 doi:10.20944/preprints201808.0350.v2

Article

Integration of Data Mining Clustering Approach with the Personalized E-Learning System

Samina Kausar [1], Xu Huahu [1], Iftikhar Hussain [2,3,*], Zhu Wen Hao [1] and Misha Zahid [3]

[1] School of Computer Engineering and Science, Shanghai University, Shanghai, 200444, China;
[2] Instituut Voor Mobiliteit (IMOB), Hasselt University, Hasselt, 3500, Belgium;
[3] School of Computer and IT, Beaconhouse National University, Lahore, 54400, Pakistan;
* Correspondence: iftikhar.hussain@uhasselt.be; Tel.: +32-486207755

Abstract: Educational data-mining is an evolving discipline that focuses on the improvement of self-learning and adaptive methods. It is used for finding hidden patterns or intrinsic structures of educational data. In the arena of education, the heterogeneous data is involved and continuously growing in the paradigm of big-data. To extract meaningful information adaptively from big educational data, some specific data mining techniques are needed. This paper presents a clustering approach to partition students into different groups or clusters based on their learning behavior. Furthermore, personalized e-learning system architecture is also presented which detects and responds teaching contents according to the students' learning capabilities. The primary objective includes the discovery of optimal settings, in which learners can improve their learning capabilities. Moreover, the administration can find essential hidden patterns to bring the effective reforms in the existing system. The clustering methods K-Means, K-Medoids, Density-based Spatial Clustering of Applications with Noise, Agglomerative Hierarchical Cluster Tree and Clustering by Fast Search and Finding of Density Peaks via Heat Diffusion (CFSFDP − HD) are analyzed using educational data mining. It is observed that more robust results can be achieved by the replacement of existing methods with CFSFDP-HD. The data mining techniques are equally effective to analyze the big data to make education systems vigorous.

Keywords: Big data; clustering; data-mining; educational data-mining; e-learning; profile learning.

---

1. Introduction

Educational data-mining (EDM) is a new perspective in modern educational systems. It is concerned with the study and development of new adaptive methods, instruments to artificially analyze and visualize the hidden patterns or intrinsic structures in educational datasets. Mostly, education related datasets contain structured, semi-structured and un-structured data with different geographical distribution [1]. EDM has emerged as a promising area of research aimed to analyze the intrinsic data structures, extracting hidden predictive information and finding insights into educational datasets [2]. In the field of education, EDM can be demarcated as an application of data-mining methods to exploit novel patterns and artificially analyze big data efficiently and effectively.

Recently, frontier technologies i.e. the Internet-of-Things (IoT), sensors, artificial intelligence, and social networks are being integrated with educational system for effective learning [3,4]. Web based systems are computer-aided virtual forms of instructions that are independent of geographical location. Sensors and IoT produce massive amount of data that lead towards the big data dilemma [5]. However, big data has significant impact in scientific studies, public health, industrial applications, and in the field of education [6–10]. In educational field, the huge amount of data provides a new insight to improve the learning capabilities and decision making skills of teachers and students. The educational data mining may play an important role to improve the education system by (1) refining the individual based quality education, (2) discovering new areas of knowledge and finding associations among different fields and (3) finding the new perspective in experimental data.

With the advancement in communication technologies, nowadays many smart devices and sensors [11] are incorporated into educational systems to observe the overall behavior of the education system. It contains rich information of people's thoughts about different events in semi-structured or unstructured form. Most of the web based learning methods are static and fail to take into account the diversity of students. These virtual educational systems can be improved by utilizing data mining techniques, in order to effectively meet the needs of diverse learners. In general, there is a wide variety of data mining methods that can be applied in the field of education. These methods can be categorized into classification, clustering, neural network, and relationship manning. Clustering is a primary unsupervised method to partition datasets into separate groups (clusters) based on the estimated intrinsic characteristics or similarities [12] and has been applied in several fields [13–18]. Clustering methods can be considered and categorized as: partition-based, density-based, model-based and hierarchy-based [19–23] and cannot be directly used to handle the complexities of big data.

1.1. Research Objectives

This paper presents a data-mining clustering approach to partition students into different groups based on their learning behavior. The offered approach establishes on the basis of big relational databases. A personalized e-learning system architecture integrating undertaken data mining technique is also presented; which creates and responds teaching content according to students' learning capability. The primary objective includes the discovery of optimal settings, in which learners can improve their learning capabilities to boost up their outcomes. Additionally, the administration can find essential hidden patterns to bring the effective reforms in the existing system. This paper also analyzes the K-Means, K-Medoids, Density-based Spatial Clustering of Applications with Noise (DBSCAN), and Agglomerative Hierarchical Cluster Tree (AHCT) approaches for clustering and compares with the recommended approach "Clustering by Fast Search and Finding of Density Peaks via Heat Diffusion (CFSFDP-HD)". A contrast between existing approaches and CFSFDP-HD in regard to academic performance of students is also examined.

1.2. Paper Organization

The organization of this paper is as follows: Section 2 presents the literature review of data mining clustering techniques with some specific tools to deal with education data. Some of the literature on big data and e-learning systems is also presented. Section 3 describes the recommended data mining clustering approach and its integration with the e-learning system architecture. Some of the existing data mining clustering approaches are also discussed. Section 4 presents the experiments and some of the results with discussion by considering a specific case study. Finally, the conclusion and the recommendations for future research are discussed in Section 5.

2. Literature Review

This section presents some of the literature on: (1) educational data-mining approaches with some specific tools to deal with educational data and (2) big data and e-learning systems.

2.1. Clustering techniques in EDM

"Educational data-mining is emerging as a research area with a suite of computational and psychological methods, and research approaches for understanding how students learn" [24]. Various clustering methods have been applied in recent studies to predict students' performance. Alana M. et *al.* in [25] researched techniques to support the teacher decision-making process by grouping students and planning challenges accordingly and reached a positive conclusion. B. R. Prakash, et al. [26] have researched learning analytic techniques for big data in educational data-mining to find out the Adaptive learning systems (ALS). The ALS empowers teachers to rapidly observe the adequacy of their adjustments and mediations, giving input to persistent change. The outcomes of the study are coherent with the conclusions of the study presented by Abdulmohsen Algarni [27].

In [27] author explored various studies and datasets revolving around the field of EDM. Author derived that EDM can be utilized as a part of a wide range of zones including recognizing the students who are at risk, distinguishing needs for the adapting needs of various students' groups, expanding graduation rates, adequately surveying institutional execution, boosting grounds assets, and upgrading subject educational modules reestablishment**.** The outcomes of the study are consistent with the conclusions of the study presented by Bovo, et *al.* [28] after studying student records on different trainings and successfully predicting students who are failing behind. Another research study [29] consistent with [27] is conducted by Amjad Abu Saa examines and predicts student performance in different scenarios using clustering methods. In the similar study [30], Tommaso and Alex Bowers have analyzed various analytical techniques: Educational data-mining, Learning and Academic Analytics, and have reached the conclusion that the applications of data-mining methods (with responsibility and professionalism) yield positive results.

The K-means is a state-of-the-art partition based clustering algorithm and have been applied in EDM. Such as, special selection of student's seat in lab or classroom and its impact on student's assessment has been evaluated by Ivancevic, Celikovic & Lukovic [31]. Another study presented by Ying, et *al.* [32] has utilized K-means to understand the behavior of students based on the annotation dataset of 40 students. In a study conducted by Eranki & Moudgalya [33], K-means was applied to examine the influence of human characteristics on student's performance while listening to music, yielding evident classifications.

2.2. Big Data and E-learning Systems

Big data has the capability to benefit students distinctly by providing them with a modern and dynamic education system. In the study [34], Athanasios and Panagiotis analysed the goals, purposes, and benefits of big-data and open-data in Education. Authors concluded that the education system can be enhanced by embracing new learning approaches to make it more effective and focused on. Moreover, Annapoorna et al. [35], support the same idea and anticipated that the big data can be effectively used in predicting student results, and improving both the teaching and the learning experience. The research conducted by Tulasi [36] and Ben Daniel [37], targeted the higher education and explored the solutions proposed by big data systems to the challenges faced by higher education. Chris Dede [38] further advanced the topic by studying "next steps" that can be taken using big data in education and concluded that the field has a lot of potential in the betterment of the individual learning experiences.

Numerous researchers have expressed that personalization, in an academic setting, permits executing more proficient and viable learning forms. Various works are attempting to enhance the quality and viability of e-learning by utilizing standards of other research zones. This pattern of personalization advancement additionally shows up in e-learning. Matteo et *al.*, [39] have introduced a new tool: Intelligent Web Teacher (IWT) to support Personalized E-Learning in their study on personalized e-learning. The comparison of traditional methods with IWT deduce that personalization permits executing more proficient and powerful e-Learning forms, featuring an expanding level of fulfilment by both educator and students. The system developed by Prawira et al. [40] using Moodle proved to be capable of improving the learning process and collaboration between the teacher and student in higher education.

Maryam Yarandi, et al. [41] take individual learning capabilities of students to present an ontology-based technique to improve an adaptive e-learning scheme. The proposed e-learning system creates content permitting to the learner's knowledge. Andino et al. [42] present a personalized e-learning framework and argue that the system encourages strong learning condition due to the realization of individual needs. The results are coherent with a study by Chun-Hui Wu et al. [43] which presents a theoretical framework of adaptive e-learning, self-assessment and dynamic scaffolding theory. The system provides tailored learning material to students based on the student ability. A grid agent model was proposed by Zhen and Yuying in their study [44] for effective adaptation of e-learning systems using artificial psychology to individual students who would benefit from this personalization. Furthermore, Xin Li and Shi-Kuo Chang [45] have proposed another personalized e-learning system which is a feedback extractor with fusion capability to adjust the user preferences.

The significance of the above mentioned literature being that personalized e-learning schemes are effective tools in individual learning. In the e-learning, a significant huge amount of data is continuously generated and ubiquitously available on the web. Therefore, more sophisticated and frontier clustering methods are required to benchmark on EDM data to get intrinsic insights. To cope with aforementioned issues, a data mining clustering approach is recommended and integrated with the personalized e-learning system. The integration of data-mining approaches makes the learning system more interesting.

3. Data mining Clustering approach and the Personalized E-learning System Architecture

In this paper, a data mining based clustering approach “CFSFDP-HD” is presented to partition students into groups and is established on the basis of big relational databases. The undertaken approach finds possible groups of students by comparing their similar learning behavior. It is sensitive to detect the understanding levels of students. Moreover, Personalized E-Learning System Architecture (PESA) integrating commenced data mining clustering approach is described; which creates and responds teaching content according to students’ learning capability. For each group, system generates different quizzes, assignments, study related games, and books’ contents to improve the learning capabilities of students. To make groups and select appropriate teaching methods, system uses artificial intelligence and adaptive clustering methods. In proposed architecture, the methods K-Means, K-Medoids, Density-based Spatial Clustering of Applications with Noise, Agglomerative Hierarchical Cluster Tree and CFSFDP-HD are used as a profiling and content filtering method to group student into appropriate classes. The traditional e-learning systems are mostly query-based and the queries are responded without any intelligence or heuristics.

3.1. The Educational Environment

A primary agenda of higher education is to harness cross-disciplinary intelligence to improve syllabus, content and delivery, enhancing learners’ experiences and creating an atmosphere that integrates them with the skills and knowledge required to cope with the changes and challenges posed by the big data. In such a complex educational environment, it is tough for human mind to identify patterns manually, but database projects have the abilities to incorporate and link traditional and new data sources. Such compactness can create deeper insights into students learning capabilities and enhance classroom activities.

Grade Point Average (GPA) and percentage score are important indicators for the measurement of students’ academic performance and capabilities. GPA is an important factor for academic planner to setup and analyse the learning environment in the academic organizations [46]. The GPA or percentage score of students can be affected by different factors such as teaching methodology and attention of teachers towards some particular students. It is a general phenomenon that teachers mostly focus on those students that take part in class activities and show satisfactory outputs. Moreover, there are some intrinsic hidden patterns that exist among the students. Students can be separated into different clusters on the basis of their progress. The same teaching method may not be effective for different clusters of students.

Institutional databases, having the teaching material and users’ queries, are entertained according to the stored data. However, most of updated knowledge lies on the Internet at different places. To reinforce the student learning capabilities, it might be credible to integrate the rest of data sources with e-learning system [47]. The data-mining clustering approaches can play significant role to find the relationships among different subjects available over the Internet, specifically in the e-learning systems.

The similarity measures and clustering are important tasks to find the similar groups in educational big data. The similar patterns of data in different fields may be useful for researchers and learners to gain knowledge easily from various fields. For example, we can use partition based clustering, density based clustering, and hierarchical based clustering for text mining, to find the similarity between data points, outliers, and similar or related fields by clustering big data.

3.2. Existing Clustering Methods

Various clustering methods have been used in EDM such as Mean-shift clustering, K-means, K-medoids, Densit Based Spatial Clustering of Applications with Noise (DBSCAN) in [21,48], and Hierarchical clustering in [49–52], however, these approaches are also not robust to identify significant clusters in ambiguous and noisy datasets [22]. A short description on these methods is as follows:

3.2.1. Mean-shift clustering

$\mathrm{Mean-shift\ Clustering}$ is a sliding-window based approach; tries to discover condensed areas of $data - points$. The goal of this approach is to detect the $center - points$ of each cluster based on a centroid method. In the Mean-shift clustering method the candidates are updating for the center points (the mean of the points) within the sliding window. In the post processing stage, the candidate windows are filtered to remove near-duplicates and forms the final set of center-points and their matching clusters.

The Mean-shift clustering having radius 'r' (as the Kernel) begins with a circular sliding window centered at a randomly selected point C. Mean shift method shifts 'r' iteratively to a higher density region (on each step) until convergence. The density of each sliding window is proportional to its size (the points inside it). By shifting, the density of the pints gradually moves towards areas of the higher point density. The shifting of the sliding window continues until a shift cannot accommodate more points inside the kernel (no longer increasing the density). In case, when multiple sliding windows are overlapped; the data points are clustered according to the sliding window in which they reside and the sliding window containing the most well-maintained points.

$\mathrm{Mean-shift\ Clustering}$ method has two disadvantages: (1) it is pretty calculation exhaustive, and (2) it trusts on satisfactory high $\mathrm{data-density}$ (with perfect gradient to find the $cluster - centers$).

3.2.2. K-means

The K-means [19] is a state-of-the-art partition based clustering algorithm. It is considered as a better approach than the Mean-shift clustering because it does not have the above mentioned problems. In K-means, input data is divided into k distinct groups, where k is an input parameter used to specify the number of output clusters. K-means iteratively improves the initial partitions until the optimized clusters are not found. Mathematically we can express K-means using the following expression:

$$\underset{S}{argmin} \sum_{i=1}^{n} \sum_{x \in S_i} ||x - \mu_i||^2 \quad (1)$$

where, μ_i is mean of data-points in S. $\mathrm{S_i}$ is initial partition of dataset $\{\mathrm{x_1, x_2, x_3, \dots, x_n}\}$.

K-means is considered as a best choice to discover the signal of interest from educational datasets if significant number of clusters is already defined. However, it might be a hectic job to discover appropriate groups using K-means without prior knowledge of existing number of clusters or in presence of noisy or complex data. In EDM, it is rigid to setup the selection of clusters, and initial centroids setting of K-means. It is also obscure to find significant signal of interest.

3.2.3. K-medoids

The k-medoids [53] is used to partition the dataset into clusters similar to the k-means. The aim of both (K-means and K-medoids) methods is to minimize the sum of distances between data-points of a cluster and a central data-point of the same cluster. In distinction to the k-means, k-medoids picks data-points by way of centers (the medoids) and workings by a generality of the Manhattan Norm to express distance among data-points. It clusters the dataset of 'n' data-points into 'k' groups or clusters. To determine 'k', the silhouette is considered as a useful tool. K-medoids reduces an amount of pairwise variations instead of summation of squared Euclidean distances. That's why it is considered as more robust (to noise and outliers) than the k-means. The k-medoids method is effective to returns the actual data-points (the medoids) of the dataset. It is also suitable for clustering the definite data where a mean is not exist.

3.2.4. DBSCAN

DBSCAN [21] is a density − based clustering method, begins through a random starting data point that has not stayed visited. The neighbors of the data-point are mined by a Distance − Epsilon ε. If the sufficient amount of $data - points$ is available within neighborhood the clustering procedure starts and the current $data - point$ is considered as the first $data - point$. The other data-points will be labeled as $noisy$. Later on, those noisy data-points may become the portion of the cluster. In both situations the data-points are noted as $visited$. For current cluster, the procedure of making all the data-points in the ε neighborhood is repeated to add all the new data-points. This process is recurring until all data-points in the current cluster are recognized and labeled as $visited$. Same process is repeated for all the clusters.

DBSCAN doesn't execute sound when the clusters of variable density are established. When the density varies, (1) the distance threshold epsilon ε and minimum data-points for identifying the neighborhood data-points will differ from cluster to cluster (2) for very high-dimensional the distance threshold epsilon ε becomes challenging to estimate.

3.2.5. Hierarchical clustering

In this section we will discuss only the Bottom-up hierarchical clustering. It treats each data-point as a single cluster at the beginning and then continuously combines pairs of clusters till entire clusters have been combined into the solo cluster. It is also known as Agglomerative Hierarchical Cluster Tree (AHCT) [54]. The AHCT is represented as a tree. The roots of the tree are considered as a unique cluster.

In the beginning, each data-point is treated as a single cluster; meaning that 'k' data-points are treated as 'k' clusters. In next step a *distance metric* is selected to measure the distance between two clusters. Furthermore the two clusters are combined into one iteratively for each pair. The combined clusters are selected with the smallest average linkage; both the clusters having (i) the smallest distance and (ii) the most similar data-points. This step will continue until the root of the tree which is explicitly given in the start. The number of clusters can be selected by recognizing the given root number, which helps to stop combining the clusters.

Bottom-up hierarchical clustering method does not need to specify the quantity of clusters and has the ability to select the best cluster because of using a tree.

3.3. Recommended Data Mining Clustering Method

The Clustering by Fast Search and Finding of Density Peaks ($CFSFDP$) has been proposed by Alex and Laio [22]. It has characteristics to discover significant clusters in a more spontaneous manner as compared with the K-means. It empowers clustering procedure, in which high-density regions are identified as potential clusters, outliers are automatically identified and arbitrary shape of clusters is organized. In K-means to obtain meaningful clusters users are required to repeat clustering process multiple times with different parametric setting. While the unique approach utilized in $CFSFDP$ to discover clusters and noise adaptively would be a significant clustering tool to analyze the educational data. The $CFSFDP$ uses the following given methodology to discover significant clusters.

$CFSFDP$ calculates local density (ρ_i) and a minimum distance (δ_i) for each given data-point i, with its nearest high density point.

The ρ_i is equivalent to the number of $data-points$ that are closer than the cut-off distance d_c to i. The d_c is a vital parameter used to estimate the ρ_i of each i. The usefulness of $CFSFDP$ depends upon the proper choice of d_c. The local density can be estimated by utilizing the Equation (2) where the d_{ij} is the distance from the $data-point$ i to $-point$ j.

$$\rho_i = \sum_j X(d_{ij} - d_c) \tag{2}$$

where,

$$X(x) = \begin{cases} 1 & x < 0 \\ 0 & otherwise. \end{cases} \tag{3}$$

The local distance δ_i, can be computed by Equation (4), of a $data-point$ to the nearest extremely condensed $data-point$ max_{ρ_i}. It is achieved for the purpose of assigning i to the nearby cluster. The value of δ is considered as maximum, when the $data-points$ with high local or global density are found. On the other hand, the $data-points$ having high ρ_i and large δ_i (compared to other $data-points$) are considered as cluster centers.

$$\delta_i = \begin{cases} \min\limits_{j:\rho_j > \rho_i} (d_{ij}) & if\ \exists\ j\ s.t{:}\ \rho_j > \rho_i \\ \max\limits_{j:\rho_j > \rho_i} (d_{ij}) & otherwise \end{cases} \tag{4}$$

Cluster centers are attained by plotting estimated values of ρ_i and δ_i, which is referred as the decision graph. Moreover, the $CFSFDP$ allocates the remaining $data-points$ to the closest cluster center and based on their δ values. In cluster analysis, the key challenge is to discover correct cluster centers in the datasets [1]. However, $CFSFDP$ uses decision graph to identify the correct $cluster-centers$ with the least human interaction, which makes it more worthy to analyze big data / streaming data.

The $CFSFDP$ has characteristics to discover intrinsic hidden signal of interest from ambiguous data; it can be applied in existing education data mining systems and e-learning systems to produce more significant clusters. It can be used to cluster the similar documents, find plagiarism in documents, analyse the students' profiles and to find the similar insights in different research areas. The $CFSFDP$ via heat diffusion ($CFSFDP-HD$) [20] was proposed as a variant of $CFSFDP$, where limitations of CFSFDP are improved and users can analyse data without any prior domain knowledge. In $CFSFDP-HD$, an adaptive method was used to estimate density of underlying dataset, which is given in Equation (5):

$$\hat{f}(x;t) = \frac{1}{\mathrm{n}} \sum_{j=1}^{n} \sum_{k=-\infty}^{\infty} e^{-k^2\pi^2 t/2} \cos(k\pi x)\ \cos(k\pi x_j) \tag{5}$$

Where x represents the data-points and preparatory probability is distributed through the data-points $\{x_1, x_2, x_3, \ldots, x_n\}$. The method evolves for a time t. The function $\hat{f}$ in Equation 5 can be expressed as

$$\hat{f}(x;t) \approx \sum_{j=0}^{n-1} a_k e^{-k^2\pi^2 t/2} \cos(k\pi x), \tag{6}$$

where n is a positive large interger and a_k is

$$a_k = \begin{cases} 1 & k = 0 \\ \frac{1}{\mathrm{n}} \sum\limits_{i=1}^{n} \cos(k\pi x_i), & k = 1,2,\ldots,n-1, \end{cases} \tag{7}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 October 2018 doi:10.20944/preprints201808.0350.v2

Equation (6) is fully adaptive and may consider; firstly the *optimal bandwidth selection* and secondly the *boundary corrections*. It delivers *enhanced performance* and is also consistent with *actual density*.

3.4. Personalized E-Learning System Architecture

The e-learning architecture responds to the individual demands of users, and is able to predict user preferences or interests. E-learning not only allows the instructors and learners to meet virtually, but also makes sharing of resources possible electronically.

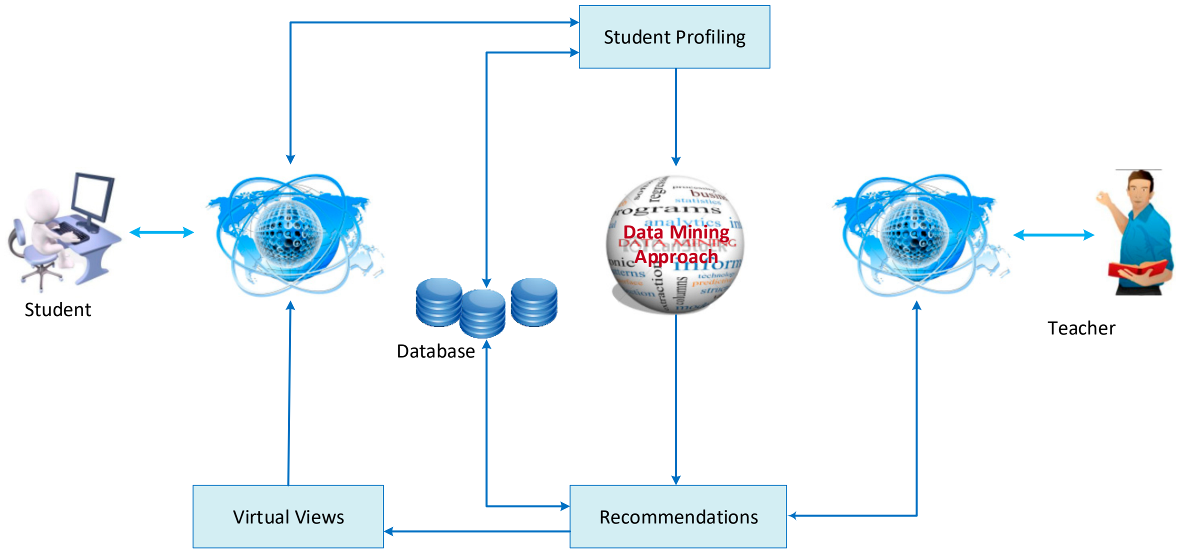


Figure 1: Personalized E-learning Architecture. A profile is created for each learner and is automatically updated based upon the activities of the learner.

The overall Personalized E-learning System Architecture is shown in Figure 1. The major steps of the PESA are described as follows:

3.4.1. Student profiling

The student interacts and manages his/her profile through the interface deployed on a desktop laptop or a smartphone. The user profile and other information seldom change through the Internet. According to [55,56], student profile or sometimes a student model refers to a typical group of students. Its function is to determine the user-learner needs and preferences automatically.

Student related data works like a seed for personalization of student queries and intelligent response of queries. Student profiling is an ongoing process which contains both static and dynamic data. Data collected in a static way [56] includes personal, personality, cognitive, pedagogical and preference data. Individual data define the bio-graphical information about the students. Personality data enlighten the students' attention, cooperation and coordination skills. Student profile reflects the overall interest and behavior of the student. Cognitive data inform about the students' cognition while pedagogical data describe different learning styles and methods. If the profile maintaining system detects any unusual behavior in student activities, it updates the profile accordingly.

3.4.2. Data Mining Clustering Approach

The data mining clustering is responsible to find association, recommendation, and intelligence to provide customized and powerful learning mechanism for students. For example, appropriate content selection on the basis of the students' interest and understanding is a big problem. This can be resolved by grouping whole contents by simply applying clustering approach to filter contents according to individual student profile. Moreover, the key inference components in such e-learning systems are based on data mining approaches, which analyze the user's profile and suggest some sort of actions with the application of artificial intelligence. Moreover, especially, when we talk about clustering methods in existing systems are mostly based on the naïve clustering approaches such as Mean-shift clustering, K-means, and DBSCAN also discussed above. Unlike existing e-learning systems, we recommended to use CFSFDP-HD methods to achieve feasible results. The recommended data mining clustering approach is already explained in the section 3.3.

3.4.3. Recommendations

This process is responsible to collect data of interest from relational databases filtered according to student profile with the help of data mining approach. It also has the ability to prevent duplication of the information created before. This process recommends or proposes the solution to the instructor implicitly based on the output of recommended data mining clustering approach.

3.4.4. Database

Database contains the large datasets of courses and other education related activities. This component contains all the information that the student received from the instructor and also recommends or proposes instructions to the instructor.

3.4.5. Virtual Views

From side to side the consideration of the large datasets from the academic databases, using recommended data mining clustering method; the intelligent analysis records and selection of appropriate contents, virtual views are created and delivered to the students in the form of electronic documents. It can be achieve to identify different patterns which will help students to study, predict and improve their academic performance. The recommended clustering method is able to find groups of items so that the items that are in a cluster are similar to each other than to the items in another cluster. This may help to arrange different items which are under consideration. The clustering data mining approach would help in analyzing different profiles and may implicitly proposes the suitable educational items/materials to each student.

3.5. Implementation: Steps Involved in the Recommended EDM Clustering Approach

The recommended data mining approach (CFSFDP-HD) is implemented using MATLAB to analyze the behavior and to simulate the educational data. The simulated educational data consists of students' (1) obtained marks and (2) class-attendance. The obtained mark consists of (1) three quizzes, (2) two assignments, (3) one midterm, and (4) one final-term exams while the class attendance is calculated on the basis of students' numbers of *presents* and *absents*. Only two percent weightage of class attendance is considered. The presented approach takes *distance matrix D* of dataset as input: *D* is the pairwise distance matrix of educational data.

The key steps of CFSFDP-HD along with the flow control are as follows (see Figure 2):

- Step 1: In the first step, the proposed approach estimates the density ρ_i via heat diffusion using Eq. (5).
- Step 2: the proposed approach calculates the minimum distance δ_i from the higher nearest dense points by using Eq. (3).
- Step 3: the identification of cluster centers is achieved by the use of decision graph. In the decision graph, the ρ_i and δ_i are plotted. The output of this step is the *Cluster Centers*.
- Step 4: The assignation of the remaining points to the identified cluster centers. The output of this step is the *organized clusters* with noise and overlapping clusters.
- Step 5: In this step, the presented approach identifies and fixes the misclassified points and also identifies the noisy or outliers of the organized clusters (noisy and overlapping clusters).

The output of the proposed approach is presented by the *organized clusters*.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 October 2018 doi:10.20944/preprints201808.0350.v2

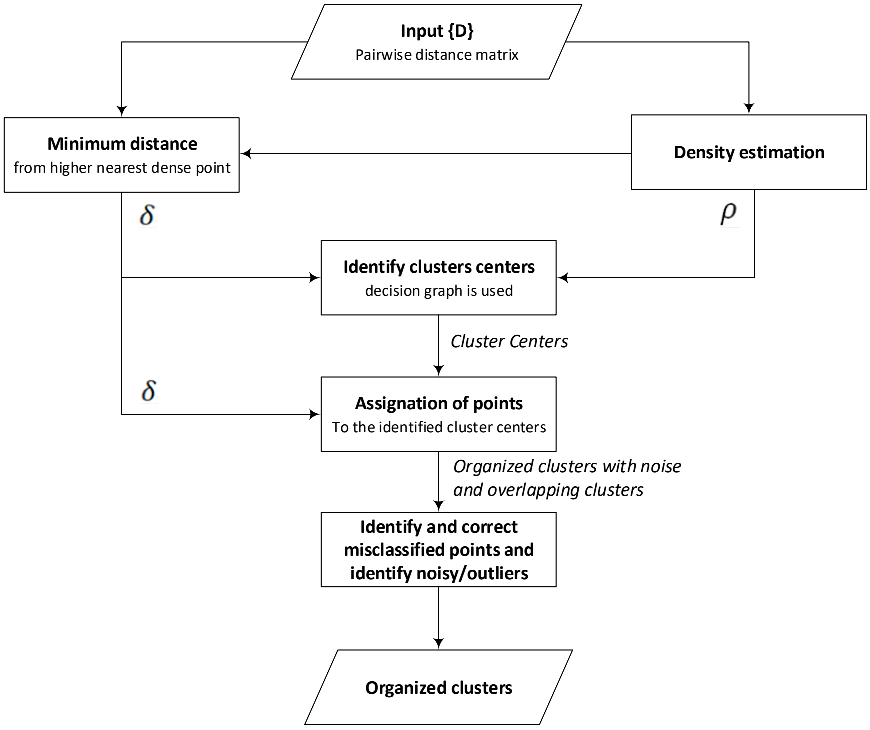


Figure 2: Key steps involved in the presented data mining approach (CFSFDP-HD).

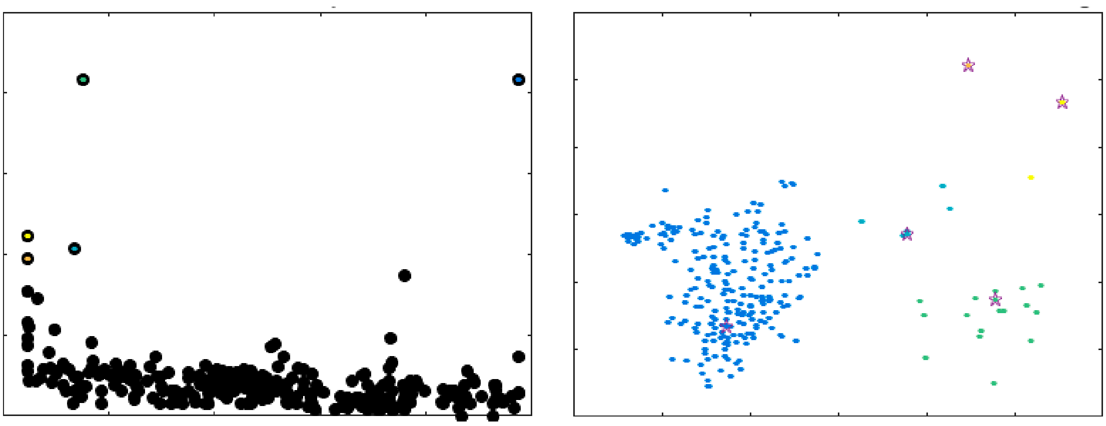
4. Results

The dataset of 600 students (enrolled in different sessions) is simulated by CFSFDP-HD approach to partition students into appropriate groups and is based on the students' obtained marks and class-attendance. The progress-based segmentation of students is necessary to design appropriate teaching methods to address the weakness of a particular group in the class. In the Figure 3(a), the decision graph based heuristic approach is visualized to select the exact number of clusters intuitively. The full black points in Figure 3(a) are treated as non-cluster central-points while the colored points (the outliers) are considered as the central-points of the expected cluster. The decision graph is established after (1) the estimation of density ρ_i via heat-diffusion and (2) the calculation of the minimum distance δ_i from the higher nearest dense-points. The identification of cluster centers is achieved by the use of decision graph shown in the Figure 3(a) where the ρ *and* δ are plotted.

With the minimum interpretation of heuristic approach to select the exact number of clusters, four distinct clusters are identified effectively, as shown in Figure 3(a), where outliers are treated as potential cluster centers and are represented with different colors. After identification of potential cluster centers, the remaining data-points are assigned to the identified cluster centers. The discovered clusters are shown with different colors scheme in Figure 3(b), where 2D Non-classical multidimensional scaling is used to visualize the dataset. The recommended approach "CFSFDP-HD" is adaptive in nature; so there is no need to set any parameter explicitly.

The aforementioned partition of students into four significant groups can play an important role to enhance the learning skills by paying special attention to a particular group of students. The self-motivator and talented students are separated from students with low and below to the average grading students. Based on the obtained different categories of the students, the instructors can adapt different teaching approaches to deal with appropriate group of students. Hence performance of students can be enhanced by applying different methods for each group of students. From the aforementioned case study of GPA, the clustering has potential to partition the education data into appropriate groups and those groups can be used for further analysis to improve the overall education system. This application is simple to understand and exercise in a class at small level effectively.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 October 2018 doi:10.20944/preprints201808.0350.v2



(a) (b)

Figure 3: (a) In the decision graph, the parameters ρ and δ are plotted. The identification of cluster centers is achieved by the use of decision graph. (b) Assigning the remaining data-points to the identified cluster centers are shown with different color schemes; where different colors represent different groups.

The existing clustering methods i.e. K-means, K-medoids, DBSCAN, and AHCT are simulated using the same dataset of 600 students, and by considering the constraints described in the table 1.

Approach		**Parameter settings**
K-Means	1.	No. of clusters **(k = 4)**,
	2.	number of iterations **(n = 50)**
K-Medoids	1.	No. of clusters **(k = 4)**,
	2.	Predefined number of iterations (n = 50).
DBSCAN	3.	Epsilon **(eps = 0.5)**: it defines the radius of neighborhood around the data-point "x".
	4.	Minimum points **(minPts = 10)**: minimum number of neighbors within "eps".
	5.	Does not need to specify the number of clusters and iterations.
AHCT	1.	No. of clusters **(k = 4)**
CFSFDP-HD	1.	Does not need to specify explicitly, the number. of clusters and the number of iterations.

Table 1: Clustering approaches with their parameters' values.

By way of comparing "CFSFDP-HD" with K-means, K-medoids, DBSCAN, and AHCT; the decision graph based approach "CFSFDP-HD" provides a deep insight to select potential clusters intuitively. In general practice, users run K-means and K-medoids more than 1000 times with various input settings (i.e. number of clusters and iterations) to get the meaningful clusters, however, the decision graph based approach used in CFSFDP-HD provides heuristics to get exact solutions within few iterations. While the DBSCAN approach shows some noisy data and also need to define some explicit parameters' values i.e. *Epsilon "eps"* and the *minimum points "minPts"*. The *Epsilon* defines the radius of neighborhood around the data-point and the *minimum data-points* represent the minimum number of neighbors within the radius of the *Epsilon* value. In AHCT approach, there is also need to mention the number of clusters explicitly. Furthermore, four distinct groups shown in the Figure 3(b) can easily be examined, visualized and compared with the K-means, K-medoids, DBSCAN, and AHCT in the Figure 4(a), 4(b), 4(c), 4(d) respectively. The representation of data-points in Figures 3(b) and Figure 4 differs because of their displayed-layout. The students having good grades are showed at left side in Figure 3(b) while the students having good grades are showed at top right of the graphs in Figure 4. In order to get appropriate clusters using the discussed approaches, users must have prior knowledge of existing clusters (number of clusters) and in the undertaken case, unable to detect very low and below average students. This limitation makes inappropriate to discover all intrinsic hidden patterns in data. To tackle technical issues, the CFSFDP-HD method is recommended to discover the existing patterns without knowing technical knowledge of the underlying data.

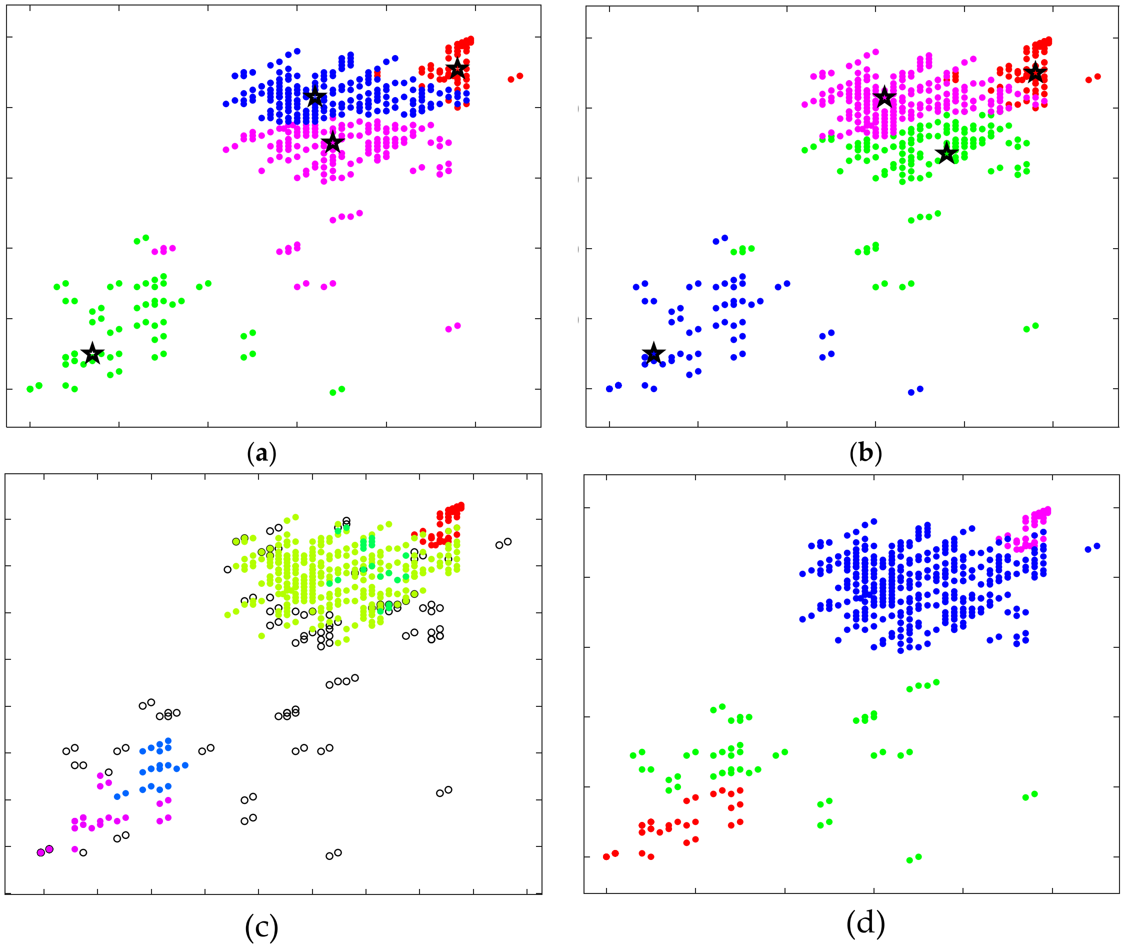


Figure 4: (a) K-means clustering results are visualized; the black-stars are the centroids of the clusters while the colored data-points represent clusters (b) K-medoids clusters are shown; the black-stars are the medoids while different colors show different clusters. Both K-means and K-medoids are of similar nature and almost have similar results. (c) The DBSCAN clusters are represented with different colors while the noisy data-points are represented by the black-circles, (d) different colors are used to represent different clusters identified by the Agglomerative Hierarchical Cluster Tree.

It is observed from the experiments that the CFSFDP-HD is more adaptive in nature and its results are more significant as compared with some of the existing approaches.

5. Conclusions

The data mining approaches provide the sense of intelligence in existing e-learning systems, efficiently and effectively. This paper has presented an adapted data mining clustering approach "CFSFDP-HD" that is integrated with the conceptual personalized e-learning system architecture. It has been observed from the literature that traditional e-learning systems are mostly query-based and the queries are responded without any intelligence or heuristics. The potential application of clustering in educational big data has also been examined. The existing discussed clustering approaches are suitable to cluster educational data where cluster numbers are known and face drawbacks when applied to unknown cluster sizes. It has been evaluated that the recommended data mining clustering approach is efficacious in analyzing the big data to make education systems robust. It also has the potential to solve the challenges of interdisciplinary research, emotional learning, and e-learning in the field of education.

For the future work; the data mining approaches can further be improved by making them more intelligent to generate knowledge and provide more intelligent assistance to the students. The larger and real datasets can be simulated to analyze the behavior of the recommended data mining approach. The learning capabilities of the students can be further improved by introducing the intelligent games integrated with the recommended approach. Student collaboration is an important aspect of learning by group discussion and by sharing personal thoughts. The intelligent techniques can be introduced in different students' groups with significant attributes for problem solving.

Author Contributions: All the authors contributed equally.

Acknowledgments: This work is supported by National Natural Science Foundation of China (No. 91630206).

Conflicts of Interest: The authors declare no conflict of interest.

References

(1) Wong, W.; Fu, A. W. Incremental Document Clustering for Web Page Classification. In *Enabling Society with Information Technology*; Jin, Q., Li, J., Zhang, N., Cheng, J., Yu, C., Noguchi, S., Eds.; Springer Japan, 2002; pp 101–110.

(2) Baker, R. S.; Yacef, K. The State of Educational Data Mining in 2009: A Review and Future Visions. *JEDM| Journal of Educational Data Mining* **2009**, *1* (1), 3–17.

(3) Yan-lin, L. L. Z. The Application of the Internet of Things in Education [J]. *Modern Educational Technology* **2010**, *2* (005).

(4) Baker, R. Data Mining for Education. *International encyclopedia of education* **2010**, *7* (3), 112–118.

(5) Blanco, T.; Casas, R.; Manchado-Pérez, E.; Asensio, Á.; López-Pérez, J. M. From the Islands of Knowledge to a Shared Understanding: Interdisciplinarity and Technology Literacy for Innovation in Smart Electronic Product Design. *International Journal of Technology and Design Education* **2017**, *27* (2), 329–362.

(6) Siemens, G.; Long, P. Penetrating the Fog: Analytics in Learning and Education. *EDUCAUSE review* **2011**, *46* (5), 30.

(7) Howe, D.; Costanzo, M.; Fey, P.; Gojobori, T.; Hannick, L.; Hide, W.; Hill, D. P.; Kania, R.; Schaeffer, M.; St Pierre, S. Big Data: The Future of Biocuration. *Nature* **2008**, *455* (7209), 47.

(8) Kim, G.-H.; Trimi, S.; Chung, J.-H. Big-Data Applications in the Government Sector. *Communications of the ACM* **2014**, *57* (3), 78–85.

(9) Chen, M.; Mao, S.; Zhang, Y.; Leung, V. C. Big Data Applications. In *Big Data*; Springer, 2014; pp 59–79.

(10) Chen, C. P.; Zhang, C.-Y. Data-Intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data. *Information Sciences* **2014**, *275*, 314–347.

(11) Noury, N.; Hervé, T.; Rialle, V.; Virone, G.; Mercier, E.; Morey, G.; Moro, A.; Porcheron, T. Monitoring Behavior in Home Using a Smart Fall Sensor and Position Sensors. In *Microtechnologies in Medicine and Biology, 1st Annual International, Conference On. 2000*; IEEE, 2000; pp 607–610.

(12) Bie, R.; Mehmood, R.; Ruan, S.; Sun, Y.; Dawood, H. Adaptive Fuzzy Clustering by Fast Search and Find of Density Peaks. *Personal and Ubiquitous Computing* **2016**, *20* (5), 785–793.

(13) Qian, G.; Wu, Y.; Ferrari, D.; Qiao, P.; Hollande, F. Semisupervised Clustering by Iterative Partition and Regression with Neuroscience Applications. *Computational intelligence and neuroscience* **2016**, *2016*.

(14) Markowska-Kaczmar, U.; Kwasnicka, H.; Paradowski, M. Intelligent Techniques in Personalization of Learning in E-Learning Systems. In *Computational Intelligence for Technology Enhanced Learning*; Springer, 2010; pp 1–23.

(15) Cordeiro, M.; Gama, J. Online Social Networks Event Detection: A Survey. In *Solving Large Scale Learning Tasks. Challenges and Algorithms*; Springer, 2016; pp 1–41.

(16) Shah, G. H.; Bhensdadia, C. K.; Ganatra, A. P. An Empirical Evaluation of Density-Based Clustering Techniques. *International Journal of Soft Computing and Engineering (IJSCE) ISSN* **2012**, *22312307*, 216–223.

(17) Engström, S. Differences and Similarities between Female Students and Male Students That Succeed within Higher Technical Education: Profiles Emerge through the Use of Cluster Analysis. *International Journal of Technology and Design Education* **2018**, *28* (1), 239–261.

(18) Nerurkar, P.; Shirke, A.; Chandane, M.; Bhirud, S. Empirical Analysis of Data Clustering Algorithms. *Procedia Computer Science* **2018**, *125*, 770–779.

(19) MacQueen, J. Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*; Oakland, CA, USA, 1967; Vol. 1, pp 281–297.

(20) Mehmood, R.; Zhang, G.; Bie, R.; Dawood, H.; Ahmad, H. Clustering by Fast Search and Find of Density Peaks via Heat Diffusion. *Neurocomputing* **2016**, *208*, 210–217.

(21) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Kdd*; 1996; Vol. 96, pp 226–231.

(22) Rodriguez, A.; Laio, A. Clustering by Fast Search and Find of Density Peaks. *Science* **2014**, *344* (6191), 1492–1496.

(23) Zappia, L.; Oshlack, A. Clustering Trees: A Visualisation for Evaluating Clusterings at Multiple Resolutions. *bioRxiv* **2018**.

(24) Anaya, A. R.; Boticario, J. G. A Data Mining Approach to Reveal Representative Collaboration Indicators in Open Collaboration Frameworks. *International Working Group on Educational Data Mining* **2009**.

(25) A. M. de Morais; J. M. F. R. Araújo; E. B. Costa. Monitoring Student Performance Using Data Clustering and Predictive Modelling. In *2014 IEEE Frontiers in Education Conference (FIE) Proceedings*; 2014; pp 1–8.

(26) Prakash, B. R.; Hanumanthappa, M.; Kavitha, V. Big Data in Educational Data Mining and Learning Analytics. *Int. J. Innov. Res. Comput. Commun. Eng* **2014**, *2* (12), 7515–7520.

(27) Algarni, A. Data Mining in Education. *International Journal of Advanced Computer Science and Applications* **2016**, *7* (6), 456–461.

(28) Bovo, A.; Sanchez, S.; Héguy, O.; Duthen, Y. Analysis of Students Clustering Results Based on Moodle Log Data. In *6th International Conference on Educational Data Mining - EDM 2013*; Memphis, Tennessee, United States, 2013; pp 306–307.

(29) Saa, A. A. Educational Data Mining & Students' Performance Prediction. *International Journal of Advanced Computer Science and Applications* **2016**, *7* (5), 212–220.

(30) Agasisti, T.; Bowers, A. J. 9. Data Analytics and Decision Making in Education: Towards the Educational Data Scientist as a Key Actor in Schools and Higher Education Institutions. *Handbook of Contemporary Education Economics* **2017**, 184.

(31) Ivancevic, V.; Celikovic, M.; Lukovic, I. The Individual Stability of Student Spatial Deployment and Its Implications. In *Computers in Education (SIIE), 2012 International Symposium on*; IEEE, 2012; pp 1–4.

(32) Ying, K.; Chang, M.; Chiarella, A. F.; Heh, J.-S. Clustering Students Based on Their Annotations of a Digital Text. In *Technology for Education (T4E), 2012 IEEE Fourth International Conference on*; IEEE, 2012; pp 20–25.

(33) Eranki, K. L.; Moudgalya, K. M. Evaluation of Web Based Behavioral Interventions Using Spoken Tutorials. In *Technology for Education (T4E), 2012 IEEE Fourth International Conference on*; IEEE, 2012; pp 38–45.

(34) Drigas, A. S.; Leliopoulos, P. The Use of Big Data in Education. *International Journal of Computer Science Issues (IJCSI)* **2014**, *11* (5), 58.

(35) Manohar, A.; Gupta, P.; Priyanka, V.; Uddin, M. F. Utilizing Big Data Analytics to Improve Education; ASEE, 2016.

(36) Tulasi, B. Significance of Big Data and Analytics in Higher Education. *International Journal of Computer Applications* **2013**, *68* (14).

(37) Daniel, B. B Ig D Ata and Analytics in Higher Education: Opportunities and Challenges. *British journal of educational technology* **2015**, *46* (5), 904–920.

(38) Dede, C. Next Steps for" Big Data" in Education: Utilizing Data-Intensive Research. *Educational Technology* **2016**, 37–42.

(39) Gaeta, M.; Miranda, S.; Orciuoli, F.; Paolozzi, S.; Poce, A. An Approach To Personalized E-Learning. *Journal of Education, Informatics & Cybernetics* **2013**, *11* (1).

(40) Prawira, A. K.; Sofianti, T. D.; Indrayadi, Y. Developing E-Learning System to Support Teaching and Learning Activities Using DSDM Approach. *PERFORMA: Media Ilmiah Teknik Industri* **2015**, *14* (1).

(41) Yarandi, M.; Jahankhani, H.; Tawil, A.-R. A Personalized Adaptive E-Learning Approach Based on Semantic Web Technology. *webology* **2013**, *10* (2), Art. 110.

(42) Maseleno, A.; Sabani, N.; Huda, M.; Ahmad, R.; Jasmi, K. A.; Basiron, B. Demystifying Learning Analytics in Personalised Learning. *International Journal of Engineering & Technology* **2018**, *7* (3), 1124–1129.

(43) Wu, C.-H.; Chen, Y.-S.; Chen, T.-C. An Adaptive E-Learning System for Enhancing Learning Performance: Based on Dynamic Scaffolding Theory. *Eurasia Journal of Mathematics, Science and Technology Education* **2017**, *14* (3), 903–913.

(44) Liu, Z.; Liu, Y. Research on Personalization E-Learning System Based on Agent Technology. In *Proceedings of the 3rd WSEAS international conference on circuits, systems, signal and telecommunications. Ningbo (China)*; 2009.

(45) Li, X.; Chang, S.-K. A Personalized E-Learning System Based on User Profile Constructed Using Information Fusion. In *DMS*; Citeseer, 2005; Vol. 2005, pp 109–114.

(46) Hedayetul, M.; Shovon, I.; Haque, M. An Approach of Improving Student's Academic Performance by Using K-Means Clustering Algorithm and Decision Tree. **2012**.

(47) Shen, L.; Wang, M.; Shen, R. Affective E-Learning: Using" Emotional" Data to Improve Learning in Pervasive Learning Environment. *Journal of Educational Technology & Society* **2009**, *12* (2).

(48) Dutt, A.; Ismail, M. A.; Herawan, T. A Systematic Review on Educational Data Mining. *IEEE Access* **2017**, *5*, 15991–16005.

(49) Romero, C.; López, M.-I.; Luna, J.-M.; Ventura, S. Predicting Students' Final Performance from Participation in on-Line Discussion Forums. *Computers & Education* **2013**, *68*, 458–472.

(50) Chang, W.-C.; Chen, S.-L.; Li, M.-F.; Chiu, J.-Y. Integrating IRT to Clustering Student's Ability with K-Means. In *Innovative Computing, Information and Control (ICICIC), 2009 Fourth International Conference on*; IEEE, 2009; pp 1045–1048.

(51) Drážďilová, P.; Martinovic, J.; Slaninová, K.; Snášel, V. Analysis of Relations in ELearning. In *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on*; IEEE, 2008; Vol. 3, pp 373–376.

(52) Cobo, G.; García-Solórzano, D.; Santamaria, E.; Morán, J. A.; Melenchón, J.; Monzo, C. Modeling Students' Activity in Online Discussion Forums: A Strategy Based on Time Series and Agglomerative Hierarchical Clustering. In *EDM*; 2011; pp 253–258.

(53) Kaufman, L.; Rousseeuw, P. J. *Clustering by Means of Medoids, in the Statistical Data Analysis Based on the L1-Norm and Related Methods /*; Dodge, Y., Ed.; North-Holland ; Amsterdam ;, 1987.

(54) Rokach, L.; Maimon, O. Clustering Methods. In *Data Mining and Knowledge Discovery Handbook*; Maimon, O., Rokach, L., Eds.; Springer US: Boston, MA, 2005; pp 321–352.

(55) Esposito, F.; Licchelli, O.; Semeraro, G. Extraction of User Profiles in E-Learning Systems. *Proceedings of I-KNOW'0, Graz, Austria* **2003**, 238–243.

(56) Gomes, P.; Antunes, B.; Rodrigues, L.; Santos, A.; Barbeira, J.; Carvalho, R. Using Ontologies for Elearning Personalization. *Communication & Cognition* **2008**, *41* (1), 127.