students in each cluster**Fig. 13** Cluster-wise bar graphs representing the effect of parent's participation on student's grade marks



## 4 Conclusion and future scope

The above case study shows that both the features, Students' Interaction with the LMS and Parents' Involvement in Students' education, are important features of Learning Analytics. It was also observed that the Parental Involvement features and Interactional features vary with Demographic features, and they also affect the Students' Learning Behavior, and in turn it affects the Students' Overall performance by affecting the Students' Grade Marks and Absenteeism.

Thus, it can be concluded that the Students' Dispositions Analytics are responsible for Students' Performance and these dispositions are highly dependent upon the Parental Involvement features and Demographic features.

This cluster-based approach to LA can be used in practice and is applicable at classroom-level as well as distance learning level. At classroom level, a LMS can be used in a lecture to improve the learning process and a similar data can be collected from this LMS as given in the used dataset using an activity tracker tool. Also, a distance learning management system can be designed with an in-built activity tracker tool, and access to this system can be provided to all the students enrolling in distance learning programs. This cluster-based approach to LA is very beneficial in order to collect heterogeneous data of learners and can be helpful in suggesting possible improvements in course design and delivery.

In future, further analysis can be carried on to use mixed-method evaluation approaches to study the inter-relationships between the different features. Also, various other clustering techniques like DBScan, Agglomerative etc. can be further used to improve the clusters and help achieve different outcomes for numerous other Learning management systems. Further analysis can also be performed to study the trends in various educational systems which in turn can help in improving the learning systems and quality of education.

The other features which can be incorporated for Disposition Analysis are Behavioral features, Learner guidance features, Relationship mining features and Administrative LMS usage features. These features can be considered in the extension work. Also, the application of Learning Analytics in the field of Curriculum Design maybe considered for future scope.

## References

- Agus, A., & bin Mohamed Makhbul, Z. K. (2002). An empirical study on academic achievement of business students in pursuing higher education: An emphasis on the influence of family backgrounds. *New paradigm of borderless education: challenges, strategies, and implications for effective education through localization and*, 168.
- Aher, S. B., & Lobo, L. (2012, August). Applicability of data mining algorithms for recommendation system in e-learning. In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics* (pp. 1034–1040). ACM.
- Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining educational data to predict Student's academic performance using ensemble methods. *International Journal of Database Theory and Application*, 9(8), 119–136.
- Antonenko, P. D., Toy, S., & Niederhauser, D. S. (2012). Using cluster analysis for data mining in educational technology research. *Educational Technology Research and Development*, 60(3), 383–398.

- Arnold, K. E., & Pistilli, M. D. (2012, April). Course signals at Purdue: Using learning analytics to increase student success. In Proceedings of the 2nd international conference on learning analytics and knowledge (pp. 267–270). ACM.
- Arora, S., Goel, M., Sabitha, A. S., & Mehrotra, D. (2017). Learner groups in massive open online courses. *American Journal of Distance Education*, 31(2), 80–97.
- Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In Learning analytics (pp. 61–75). Springer New York.
- Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM-Journal of Educational Data Mining*, 1(1), 3–17.
- Banumathi, A., & Pethalakshmi, A. (2012, January). A novel approach for upgrading Indian education by using data mining techniques. In Technology Enhanced Education (ICTEE), 2012 I.E. International Conference on (pp. 1–5). IEEE.
- Bharara, S., Sabitha, A. S., & Bansal, A. (2017, January). A review on knowledge extraction for business operations using data mining. In Cloud Computing, Data Science & Engineering-Confluence, 2017 7th International Conference on (pp. 512–518). IEEE.
- Bovo, A., Sanchez, S., Héguy, O., & Duthen, Y. (2013, September). Clustering moodle data as a tool for profiling students. In e-Learning and e-Technologies in Education (ICEEE), 2013 Second International Conference on (pp. 121–126). IEEE.
- Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2012). A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, 4(5–6), 318–331.
- Chatti, M. A., Lukarov, V., Thüs, H., Muslim, A., Yousef, A. M. F., Wahid, U., ... & Schroeder, U. (2014). Learning analytics: Challenges and future research directions. Retrieved June, 24, 2016 on, vol. 37, pp. 1349–1359, 2007.
- Chellatamilan, T., Ravichandran, M., Suresh, R. M., & Kulanthaivel, G. (2011, July). Effect of mining educational data to improve adaptation of learning in e-learning system. In Sustainable Energy and Intelligent Systems (SEISCON 2011), International Conference on (pp. 922–927). IET.
- Chen, C. M., Chen, Y. Y., & Liu, C. Y. (2007). Learning performance assessment approach using web-based learning portfolios for e-learning systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 37(6), 1349–1359.
- Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, 36(4), 1165–1188.
- Chu, H. C., Chen, T. Y., Lin, C. J., Liao, M. J., & Chen, Y. M. (2009). Development of an adaptive learning case recommendation approach for problem-based e-learning on mathematics teaching for students with mild disabilities. *Expert Systems with Applications*, 36(3), 5456–5468.
- Cobo, G., Garcia-Solórzano, D., Santamaría, E., Morán, J. A., Melenchón, J., & Monzo, C. (2010, June). Modeling students' activity in online discussion forums: A strategy based on time series and agglomerative hierarchical clustering. In Educational Data Mining 2011.
- Cobo, G., Garcia-Solórzano, D., Morán, J. A., Santamaría, E., Monzo, C., & Melenchón, J. (2012, April). Using agglomerative hierarchical clustering to model learner participation profiles in online discussion forums. In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (pp. 248–251). ACM.
- Cohn, D., & Hull, R. (2009). Business artifacts: A data-centric approach to modeling business operations and processes. *IEEE Data Engineering Bulletin*, 32(3), 3–9.
- DeKalb, J. (1999). *Student absence without permission (Student Truancy)*. ERIC Digest.
- Desire2Learn (2012). Desire2Learn Client Success Story: Austin Peay State University. Retrieved from [http://content.brightspace.com/wp-content/uploads/Desire2Learn\\_Success\\_Story-Degree-Compass-APSU.pdf](http://content.brightspace.com/wp-content/uploads/Desire2Learn_Success_Story-Degree-Compass-APSU.pdf).
- Dráždilová, P., Martinovic, J., Slaninová, K., & Snášel, V. (2008, December). Analysis of relations in eLearning. In Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 03 (pp. 373–376). IEEE computer society.
- Duval, E. (2011, February). Attention please!: Learning analytics for visualization and recommendation. In Proceedings of the 1st International Conference on Learning Analytics and Knowledge (pp. 9–17). ACM.
- Eranki, K. L., & Moudgalya, K. M. (2012, July). Evaluation of web based behavioral interventions using spoken tutorials. In Technology for Education (T4E), 2012 I.E. Fourth International Conference on (pp. 38–45). IEEE.
- Ermisch, J., & Francesconi, M. (2001). Family matter: Impacts of family background on educational attainment. *Economica*, 68, 137–156.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), 37.

- Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5–6), 304–317.
- Ghorbani, F., & Montazer, G. A. (2012, February). Learners grouping improvement in e-learning environment using fuzzy inspired PSO method. In E-Learning and E-Teaching (ICELET), 2012 Third International Conference on (pp. 65–70). IEEE.
- Govindarajan, K., Somasundaram, T. S., & Kumar, V. S. (2013, December). Continuous clustering in big data learning analytics. In Technology for Education (T4E), 2013 I.E. Fifth International Conference on (pp. 61–64). IEEE.
- Greller, W., & Drachsler, H. (2012). Translating learning into numbers: A generic framework for learning analytics. *Educational Technology & Society*, 15(3), 42–57.
- Gunuc, S., & Kuzu, A. (2015). Student engagement scale: Development, reliability and validity. *Assessment & Evaluation in Higher Education*, 40(4), 587–610.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>.
- Heiner, C., Heffernan, N., & Barnes, T. (2007, July). Educational data mining. In Supplementary Proceedings of the 12th International Conference of Artificial Intelligence in Education.
- Jili, C., Kebin, H., Feng, W., and Huixia, W. (2009). E-learning behavior analysis based on fuzzy clustering. In genetic and evolutionary computing, 2009. WGECC '09. 3rd International conference on, Guilin, 2009, (pp. 863–866).
- Kaggle (2016). Students' Academic Performance Dataset: xAPI-Educational Mining Dataset for Data Science. <https://www.kaggle.com/aljarah/xAPI-Edu-Data>.
- Kizilcec, R. F., Piech, C., & Schneider, E. (2013, April). Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. In Proceedings of the third international conference on learning analytics and knowledge (pp. 170–179). ACM.
- Kuhlmann, M., Shohat, D., & Schimpf, G. (2003, June). Role mining-revealing business roles for security administration using data mining technology. In Proceedings of the eighth ACM symposium on Access control models and technologies (pp. 179–186). ACM.
- Lahane, S. V., Kharat, M. U., & Halgaonkar, P. S. (2012, November). Divisive approach of clustering for educational data. In Emerging Trends in Engineering and Technology (ICETET), 2012 Fifth International Conference on (pp. 191–195). IEEE.
- Liao, S. H., Chen, C. M., & Wu, C. H. (2008). Mining customer knowledge for product line and brand extension in retailing. *Expert Systems with Applications*, 34(3), 1763–1776.
- Lias, T. E., & Elias, T. (2011). Learning Analytics.
- Mandinach, E. B. (2012). A perfect time for data use: Using data-driven decision making to inform practice. *Educational Psychologist*, 47(2), 71–85.
- Mandinach, E. B., & Jackson, S. S. (2012). *Transforming teaching and learning through data-driven decision making*. Corwin: Thousand Oaks.
- Meit, S. S., Borges, N. J., Cubic, B. A., & Seibel, H. R. (2004). Personality differences in incoming male and female medical students. Online Submission.
- Ong, C. S., & Lai, J. Y. (2006). Gender differences in perceptions and relationships among dominants of e-learning acceptance. *Computers in Human Behavior*, 22(5), 816–829.
- Parack, S., Zahid, Z., & Merchant, F. (2012, January). Application of data mining in educational databases for predicting academic trends and patterns. In Technology Enhanced Education (ICTEE), 2012 I.E. International Conference on (pp. 1–4). IEEE.
- Putrevu, S. (2001). Exploring the origins and information processing differences between men and women: Implications for advertisers. *Academy of Marketing Science Review*, 2001, 1.
- Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 40(6), 601–618.
- Rothman, S. (2001). School absence and student background factors: A multilevel analysis. *International Education Journal*, 2(1), 59–68.
- Sabitha, A. S., Mehrotra, D., & Bansal, A. (2016a). Delivery of learning knowledge objects using fuzzy clustering. *Education and Information Technologies*, 21(5), 1329–1349.
- Sabitha, A. S., Mehrotra, D., Bansal, A., & Sharma, B. K. (2016b). A naive bayes approach for converging learning objects with open educational resources. *Education and Information Technologies*, 21(6), 1753–1767.
- Sabitha, A. S., Mehrotra, D., & Bansal, A. (2017). An ensemble approach in converging contents of LMS and KMS. *Education and Information Technologies*, 22(4), 1673–1694.

- Salazar, A., Gosalbez, J., Bosch, I., Miralles, R., & Vergara, L. (2004). A case study of knowledge discovery on academic achievement, student desertion and student retention. In *Information Technology: Research and Education*, 2004. ITRE 2004. 2nd International Conference on (pp. 150–154). IEEE.
- Scheuer, O., & McLaren, B. M. (2012). Educational data mining. In *Encyclopedia of the Sciences of Learning* (pp. 1075–1079). Springer US.
- Shahiri, A. M., & Husain, W. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, 72, 414–422.
- Siemens, G., & d Baker, R. S. (2012, April). Learning analytics and educational data mining: Towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 252–254). ACM.
- Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *Educause Review*, 46(5), 30.
- Stovall, I. (2003). Engagement and online learning. *UIS community of practice for e-Learning*, 3, 2014.
- Tian, F., Wang, S., Zheng, C., & Zheng, Q. (2008, April). Research on e-leamer personality grouping based on fuzzy clustering analysis. In *Computer Supported Cooperative Work in Design*, 2008. CSCWD 2008. 12th International Conference on (pp. 1035–1040). IEEE.
- Tie, Z., Jin, R., Zhuang, H., & Wang, Z. (2010, June). The research on teaching method of basics course of computer based on cluster analysis. In *Computer and Information Technology (CIT)*, 2010 I.E. 10th International Conference on (pp. 2001–2004). IEEE.
- Valsamidis, S., Kontogiannis, S., Kazanidis, I., Theodosiou, T., & Karakos, A. (2012). A clustering methodology of web log data for learning management systems. *Educational Technology & Society*, 15(2), 154–167.
- Wijayanto, F. (2015, November). Indonesia education quality: Does distance to the capital matter?(a clustering approach on elementary school intakes and outputs qualities). In *Science and Technology (TICST)*, 2015 International Conference on (pp. 318–322). IEEE.
- Wook, M., Yahaya, Y. H., Wahab, N., Isa, M. R. M., Awang, N. F., & Seong, H. Y. (2009, December). Predicting NDUM student's academic performance using data mining techniques. In *Computer and Electrical Engineering*, 2009. ICCEE'09. Second International Conference on (Vol. 2, pp. 357–361). IEEE.
- Zhao, J. W., Gu, S. M., & He, L. (2010, June). A novel approach to clustering access patterns in e-learning environment. In *Education Technology and Computer (ICETC)*, 2010 2nd International Conference on (Vol. 1, pp. V1–393). IEEE.
- Zheng, X., & Jia, Y. (2011, December). A study on educational data clustering approach based on improved particle swarm optimizer. In *IT in Medicine and Education (ITME)*, 2011 International Symposium on (Vol. 2, pp. 442–445). IEEE.
- Zheng, Q., Ding, J., Du, J., & Tian, F. (2007, April). Assessing method for e-leamer clustering. In *Computer Supported Cooperative Work in Design*, 2007. CSCWD 2007. 11th International Conference on (pp. 979–983). IEEE.
- Zorrilla, M. E., Menasalvas, E., Marin, D., Mora, E., & Segovia, J. (2005, February). Web usage mining project for improving web-based learning sites. In *International Conference on Computer Aided Systems Theory* (pp. 205–210). Springer Berlin Heidelberg.