doi:10.20944/preprints201808.0350.v1

Article

Personalized E-learning System Architecture Using Data Mining Approach

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

[1] School of Computer Engineering and Science, Shanghai University, Shanghai, 200444, China;
[2] School of Computer and IT, Beaconhouse National University, Lahore, 54400, Pakistan;
* Correspondence: iftikhar.hussain@bnu.edu.pk; Tel.: +92-4238100156 (Ext. 508)

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

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]. EDM can be defined as an application of data mining methods in the field of education to exploit novel patterns and artificially analyze big data efficiently and effectively.

Recently, frontier technologies such as 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 form of instructions that are independent of geographical location. Sensors and IoT generate huge 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 approach to partition datasets into distinct groups based on the estimated intrinsic characteristics or similarities [12] and has been applied in various fields [13–19]. Clustering methods can be categorized as: partition-based, density-based, model-based and hierarchy-based [20–24]. The traditional data mining techniques cannot be directly applied to cope with the complexities of big data.

1.1. Research Objectives

This paper presents a personalized e-learning system architecture integrating data mining technique 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. Moreover, the administration can find essential hidden patterns to bring the effective reforms in the existing system. The system is more robust compared to the typical e-learning systems due to the use of clustering methods. The data mining based clustering approaches are offered to partition the students into different groups based on their learning behavior. This paper analyzes K-means algorithm for clustering and compares it with Clustering by Fast Search and Finding of Density Peaks (CFSFDP). It also draws a contrast between K-means and CFSFDP via Heat Diffusion (CFSFDP-HD) in regard to academic performance of students. Both K-means and CFSFDP-HD algorithms were executed multiple times to effectively partition students into groups according to their learning capabilities.

1.2. Paper Organization

This paper organized as follows: Section 2 presents the literature review of data mining techniques with some specific tools to deal with education data. Section 3 describes the idea of personalization in e-learning system architecture using data mining approach. The existing clustering (K-means) approach and the proposed clustering approaches are also described in this section. Section 4 presents the experiments and results with discussion by considering a specific case study. Finally, the conclusion and recommendations for the future research are discussed in Section 5.

2. Literature Review

This section presents a comprehensive review of data mining techniques with some specific tools to deal with educational data.

Big data has the capability to benefit students distinctly by providing them with a modern and dynamic education system. In the study [25], Athanasios S. D. and Panagiotis L. 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 M. et *al.* [26], 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 B. Tulasi [27] and Ben Daniel [28], targeted the higher education and explored the solutions proposed by big data systems to the challenges faced by higher education. Chris Dede [29] 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.

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 [30]. B. R. Prakash, et *al.* [31] 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 this study are coherent with the conclusions of the study presented by Abdul-Mohsen Algarni [32]. In [32] 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 at risk students, distinguishing needs for the adapting needs of various groups of students, expanding graduation rates, adequately surveying institutional execution, boosting grounds assets, and upgrading subject educational modules reestablishment. Another research study [33] consistent with [32] is conducted by Amjad Abu Saa examines and predicts student performance in different scenarios using data mining methods. In the similar study [34], Tommaso Agasisti and Alex J. Bowers have analysed various analytical techniques: Educational Data Mining, Learning Analytics and Academic Analytics, and have reached the conclusion that application of data mining methods with responsibility and professionalism yields positive results.

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 G. et *al.*, [35] 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. A grid agent model was proposed by Zhen L. and Yuying L. in their study [36] 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 [37] have proposed another personalized e-learning system which is a feedback extractor with fusion capability to adjust the user preferences. Maryam Yarandi, et *al.* [38] take individual learning capabilities of students to present an ontology-based approach to develop an adaptive e-learning system. The proposed e-learning system creates content according to the learner's knowledge. The significance of the above mentioned literature being that personalized e-learning systems are effective tools in individual learning and hence this paper proposes yet a fresh intelligent personalized e-learning system. The K-means [20] is a state-of-the-art partition based clustering algorithm and have been applied in EDM [39–50]. 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 [45]. Another study presented by Ying, et al. [49] 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 [51], K-means was applied to examine the influence of human characteristics on student's performance while listening to music. Chang, et al [50] utilized Item Response Theory (IRT) to identify student's ability and discovered distinct groups based on the student's ability.

Web based education or e-learning is a new paradigm in education where a significant large amount of information defining the variety of teaching-learning interactions. It is endlessly generated and ubiquitously available. To cope with aforementioned e-learning issues, we proposed a new e-learning system architecture using the data mining techniques. The integration of data mining techniques (DMT) makes the learning system more interesting.

3. Personalized E-learning System Architecture Using Data Mining Techniques

In this section, a Personalized E-learning System Architecture (PESA) is presented. Proposed system is sensitive to detect the understanding levels of students and then respond to the students

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2018 doi:10.20944/preprints201808.0350.v1

according to their learning capabilities. Proposed system finds the possible groups in students by matching shared similarities according to their level of interest. For each group, system generates different quizzes, assignments, study related games, and books' contents to improve their learning capabilities. To make groups and select appropriate teaching methods, system uses artificial intelligence and adaptive clustering methods. In proposed architecture, the K-means 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. Problem Background and the Big Data

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 the changes and challenges posed by the big data. In such 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 [59]. 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 students those take part in class activities and show satisfactory outputs. Moreover, there are some intrinsic hidden patterns that exist among the students. Students can be divided into different categories or groups based on their progress. The same teaching method may not be effective for different groups of students.

The similarity measures and clustering are important tasks to find the similar groups in 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.

Institutional databases, having the teaching material and users' queries, are entertained according to the stored data. However, most of updated knowledge lies on internet at different places. To robust the student learning capabilities, it might be credible to integrate the rest of data sources with e-learning system [52]. The data mining techniques can play an important role to find the relationships among different subjects available over internet, specifically in the e-learning systems. Generally most of the e-learning systems are static and query based. In this domain, students' click based server logs generated valuable data. Clustering methods can be successfully utilized to analyze the click stream data. Clustering of click streams data can be further utilized to make e-learning system more attractive and intelligent to understand the students' activities and interest.

3.2. Proposed PESA

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.

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

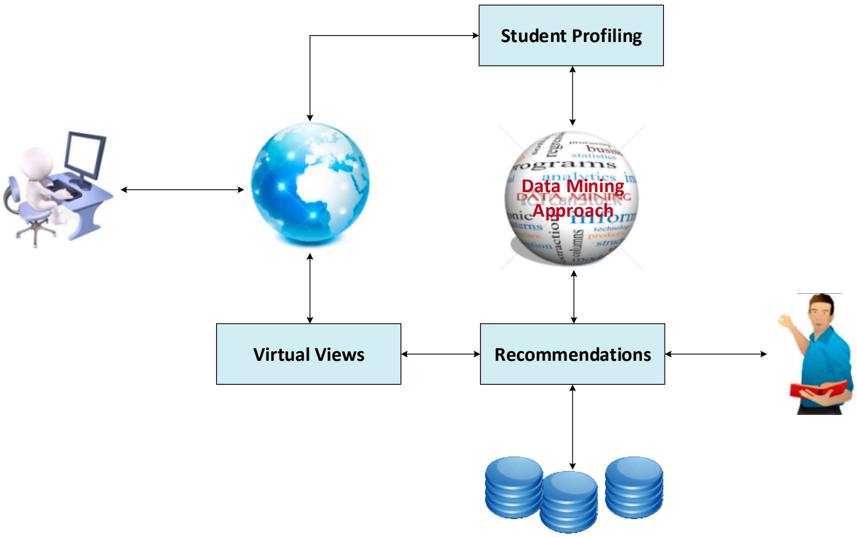

Figure 1: Personalized E-learning Architecture. A profile is created for each learner and is automatically updated based upon the activities of the learner.

3.2.1. 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 [53,54], 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 [54] includes personal, personality, cognitive, pedagogical and preference data. Personal data define the biographical 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.

3.2.2. Data Mining Techniques

The data mining 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 techniques, 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 K-means and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). Unlike existing e-learning systems, we proposed to use CFSFDP and CFSFDP-HD methods to achieve robust results. The data mining techniques (CFSFDP and CFSFDP-HD) are explained in more details as follows.

CFSFDP and CFSFDP-HD

The CFSFDP has been recently proposed by Alex and Laio [23]. It has characteristics to discover significant clusters in a more intuitive way as compared with K-means. A brand new heuristic approach is proposed that empowers clustering procedure, in which high-density regions are identified as potential clusters, outliers are automatically identified and arbitrary shape of clusters are organized. In K-means to obtain meaningful clusters users are required to repeat clustering

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2018 doi:10.20944/preprints201808.0350.v1

process multiple times with different parametric setting. However, the unique approach utilized in CFSFDP to discover clusters and noise adaptively would be a significant clustering tool to analyze the educational. The CFSFDP uses the following given methodology to discover significant clusters.

For each given data-point i, CFSFDP calculates its local density (ρ_i) and a minimum distance (δ_i) with its nearest high density point. The local density can be estimated by utilizing the following definition:

Definition-1:

$$\rho_i = \sum_j X(d_{ij} - d_c) \tag{1}$$

where,

$$X(x) = \begin{cases} 1 & x < 0 \\ 0 & otherwise. \end{cases} \tag{2}$$

However, the distance (δ_i) can be computed using the definition-2, given as follows:

Definition-2:

$$\delta_i = \begin{cases} \min\limits_{j:\rho_j > \rho_i} (d_{ij}) & if\ \exists\ j\ s.t{:}\ \rho_j > \rho_i \\ \max\limits_{j:\rho_j > \rho_i} (d_{ij}) & otherwise \end{cases} \tag{3}$$

Cluster centers are attained by plotting calculated values of ρ_i and δ_i, which is referred as the decision graph. In cluster analysis, the key challenge is to discover correct cluster centers in the datasets [1]. However, CFSFDP uses decision graph to select the correct cluster centers with the least human interaction, which makes it more worthy to analyze big data / streaming data. CFSFDP has variety of applications in education as well as in many other fields, such as bioinformatics [58], image processing and protein analysis [23].

As 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 and further it can be used to cluster the similar documents, find plagiarism in documents, and analyse the students' profiles and to find the similar insights in different research areas. The CFSFDP via heat diffusion (CFSFDP-HD) [21] 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 as follows:

$$\hat{f}(x;t) = \frac{1}{n} \sum_{j=1}^{n} \sum_{k=-\infty}^{\infty} e^{-k^2\pi^2 t/2} \cos(k\pi x)\ \cos(k\pi x_j) \tag{4}$$

Equation 5 can be expressed as

$$\hat{f}(x;t) \approx \sum_{j=0}^{n-1} a_k e^{-k^2\pi^2 t/2} \cos(k\pi x), \tag{5}$$

where n is a positive large interger and a_k is

$$a_k = \begin{cases} 1 & k = 0 \\ \frac{1}{n} \sum_{i=1}^{n} \cos(k\pi x_i), & k = 1,2,\dots,n-1, \end{cases} \tag{6}$$

3.2.3. Recommendations

This process is responsible to collect data from databases filtered according to student profile with the help of data mining techniques. It also has the ability to prevent duplication of the information created before. This process recommends or proposes the solution to the instructor.

3.2.4. Database

Database contains the rich data 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.

3.2.5. Virtual Views

After the intelligent analysis of student records and selection of appropriate contents for students, virtual views are created and delivered to the students in the form of electronic documents.

3.3. Existing Clustering Method (K-means)

The K-means [20] is a state-of-the-art partition based clustering algorithm. 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:

$$\underset{S}{argmin} \sum_{i=1}^{n} \sum_{x \in S_i} ||x - \mu_i \,||^2 \tag{7}$$

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

K-means is the 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. As, in EDM data, the selection of number of clusters and initial centroids setting of K-means are hard to setup. These are also obscure to find significant signal of interest. Therefore, more sophisticated and frontier clustering methods are required to benchmark on EDM data to get intrinsic insights. Moreover, various other clustering methods have been used in EDM such as DBSCAN in [22,55] and Hierarchical clustering in [42,50,56,57], however, these approaches are also not robust to identify significant clusters in ambiguous and noisy datasets [23].

3.4. Steps Involved in the Proposed Framework

The key steps of CFSFDP-HD along with the flow control are shown in Figure 2.

The presented approach takes *distance matrix D* of dataset as input: D is the pairwise distance matrix of educational data.

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 the *organized clusters*.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2018 doi:10.20944/preprints201808.0350.v1

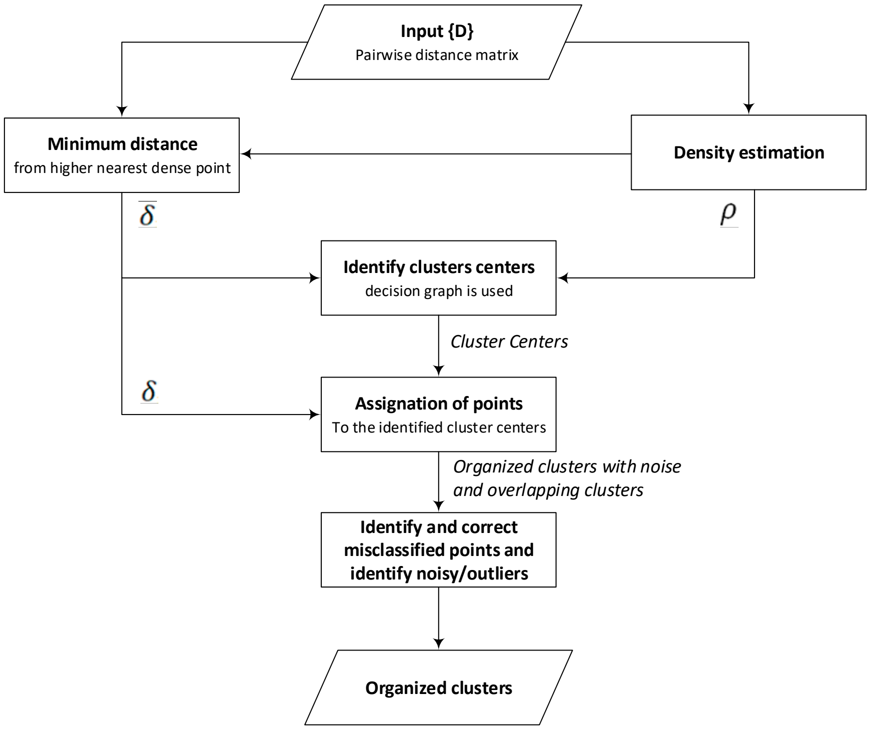

Figure 2: Key steps involved in the presented data mining approach (CFSFDP-HD).

4. Experiments and Results

The presented data mining approach (CFSFDP-HD) is implemented using MATLAB to analyse the behavior and to simulate the educational data.

4.1. Experiment 1: Using K-means clustering approach

In the first experiment, the data of 57 students is simulated using *K-means clustering approach* and executed for 1000 times. The analysis is based on the students' obtained marks of: (1) three quizzes, (2) two assignments, (3) one midterm, and (4) one final-term exams. The class-attendance and class-participation are also considered. The results are extracted by passing different values of clustering inputs. The output showed that three distinct groups of students are obtained. The aforementioned partition of students into three significant groups can play an important role to enhance the learning skills by paying special attention to a particular group of 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. According to table 1, the students in group C require special care and attention to improved their skills, group B students require only a little attention, especially in class tests and quizes, and the students of group A are self-motivated and do not require special attention by instructors or counselors, as described in table-1.

Table 1: K-means based created three different student categories of the synthetic data of 57 students. Each category needs to teach with different levels of preparation.

No. of students	Group	Efforts
18	A	Extra-ordinary students are comprised in this category and do not need special care to enhance their performance.
18	B	The students of this category are mediocre; they need to take care of their attendance and the sessional tests (i.e. class tests & Assignments).
21	C	The students if this category of below average and they needs special care and also required a lot of practice to deal with their course material.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2018 doi:10.20944/preprints201808.0350.v1

The aforementioned application is simple to understand and exercise in a class at small level. In order to get appropriate clusters using K-means, users must have prior knowledge of existing clusters. This limitation makes K-means inappropriate to discover all intrinsic hidden patterns in data. To tackle with this technical drawback of K-means we are presenting density peaks based clustering methods to discover all existing patterns in data without knowing technical knowledge of underlying data.

4.2. Experiment 2: Using CFSFDP-HD clustering approach

In this experiment, the dataset of 600 students (enrolled in different sessions) is simulated by CFSFDP-HD approach. The CFSFDP-HD is used to partition the students into appropriate groups and is based on the students' obtained marks of: (1) three quizzes, (2) two assignments, (3) one midterm, and (4) one final-term exams. The class-attendance and class-participation are also considered. 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, the decision graph based heuristic approach is visualized to select the exact number of clusters intuitively. The full black points in Figure 3 are treated as non-cluster centre points.

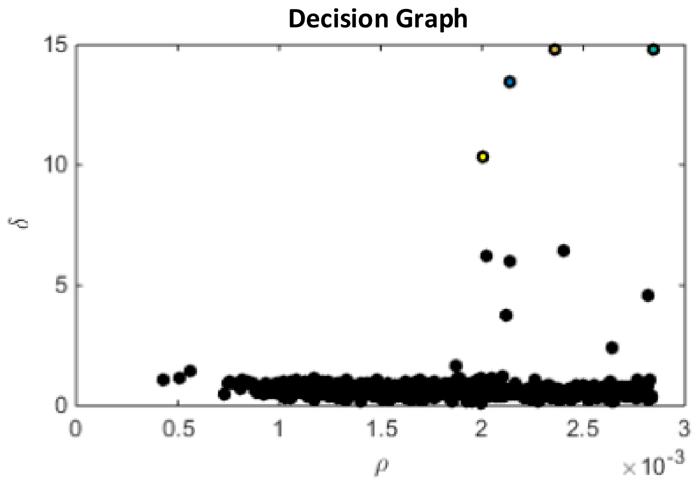

Figure 3: In the decision graph, the ρ and δ are plotted. The identification of cluster centers is achieved by the use of decision graph.

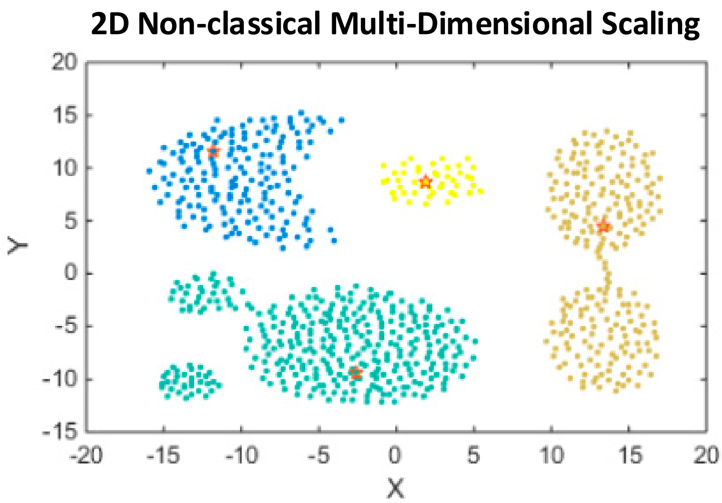

Figure 4: CFSFDP-HD analysis of 600 students' performance in Computer Application subject. Assigning the remaining points to the identified cluster centres are shown in different colour schemes, where different colours represent different groups.

With the minimum interpretation of heuristic approach to select the exact number of clusters, we successfully identified four distinct groups: Excellent (A+), Good (A), Average (B) and poor (C) in the students, as shown in Figure 4, where outliers are treated as potential cluster centres and are represented with different colours. After identification of potential cluster centers, the discovered clusters are shown with different colours scheme in Figure 4, where 2D Non-classical multidimensional scaling is used to visualize the dataset.

As compared with K-means, the decision graph based approach provides a deep insight to select potential clusters intuitively. In general practice, users run K-means more than 1000 times with various input settings to get the meaningful clusters, however, the decision graph based approach in CFSFDP-HD provides heuristics to get exact solutions within few repetitions of CFSFDP-HD. Furthermore, four distinct groups can easily be examined and visualized in Figure 4 using the heat-map.

Table 2: CFSFDP-HD based created four different student categories of the dataset of 600 students belong to different sessions. Each category needs to teach with different levels of preparation.

No. of students	Group	Efforts
90	A+	A good student who needs no extra effort.
398	A	An average student who needs to put some effort in course work.
20	B	A below average student who needs to take put in extra effort in lessons and course work.
92	C	A lowest level student who needs to put in the most effort in lessons and course work to keep up.

From the aforementioned case study of GPA, the clustering has potential to partition the education data into appropriate groups and that groups can be used for further analysis to improve the overall education system. From literature [39–50], it has been observed that K-means has been used in EDM for different purposes, however, CFSFDP and CFSFDP-HD are more adaptive in nature and their results are more significant as compared with K-means [21,23]. Therefore, more robust results can be achieved with replacement of K-means with CFSFDP and CFSFDP-HD.

5. Conclusions

As the data mining approaches provide the sense of intelligence in existing e-learning systems, efficiently and effectively. This paper has been presented the personalized e-learning architecture using the data mining techniques. The potential application of clustering in educational big data has also been examined. 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. Similarly, the K-means is suitable to cluster educational data where cluster numbers are known and faces drawbacks when applied to unknown cluster sizes. Hence, more robust data mining approaches (CFSFDP and CFSFDP-HD) are incorporated in the proposed e-learning system to find clusters in the educational data. Furthermore, it has been evaluated that data mining techniques are efficacious in analyzing the big data to make education systems robust and have 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 proposed data mining approaches. The learning capabilities of the students can further be improved by introducing the intelligent games. 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) Stevenson, J. Developing Technological Knowledge. *International Journal of Technology and Design Education* **2004**, *14* (1), 5–19.

(19) Zhang, Y.; Zhao, Y. Automated Clustering Algorithms for Classification of Astronomical Objects. *Astronomy & Astrophysics* **2004**, *422* (3), 1113–1121.

(20) 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.
(21) 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.
(22) 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.
(23) Rodriguez, A.; Laio, A. Clustering by Fast Search and Find of Density Peaks. *Science* **2014**, *344* (6191), 1492–1496.
(24) Xu, R.; Wunsch, D. Survey of Clustering Algorithms. *IEEE Transactions on neural networks* **2005**, *16* (3), 645–678.
(25) Drigas, A. S.; Leliopoulos, P. The Use of Big Data in Education. *International Journal of Computer Science Issues (IJCSI)* **2014**, *11* (5), 58.
(26) Manohar, A.; Gupta, P.; Priyanka, V.; Uddin, M. F. Utilizing Big Data Analytics to Improve Education; ASEE, 2016.
(27) Tulasi, B. Significance of Big Data and Analytics in Higher Education. *International Journal of Computer Applications* **2013**, *68* (14).
(28) Daniel, B. B Ig D Ata and Analytics in Higher Education: Opportunities and Challenges. *British journal of educational technology* **2015**, *46* (5), 904–920.
(29) Dede, C. Next Steps for" Big Data" in Education: Utilizing Data-Intensive Research. *Educational Technology* **2016**, 37–42.
(30) 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**.
(31) 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.
(32) Algarni, A. Data Mining in Education. *International Journal of Advanced Computer Science and Applications* **2016**, *7* (6), 456–461.
(33) Saa, A. A. Educational Data Mining & Students' Performance Prediction. *International Journal of Advanced Computer Science and Applications* **2016**, *7* (5), 212–220.
(34) 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.
(35) Gaeta, M.; Miranda, S.; Orciuoli, F.; Paolozzi, S.; Poce, A. An Approach To Personalized E-Learning. *Journal of Education, Informatics & Cybernetics* **2013**, *11* (1).
(36) 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.
(37) 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.
(38) Yarandi, M.; Jahankhani, H.; Tawil, A.-R. A Personalized Adaptive E-Learning Approach Based on Semantic Web Technology. *webology* **2013**, *10* (2), Art. 110.
(39) Zheng, Q.; Ding, J.; Du, J.; Tian, F. Assessing Method for E-Learner Clustering. In *Computer Supported Cooperative Work in Design, 2007. CSCWD 2007. 11th International Conference on*; IEEE, 2007; pp 979–983.

(40) Tian, F.; Wang, S.; Zheng, C.; Zheng, Q. Research on E-Learner Personality Grouping Based on Fuzzy Clustering Analysis. In *Computer Supported Cooperative Work in Design, 2008. CSCWD 2008. 12th International Conference on*; IEEE, 2008; pp 1035–1040.
(41) Antonenko, P. D.; Toy, S.; Niederhauser, D. S. Using Cluster Analysis for Data Mining in Educational Technology Research. *Educational Technology Research and Development* **2012**, *60* (3), 383–398.
(42) 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.
(43) Chang, W.-C.; Wang, T.-H.; Li, M.-F. Learning Ability Clustering in Collaborative Learning. *JSW* **2010**, *5* (12), 1363–1370.
(44) Almeda, M. V.; Scupelli, P.; Baker, R. S.; Weber, M.; Fisher, A. Clustering of Design Decisions in Classroom Visual Displays. In *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*; ACM, 2014; pp 44–48.
(45) 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.
(46) Chen, C.-M.; Li, C.-Y.; Chan, T.-Y.; Jong, B.-S.; Lin, T.-W. Diagnosis of Students' Online Learning Portfolios. In *Frontiers In Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE'07. 37th Annual*; IEEE, 2007; pp T3D-17-T3D-22.
(47) Tair, M. M. A.; El-Halees, A. M. Mining Educational Data to Improve Students' Performance: A Case Study. *International Journal of Information* **2012**, *2* (2), 140–146.
(48) Perera, D.; Kay, J.; Koprinska, I.; Yacef, K.; Zaïane, O. R. Clustering and Sequential Pattern Mining of Online Collaborative Learning Data. *IEEE Transactions on Knowledge and Data Engineering* **2009**, *21* (6), 759–772.
(49) 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.
(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) 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.
(52) 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).
(53) Esposito, F.; Licchelli, O.; Semeraro, G. Extraction of User Profiles in E-Learning Systems. *Proceedings of I-KNOW'0, Graz, Austria* **2003**, 238–243.
(54) Gomes, P.; Antunes, B.; Rodrigues, L.; Santos, A.; Barbeira, J.; Carvalho, R. Using Ontologies for Elearning Personalization. *Communication & Cognition* **2008**, *41* (1), 127.
(55) Dutt, A.; Ismail, M. A.; Herawan, T. A Systematic Review on Educational Data Mining. *IEEE Access* **2017**, *5*, 15991–16005.
(56) Dráždilová, 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.

(57) 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.
(58) Wiwie, C.; Baumbach, J.; Röttger, R. Comparing the Performance of Biomedical Clustering Methods. *Nature methods* **2015**, *12* (11), 1033.
(59) Hedayetul, M.; Shovon, I.; Haque, M. An Approach of Improving Student's Academic Performance by Using K-Means Clustering Algorithm and Decision Tree. **2012**.