EFFECTIVE STUDENT COLLBORATIVE SYSTEM BASED ON CLUSTERING ANALYSIS

Kyaw Thiha[1], Khin Myo Sett[2]

Abstract

In most recent years, many educational research workers pointed out that collaborative learning should be widely used in education systems because collaboration skill has increasingly become an essential skill of modern society. Moreover, collaborative learning can encourage the improvement of students' comprehension about their studying. However, it is vitally important to notice that students have different common interests which mean various habits, likes and dislikes, and different learning styles. From the view of educational informatization, we therefore propose Student Collaborative System (SCS) to help in grouping students in accordance with their common interests in this paper. In our proposed system (SCS), students are grouped based on the analysis of common interests using K-Means clustering method. This system is developed using Python as programming language and Spyder (Python 3.7) as integrated development environment (IDE).

Keywords: collaborative learning, common interests, educational informatization, K-Means clustering

Introduction

Alongside with the influence of educational informatization, Data Mining techniques are widely used in education industry to gain the valuable insight from data. This is namely called Educational Data Mining (C. Romero, S. Ventura (2007)). Following the steps of most Data Mining system, Educational Data Mining also has three steps; data preprocessing, mining or using one of Data Mining techniques and providing finalized results.

Although the use of Educational Data Mining is increasingly spread, it is mainly convenient in developed countries because data collecting is the indispensible process of Data Mining and it is easily performed in developed countries through many information system and successful E-Learning classrooms. Educational Data Mining is rarely used in developing countries as traditional classrooms are essential roles of education system and data collection is not that much easy.

On the one hand, collaborative learning which can provide collaborative skill in real life and has a lot of benefits for students has become one of popular learning styles. Moreover, it can help student-center learning which can promote student engagements (Singhal, Divya. (2017)). On top of that, student-center learning is the education system which is used in our country, Republic of the Union of Myanmar.

Due to the above reasons, student collaborative system (SCS) is designed to meet the requirements in developing countries, especially in Myanmar. Regardless of being homogeneous of heterogeneous grouping, our paper mainly focus on grouping students according to their common interests. In our research work, questionnaire approach is used as primary data collection method and K-Means clustering method is applied to classify students into proper groups.

Related Works

A lot of educational data mining researches are developed to enhance collaborative learning although there many open source tools such as Keel, Weka and we can adopt many seldom paywares, e.g. SPSS and DBMiner. Among these previous researches, a research

[1] Part-Time Demonstrator, Department of Computer Studies, University of Mandalay
[2] Professor(Head), Department of Computer Studies, University of Mandalay

paper by TangJie et.al., (2012) explored a peer-model that was designed for Mobile Computer Supported Learning in order to find partners for students in 2012. Moreover, there are many researches about educational data mining tools such as Pdinamet (E. Gaudioso et al., (2009)) and E-learning Web Miner (Zorrilla, Marta & García-Saiz, Diego. (2013)) which apply clustering or association rules.

In 2009, Wen-Yan Kao developed a learning style classification mechanism for e-learning. A hybrid approach which is composed of genetic algorithm and k-NN classification is used in this research. And In 2014, Enhanced Ant Colony Optimization (EACO) algorithm is proposed by Hu Hui based on the ability, interests and comprehensions of students in order to enhance cooperative learning. On top of that, Ma Yanyun measured students' learning ability to improve the effectiveness of collaborative learning in 2016.

Through these previous researches, we can clearly see that many educational researchers applied Educational Data Mining in terms of improving effectiveness and efficiency of student collaboration.

Proposed System

Our proposed system, student collaborative system (CSC) is inspired by these related works. However, the main difference between our work and these related researches are data sources. The data sources of these related studies are mostly collected from E-Learning systems and computerized school management systems which can acquire data of student learning activity. Apart from these researches, we developed a questionnaire to collect data because E-Learning system and computerized school management systems cannot be adopted widely in developing countries and so teachers in those portions of the world need a questionnaire as a mean to gather students' data. After we gathered students' data through the developed questionnaire, K-means clustering method was employed to analyze collected data. The system flow diagram of our proposed system is illustrated in figure 1.

Data Collecting

Following the rules of Data Mining processes, data collecting came in first place of our research work. To collect data, we developed a questionnaire that can catch students' opinions which can be used to identify the groups. Moreover, this questionnaire was comprised in accordance with students in Myanmar currently enrolling in computer science because it is vitally important to guarantee comprehensiveness, usability, quality and to reflect the targeted people and to meet requirements in research areas. Therefore, the questionnaire was developed with 20 questions in which number one to ten are about lifestyle and habits while number eleven to fifteen are about their opinions on their specialization and last five question are about their learning styles that are extracted from Dunn's theory of learning style (Dunn, R. (1984)). Each question consists of three answers to be chosen as shown in Table 1.

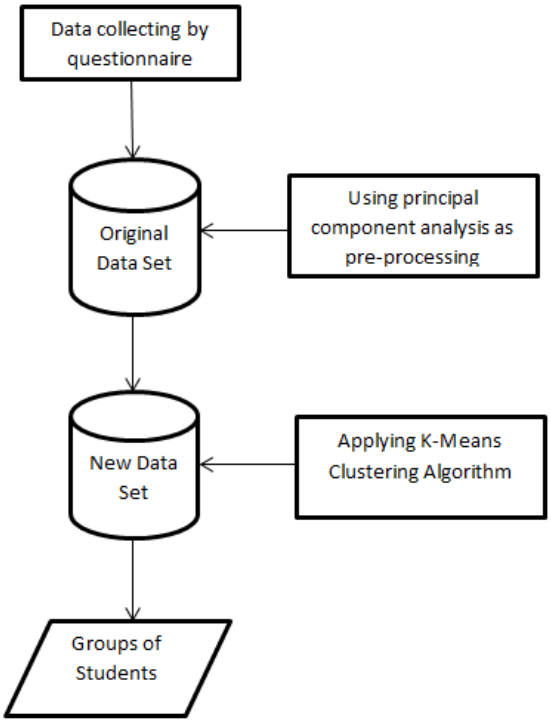

Figure 1 System Flow Diagram

Table 1 Questionnaire to collect students' opinions

Questions	Choices	Answers
1. Which is the most and usually habit you would like to do in the following?	(a) Gaming (b) Music/Watching TV (c) Photography	
2. What time of weekend day would you be most available?	(a) Morning (b) Afternoons (c) Evening	
: :	: :	: :
11. Which programming language you would prefer the most?	(a) Java (b) C# (c) C++	
: :	: :	: :
20. What will you do when you need to make a decision while shopping?	(a) Make the decision right away (b) Make the decision after deep consideration (c) Hesitate to make decision	

These questionnaires are handed out to 30 computer science students from Department of Computer Studies, University of Mandalay. And 27 questionnaires were collected with a collecting rate of nearly 90% (exactly 89.99%).

Data Pre-processing

After collecting data sources, next step we must perform is data pre-processing on original data set, which is an indispensable and important initial stage of data mining. Moreover, we can solute the noises, missing values and inconsistent data via data pre- processing. In other words, we can say that the main objective of data pre-processing is to minimize the dimensionality or size of data, normalize the original data set, discover the interconnection between data, detect outliers and extract features for data. Data pre- processing consists of data cleaning, data integration, data transformation. These stages are illustrated by figure 2 and figure 3 (Tamilselvi et al.,2015). In our research work, we performed data conversion and principal component analysis for dimensionality reduction on original data set.

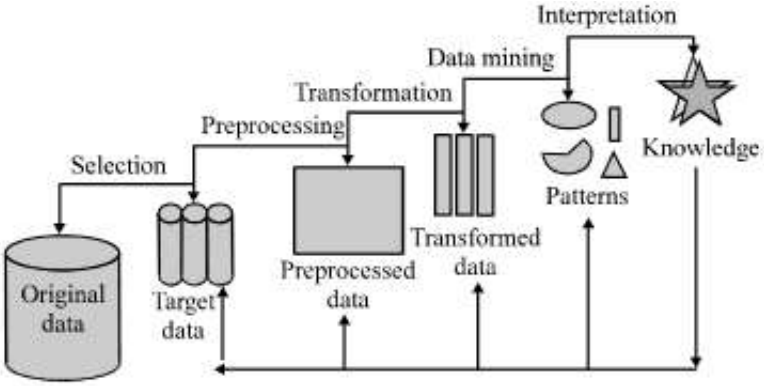

Figure 2 Data pre-processing in Knowledge discovery

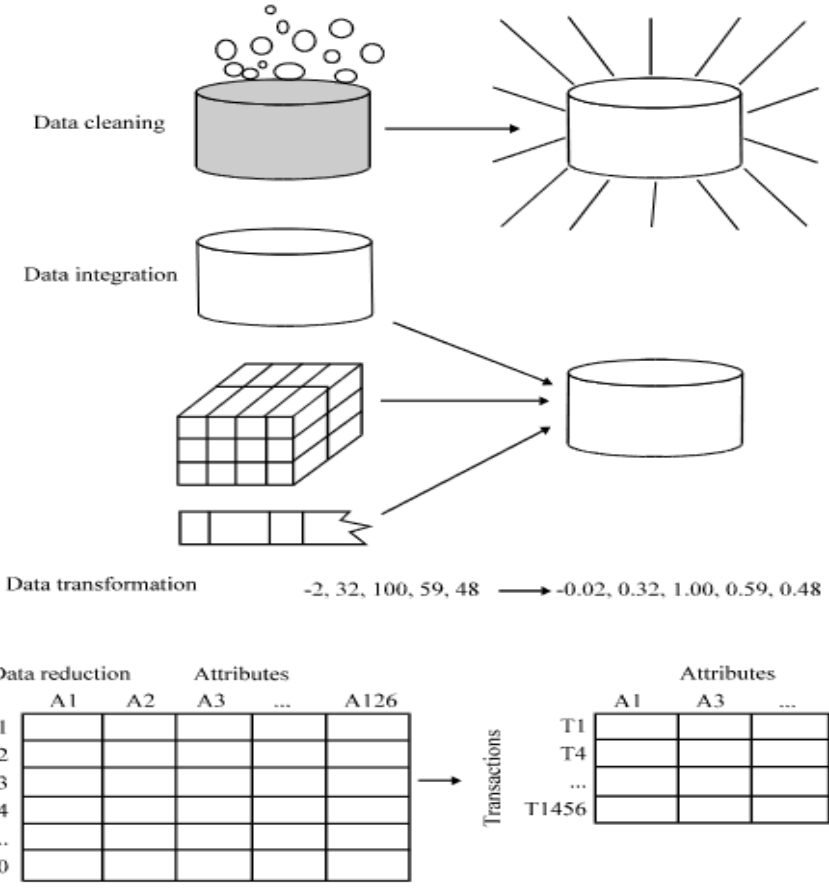

Figure 3 Data pre-processing Form Data Conversion

According to the structure of questionnaire we designed as shown in Table 1, every students staring from the sequences of their roll number R possess 20 attributes from A1 to A20. A range of choice for each answer A is{a, b, c}.Therefore, we can say that R_i={A_1,A_2,A_3,……,A_{19},A}, i= number of students {1,2,3,…..,40} as shown in Table 2. Due to the fact that our proposed system (SCS) need numeric values to conduct clustering analysis, a, b, and c is transformed to the values of 1, 2, and 3 respectively as shown in Table 3.

Table 2 Original data

Roll Num	Q1	Q2	……..	Q10	…….	Q19	Q20
R1	B	b	…….	B	……..	C	A
R2	C	c	……..	C	……..	C	a
:	:	:	……..	:	……	:	:
R10	B	a	…….	B	…….	C	A
:	:	:	……..	:	……	:	:
R26	B	c	……..	B	…….	B	a
R27	B	a	……	C	……	C	B

Table 3 Converted data

Roll Num	Q1	Q2	……..	Q10	…….	Q19	Q20
R1	2	2	…….	2	……..	3	1
R2	3	3	……..	3	……..	3	1
:	:	:	……..	:	……	:	:
R10	2	1	…….	2	…….	3	1
:	:	:	……..	:	……	:	:
R26	2	3	……..	2	…….	2	1
R27	2	1	……	3	……	3	2

Dimensionality Reduction of Original Data by Principal Component Analysis

Principal component analysis can be defined as a process to reduce the size of data that has multi dimension; in other words, data which has a lot of attributes with the minimum loss of data. The goals of principal component analysis are (1) to extract the most important information from the data table;(2) to compress the size of the data set by keeping only this important information;(3) to simplify the description of the data set; and (4) to analyze the structure of the observations and the variables;(5) to compress the data, by reducing the number of dimensions, without much loss of information (Mishra et al., 2017). In our research work, principal component analysis was applied as pre stage to K-Mean Clustering algorithm which need the least data size as much as possible.

Although principal component analysis has several stages which are also related to mathematical backgrounds, we applied the Python programming libraries built in Python 3.7 supported by Spyder IDE to perform the principal component analysis on original data. As a result, we can reduce the data size of original data from 20 attributes to 15 attributes. As shown in figure 4, there is no variation among data when the variation line reaches 15 principal components. This means that the original data which has 20 components can be covered by first 15 components. This leads to R_1 ={ A_1,A_2,A_3,…..,A_{14},A_{15}}.

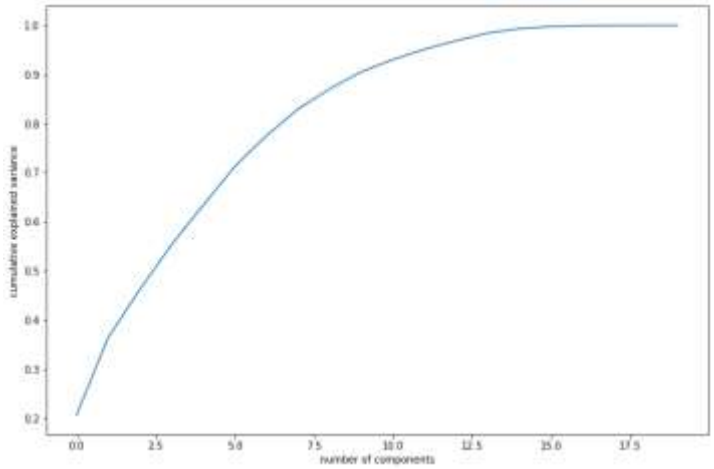

Figure 4 Variance of data

After deciding number of principle component as 15, we derived a new data set as shown in figure 5.

K-Means Clustering Algorithm

Clustering Analysis can be defined as partitioning the data into different groups according to their attributes or characteristics. We can also say that data in one group are similar and have same properties compared to data in different groups (Han et al., 2004). Unlike supervised learning, clustering is an unsupervised learning method in which we analyze the data to be put into similar groups and to create different groups instead of predicting a target. There are many types of clustering; connectivity-based clustering, also known as hierarchical clustering, centroid-based clustering, also known as k-means clustering, distribution-based clustering, density-based clustering, gird-based clustering.

With the rapid development of data mining technology, there are a lot of areas in which clustering method is highly applied. It is widely used not only in business and market research but also in image segmentations. Moreover, clustering analysis can identify Web files by looking at their attributes such as their subjects. In education sector, clustering analysis is applied by many research workers as we discussed in related works section.

PC1	PC2	...	PC13	PC14	PC15
0.365854	9.985776e-02	...	5.367860e-01	4.289935e-02	-2.091347e-01
0.228457	-1.976134e-01	...	-1.864872e-01	-3.129605e-01	-1.545044e-01
0.240936	2.272357e-01	...	2.562431e-02	-3.072779e-01	3.729338e-01
-0.223152	-1.778080e-01	...	5.115909e-01	-3.347983e-01	-1.468264e-01
-0.306658	2.655403e-01	...	-1.414196e-01	-5.341031e-01	-3.054365e-01
-0.279119	2.836098e-01	...	1.588858e-01	1.548743e-01	4.285242e-01
-0.269030	1.417237e-01	...	-1.710357e-01	-2.862694e-01	1.748827e-01
0.015183	1.883671e-01	...	-3.601613e-02	9.955013e-03	-4.493631e-01
0.308539	-1.669531e-01	...	-7.023506e-02	-3.336279e-01	5.829103e-03
-0.209130	-2.105387e-01	...	7.172767e-02	-1.656067e-01	8.181922e-02
-0.129787	2.876661e-01	...	2.602901e-01	1.585640e-01	-3.538935e-01
0.000000	-3.388132e-21	...	2.775558e-17	-2.090342e-16	5.551115e-17
-0.092111	3.357913e-01	...	-5.945705e-02	1.304645e-02	3.872300e-02
-0.205322	-3.288269e-01	...	-1.753506e-01	1.690723e-01	-2.050590e-01
0.172279	4.370517e-02	...	1.411913e-01	-3.058368e-01	1.518661e-01
0.129435	-1.249319e-01	...	4.477161e-02	-2.326388e-02	4.124667e-02
-0.246579	2.143753e-01	...	3.709259e-02	-6.813678e-02	-2.923280e-02
0.308030	3.826778e-01	...	6.075898e-02	1.144206e-02	-1.328249e-02
0.185212	2.497981e-01	...	-4.039612e-01	1.898060e-02	-2.375823e-01
0.136815	1.142839e-01	...	-1.913203e-01	1.027712e-01	6.752292e-02

Figure 5 New Data Set Via Principal Component Analysis

In our proposed system, K-Means clustering algorithm was adopted because it is much better than other clustering algorithm in term of computation. Another reason is that K- means clustering is very convenient if the number of student is very large and is easy to implement. Last but not least, K-mean clustering is less time consuming compared to other algorithm. John A. Hartigan (1975) proposed five steps for K-Means algorithm: (1) Arbitrarily choose k objects from D as the initial cluster centers; (2) Repeat; (3) (Re) assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster; (4) Update the cluster means, i.e., calculate the mean value of the objects for each cluster; and (5) Until no change; where k is the number of clusters, and D is a data set containing n objects.

The most important stage of K-Mean clustering algorithm is to determine the number of cluster; in other word deciding the value of K. In our proposed system, we decided the value of K as 6 targeting that the number of student in one group is about four due to the reason that at least 2 and at most 5 students should be in a group collaborative learning according to Johnson & Johnson (1999).

Clustering Analysis Result and Discussion

After we had done all data pre-processing stages and get a new data set via principle component analysis as shown in Figure 5, we applied K-Meaning Clustering algorithm to this new data set. As a result, we got an output in which clusters of components of new data set can be seen as shown in Figure 6. After that, we integrated this result to initial data which means data with same size to original data in order to get finalized data. According to final results shown in figure 7, we found that cluster 0 and cluster 4 had two students respectively while cluster 2 and cluster 5 stands with six students in each. Unlike other clusters, cluster 3 has seven students and cluster 1 possesses four students.

PC1	PC2	PC3	...	PC14	PC15	ClusterID
1.595324	0.105983	-0.875856	...	0.219723	0.030235	3
2.496133	0.852213	2.382664	...	-0.321496	0.087751	3
-1.351051	1.429692	0.281185	...	0.612708	-0.563851	5
-1.173077	-1.168322	2.354893	...	-1.047280	0.339959	1
-1.466784	-1.980877	-1.651553	...	-0.174610	-1.250551	1
1.284443	0.098350	1.987432	...	0.409770	-0.880005	3
0.408275	-0.294704	-0.576540	...	0.230531	0.511582	3
0.450440	0.494688	-0.942574	...	-0.484379	-0.031278	5
-0.765306	-0.242670	2.498679	...	-1.030195	-0.404444	2
-1.056105	1.379964	0.449117	...	0.007582	0.137108	0
-3.247024	0.269276	-2.105386	...	-0.282511	-0.269802	5
0.886622	-2.328942	0.007489	...	0.045342	-0.117104	2
2.163281	0.981455	-1.468905	...	0.688552	0.243300	3
0.857690	-2.036056	0.002346	...	-0.588359	-0.053659	2
-1.638978	-0.064278	-1.945133	...	-0.348003	0.377057	5
-2.368671	-2.230059	1.356742	...	0.571551	0.202716	1
-0.994854	2.726695	0.554623	...	-0.513644	0.566621	4
4.597196	0.921338	-1.073432	...	-0.438596	-0.195548	3
-1.097413	3.173216	1.045240	...	0.772549	-0.303755	4
1.706604	-1.982412	0.855020	...	0.466852	0.223135	2
-0.873590	1.532422	0.584746	...	-0.129518	0.047030	0
-3.247024	0.269276	-2.105386	...	-0.282511	-0.269802	5
0.521591	-2.633858	-0.263768	...	0.319542	0.063051	2
1.980765	0.828998	-1.604534	...	0.825652	0.333378	3
0.675175	-2.188514	-0.133282	...	-0.451259	0.036419	2
-1.821493	-0.216736	-2.080761	...	-0.210903	0.467135	5
-2.551186	-2.382516	1.221113	...	0.708651	0.292793	1
-0.994854	2.726695	0.554623	...	-0.513644	0.566621	4

Figure 6 Principal components with cluster ID

Roll Num	Q1	Q2	Q3	Q4	Q5	Q6	...	Q15	Q16	Q17	Q18	Q19	Q20	ClusterID
R1	2	2	2	2	2	1	...	2	3	2	2	3	1	3
R2	3	3	3	1	1	1	...	1	3	2	2	3	1	3
R3	2	1	2	2	3	2	...	3	1	2	1	3	3	5
R4	2	3	1	2	1	1	...	2	1	1	1	2	2	1
R5	1	1	1	2	3	1	...	2	3	1	1	2	1	1
R6	3	2	3	2	1	1	...	2	2	2	2	3	1	3
R7	2	3	2	2	3	1	...	2	2	2	2	3	1	3
R8	2	3	3	1	3	1	...	2	3	2	2	3	2	5
R9	2	3	3	3	1	3	...	1	1	2	1	3	3	2
R10	2	1	3	3	3	1	...	1	2	1	2	3	1	0
R11	1	2	2	3	3	3	...	1	3	2	1	3	1	5
R12	2	3	1	3	1	1	...	2	3	1	1	3	3	2
R13	2	3	3	2	3	1	...	2	2	1	2	3	3	3
R14	2	3	2	3	1	1	...	2	2	1	1	3	2	2
R15	2	1	2	3	3	2	...	2	3	2	1	3	2	5
R16	2	3	1	3	3	1	...	1	2	2	1	2	2	1
R17	2	1	1	2	3	2	...	2	2	2	2	3	2	4
R18	3	3	3	2	1	1	...	2	3	1	3	3	2	3
R19	2	3	3	3	3	3	...	2	2	2	2	3	2	4
R20	2	3	3	3	1	1	...	2	3	1	1	3	3	2
R21	2	1	3	3	3	1	...	1	2	1	2	3	2	0
R22	1	2	2	3	3	3	...	1	3	2	1	3	1	5
R23	2	3	1	3	1	1	...	2	3	1	1	3	1	2
R24	2	3	3	2	3	1	...	2	2	1	2	3	2	3
R25	2	3	2	3	1	1	...	2	2	1	1	3	1	2
R26	2	1	2	3	3	2	...	2	3	2	1	3	1	5
R27	2	3	1	3	3	1	...	1	2	2	1	2	1	1

Figure 7 Finalized Result

These are the result of our proposed system (SCS), which applied K-Means clustering to group the students based on students' common interests with a goal to promote collaborative learning. By the final results, our proposed system focus on grouping students based on their opinions, their answers to questionnaire. Moreover, the data size of our data set can be effectively reduced by principal component analysis and it is also an advantage to K- Means clustering algorithm. K-Means clustering algorithm is also better compared to other algorithm in terms of speed and simplicity.

Conclusion and Outlook

Although educational informatization is widely spread and improved in developed countries in order to promote collaborative learning, it is still hard to implement in developing countries because of the lack of data sources. In developed countries, it is easy and convenient to acquire the required data sources for collaborative learning because of the availability and mushroom existence of E-Learning classrooms. However, E-Learning classrooms and school information system are not prospered in developing countries. Therefore, we designed a questionnaire that reflects the common interests of the students especially in computer science to collect data and then we applied K-Means clustering to data in order to group students in accordance with their common interests after principal component analysis was performed as data pre-processing. By the results of our proposed system, teachers can group their students based on their common interests and enhance the collaborative learning.

In the future, researchers can perform more clustering algorithm by combining two or more them as a hybrid approach to make more reliability of clustering. Moreover, the research material (30 students) we used in this study is relatively small because we use the students in one classroom as research objects. Therefore, further studies should be carried out with larger amount of class size improving the proposed system. By conducting these researches more, our education system should be moved forward and better with the help of educational informatization.

Acknowledgements

My sincerest gratitude to my supervisor, Professor Khin Myo Sett, for her insights and guidance throughout the research. She was always happy to meet me and discuss our work. She encouraged and motivated me throughout the process. I would also like to thank all the teachers in our department, for guiding me throughout my this study. I also thank to final students who are cooperative in my questioner approach to obtain data. Finally, I also owe much to my family; my special appreciation and gratitude go to my parents for being a source of encouragement.

References

C. Romero, S. Ventura. (2007) Educational data mining: A survey from 1995 to 2005[J]. Journal of Expert Systems with Applications, (1-33): 135-146.

Chang, Y.C., Kao, W.Y., Chu, C.P., Chiu, C.H. (2009): A learning style classification mechanism for e-learning. Comput. Educ. 53, 273–285.

Dunn, R. (1984): Learning style: state of the science. Theory Pact. 23(1), 10–19.

E. Gaudioso, M. Montero, L. Talavera, and F. Hernandez-del-Olmo. (2009) Supporting teachers in collaborative student modeling: A framework and an implementation [J]. Expert System with Applications, (36): 2260-2265.

Han, J.W., Kamber, M., Pei, J. (2004): Data Mining, 3rd edn. Morgan Kaufmann Publishers Inc., San Francisco.

Hu, H., He, J.H. (2014): Research of composing cooperative learning group based on enhanced ant colony optimization algorithm. Comput. Eng. Appl. 50(13), 137–141.

John A. Hartigan. (1975): Clustering Algorithms, John Wiley & Sons New York, London, Sydney, Toronto.

Johnson, D. W., & Johnson, R. T. (1999). Making cooperative learning work. *Theory Into Practice*, *38*, 67–73. 10.1080/00405849909543834

Ma, Y.Y., Yuan, J. (2016): Based on GSDBK – means grouping algorithm research for networked collaborative learning. Electron. Sci. Technol. 29(12), 89–92.

Mishra, Sidharth & Sarkar, Uttam & Taraphder, Subhash & Datta, Sanjoy & Swain, Devi & Saikhom, Reshma & Panda, Sasmita & Laishram, Menalsh. (2017). Principal Component Analysis. International Journal of Livestock Research. 1. 10.5455/ijlr.20170415115235.

Singhal, Divya. (2017). Understanding Student- Centered Learning and Philosophies of Teaching Practices. International Journal of scientific research and management. 10.18535/ijsrm/v5i2.02.

Tamilselvi, R., Sivasakthi, B. and Kavitha, R., (2015). AN EFFICIENT PREPROCESSING AND POSTPROCESSING TECHNIQUES IN DATA MINING.

Tang, J., Li, H.J., Qiu, F.Y. (2012): A research on collaborative grouping peer-model in mCSCL. Distant Educ. China 2, 48–51.

Zorrilla, Marta & García-Saiz, Diego. (2013). A service oriented architecture to provide data mining services for non-expert data miners. Decision Support Systems. 55. 399-411. 10.1016/j.dss.2012.05.045.