Journal of Theoretical and Applied Information Technology
15th April 2018. Vol.96. No 7

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CLASSIFYING STUDENTS' ANSWERS USING CLUSTERING ALGORITHMS BASED ON PRINCIPLE COMPONENT ANALYSIS

ALAA KHALAF HAMOUD
Computer Information System Department, University of Basrah, Iraq
E-mail: alaak7alaf@gmail.com

ABSTRACT

Recently, almost all the academic institutions focus on finding the factors which increase the educational outcomes. Due to its importance in reflecting achievement of the academic organizations, students' success is a significant goal pursued by all educational institutions and grabbed their wide attention. Different data mining tools and algorithms are implemented and used for increasing students' academic success and analysing the factor affect the student's performance. The paper introduces a proposed model for classifying students' answers using four clustering algorithms (EM, Hierarchical Clustering, Make Density Based and K-Means). PCA is used for attributes selection for two reasons; the first one is to reduce the number of attributes in order to increase the results accuracy, and second reason to find the most relevant attributes which affect on the final class. A comparison is made between clustering algorithms based on specific classification performance to find the optimal one for clustering. The questionnaire consists of 62 questions that cover the most related fields such as health, social activity, relationships and academic performance. Google form and LimeSurevey questionnaire (open source application) are used to build the questionnaire and the total number of students' answers is 161 answers. The answers of students are collected from two departments (Computer Science and Computer Information Systems) in the college of Computer Science and Information Technology, University of Basrah, Iraq. Weka 3.8 tool is used to build and implement the model. The overall model design process can be divided into four stages, the first stage is data pre-processing, and the second one is applying PCA to find the correlated attributes. In the third stage, four proposed clustering algorithms are applied and the final stage the optimal algorithm is selected based on a comparison between algorithms.

Keywords: *Educational Data Mining (EDM), Student's Answers, Clustering Algorithm, Principles Component Analysis (PCA), Classification, Weka.*

1. INTRODUCTION

Monitoring and evaluation of student's performance is an important task in learning process. Finding the factors which affect on the performance of student's can be very helpful in detecting the relationship between student's performance and other factors in education setting. This process can help to categorize students and provide more focus towards improving their performance [1].

The main aim of educational institutions is to provide student with the evaluation reports regarding their test/ examination as best as possible with minimum errors. Some factors other than academic have been reported to creates/pose barrier to students attaining and maintaining their high performance [2][3]. Many works related to student's academic performance used different classification and prediction algorithms to help teachers/students to know the most related factors which affect student's success/failure [4][5].

The aim of the model is to classifying student's answers in order to analyzing them to get analytical results. The data set consists of students' answers of questionnaire related to assessment students'

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

performance. The questionnaire consists of 61 questions which built based on different student's questionnaires [6][7]. The final data set of answers held many uncompleted answers with uncleaned rows. The paper proposed a model based on a comparison between four different clustering algorithms to classify student's answers. The model passed through different stage starting with preprocessing data (clean data), selecting attributed based on PCA, applying algorithms and evaluating results. The objectives of the model are finding the most attributes (questions) which affect the final class (Failed) and exploring the instances which construct the clusters. The result of the model can be used for decision support related to academic stuff and students. The paper organized as follow, section 2 lists several papers which list the related works to the proposed model. Section 3 defines and explains briefly Educational Data Mining (EDM) and its stages. Section 4 explains theoretically the proposed clustering algorithms. Section 5 lists the steps of implementing proposed model while the final section lists the conclusion points.

2. RELATED WORKS

In [8], M. Singh, A. Rani, and R. Sharmak proposed a study in which K-means clustering technique is applied to analyze student academic performance. This study makes use of cluster analysis to segment students into groups according to their characteristics. This include many factors like class internal marks, GPA, mid and final exam, assignment, lab–work are studied. It is recommended that all these correlated information should be conveyed to the class teacher before the conduction of final exam. The paper presented an optimal procedure based on K-Means Clustering algorithm using Weka Interface that enables academicians to enhance the student's education quality and the instructor can take necessary steps to improve student academic performance based on it.

In [9], Chady El Moucary, Marie Khair, and Walid Zakhem presented a hybrid procedure based on Neural Networks (NN) and Data Clustering that enables academicians to predict students' GPA according to their foreign language performance at a first stage, then classify the student in a well-defined cluster for further advising and follow up by forming a new system entry. This procedure has mainly a twofold objective in which it allows meticulous advising during registration and thus, helps maintain high retention rate, acceptable GPA and grant management. Additionally, it provides the instructors an anticipated estimation of their students' capabilities during team forming and in-class participation. The results demonstrated a high level of accuracy and efficiency in identifying slow, moderate and fast learners and in endowing advisors as well as instructors an efficient tool in tackling this specific aspect of the learners' academic standards and path.

R. Sasi regha and R. Uma in [10] introduced a model that establishing the novel technique for feature or attributes selection process by hybrid of Artificial fish swarm-Cuckoo Search optimization algorithm to remove the irrelevant features or obtaining relevant features. This study used to detect and remove the both irrelevant and redundant features that can be used to enhance the classification accuracy in predicting the student performance. This goal is achieved by Also, Non-negative Matrix Factorization Clustering algorithm (NMFC) performs the removal of redundant feature or attributes which are presented in the relevant features. The performance of this technique is analyzed by using the student database which comprises the gathering of student's information from different colleges. For analyzing the performance of this technique, the comparative evaluation is carried out between the classifiers used in this research such as Prism and J48 without the feature selection and classifiers with our proposed technique. The experimental consequences illustrate that hybrid of artificial fish swarm-cuckoo search optimization feature selection along with NMFC approach is accomplishing high accuracy rate than other techniques. This study facilitates us to enhance the performance of the student's failure and dropout prediction. In other words, this helps to increase the accuracy of the classification result.

In [11] Hillol Kargupta, Weiyun Huang, Krishnamoorthy Sivakumar, and Erik Johnson proposed a distributed clustering of high-dimensional heterogeneous data using a distributed

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

principal component analysis (PCA) technique called the collective PCA. They presented the collective PCA technique, which can be used independent of the clustering application. It shows a way to integrate the Collective PCA with a given o–off-the-shelf clustering algorithm in order to develop a distributed clustering technique. It also presented experimental results using different test data sets including an application for web mining.

In [12] R. K. Arora and D. Badal, [13] J. J. Manoharan, S. H. Ganesh and M. L. P. Felciah and [14] I. Singh, A S. Sabitha, and A. Bansal proposed a system for analyzing the performance of students using k-means clustering algorithm. The result of analysis can be used to assist the academic planners in evaluating the performance of students during specific period of time and steps that need to be taken to improve students' performance from next batch onwards.

3. EDUCATIONAL DATA MINING (EDM)

The Educational Data Mining (EDM) community website, www.educationaldatamining.org, defines educational data mining as follows: "Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in". Applying data mining (DM) in education is an emerging interdisciplinary research field also known as EDM. It is concerned with developing methods for exploring the unique types of data that come from educational environments. Its goal is to improve students' outcome and better understand how students learn and identify the settings in which they learn to gain insights into and explain educational phenomena [15][16][17].

EDM is a young research area and it is necessary more specialized and oriented work educational domain in order to obtain a similar application success level to other areas, such as medical data mining, mining e-commerce data, mining spatial data, and web mining [18]. EDM can be seen in two ways; either as a research community or as an area of scientific inquiry. As a research community, EDM can be seen as a sister community to learning analytics. EDM first emerged in a workshop series starting in 2005, which became an annual conference in 2008 and spawned a journal in 2009 and society, the International Educational Data Mining Society, in 2011[19].

The first stage corresponds to the provision of EDM proactive support for adapting the educational setting according to the student's profile prior to deliver a lecture. During the student system interaction stage, it is desirable that EDM acquires log data and interprets their meaning in order to suggest recommendations to help students and lecturers. In the next stage, EDM should carry out the evaluation of the provided education concerning: delivered services, achieved outcomes, degree of user's satisfaction, and usefulness of the resources employed. What is more, several challenges (i.e., targets, environments, modalities, functionalities, kinds of data) wait to be tackled or have been recently considered by EDM, such as: big data, cloud computing, social networks, web mining, text mining, virtual 3-D environments, spatial mining, semantic mining, collaborative learning, learning companions [20].

4. CLUSTERING ALGORITHMS

Clustering techniques apply when there is no class to be predicted but the instances are to be divided into natural groups. These clusters presumably reflect some mechanism that is at work in the domain from which instances are drawn, a mechanism that causes some instances to bear a stronger resemblance to each other than they do to the remaining instances. Clustering naturally requires different techniques to the classification and association learning methods that we have considered so far [21].

Clustering techniques consider data tuples as objects. They partition the objects into groups, or clusters, so that objects within a cluster are "similar" to one another and "dissimilar" to objects in other clusters. Similarity is commonly defined in terms of how "close" the objects are in space, based on a distance function. The "quality" of a cluster may be represented by its diameter, the maximum distance between any two objects in the cluster. Centroid distance is an alternative measure of cluster quality and is defined as the average distance of each cluster object from the cluster centroid (denoting the "average object," or average point in space for the cluster). Figure 3.3 showed a 2-D plot of customer data with respect to customer locations in a city. Three data clusters are visible. In

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

data reduction, the cluster representations of the data are used to replace the actual data. The effectiveness of this technique depends on the data's nature. It is much more effective for data that can be organized into distinct clusters than for smeared data. There are many measures for defining clusters and cluster quality [22].

4.1. K-Means Algorithm

Figure (1) lists the steps of the K-means clustering algorithm. The K-means clustering algorithm starts with a given K value and the initially assigned centroids of the K clusters. The algorithm proceeds by having each of n data points in the data set join its closest cluster and updating the centroids of the clusters until the centroids of the clusters do not change any more and consequently each data point does not move from its current cluster to another cluster. In Step 7 of the algorithm, if there is any change of cluster centroids in Steps 3–6, we have to check if the change of cluster centroids causes the further movement of any data point by going back to Step 2. To determine the closest cluster to a data point, the distance of a data point to a data cluster needs to be computed. The mean vector of data points in a cluster is often used as the centroid of the cluster. Using a measure of similarity or dissimilarity, we compute the distance of a data point to the centroid of the cluster as the distance of a data point to the cluster. One method of assigning the initial centroids of the K clusters is to randomly select K data points from the data set and use these data points to set up the centroids of the K clusters. Although this method uses specific data points to set up the initial centroids of the K clusters, the K clusters have no data point in each of them initially. There are also other methods of setting up the initial centroids of the K clusters, such as using the result of a hierarchical clustering to obtain the K clusters and using the centroids of these clusters as the initial centroids of the K clusters for the K-means clustering algorithm.

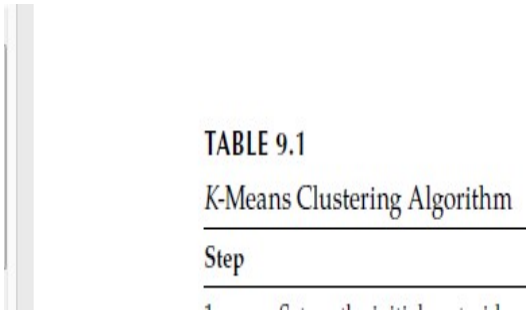
TABLE 9.1
K-Means Clustering Algorithm
Step

Figure 1:K-Means Algorithm

For a large data set, the stopping criterion for the REPEAT-UNTIL loop in Step 7 of the algorithm can be relaxed so that the REPEAT-UNTIL loop stops when the amount of changes to the cluster centroids is less than a threshold, e.g., less than 5% of the data points changing their clusters[23].

4.2. Expectation Maximisation (EM)

The EM algorithm was discovered and employed independently by several different researchers until Dempster et al. brought their ideas together, proved convergence, and coined the term "EM algorithm." Since that seminal work, hundreds of papers employing the EM algorithm in many areas have been published. A typical application area of the EM algorithm is in genetics, where the observed data (the phenotype) is a function of the underlying, unobserved gene pattern (the genotype). Another area is estimating parameters of mixture distributions. The EM algorithm has also been widely used in econometric, clinical, and sociological studies that have unknown factors affecting the outcomes [24].

The steps of the EM algorithm for this case are [25]:

(1) Initialise with the starting distribution of the state ¯ X1 (mean and variance), and an initial estimate for the dynamics (¯x, A0,A1,B0);

(2) Run the Kalman filter on the measurement set to produce the filtered estimates ¯ Xkk;

(3) Run the smoothing algorithm on the measurement set (using the filtered estimates) to produce the smoothed estimates ¯ XkN;

(4) Using the results of the smoothing, find the expected values ∑[Si] and ∑[Sij] the moments;

(5) From the expected values of the moments, estimate the system parameters (¯x, A0,A1,B0);

(6) Using the system derived in the previous step, go back to step (2).

The algorithm should be run, as is usual for EM applications, until convergence of the parameters of interest, namely (¯x, A0,A1,B0).

4.3. Hierarchical Clustering

Hierarchical clustering produces groups of similar data points at different levels of similarity. This chapter introduces a bottom-up procedure of hierarchical clustering, called agglomerative hierarchical clustering. A list of software packages that support hierarchical clustering is provided. Some applications of hierarchical clustering are given with references. Given a number of data records in a data set, the agglomerative hierarchical clustering algorithm produces clusters of similar data records in the following steps:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

1. Start with clusters, each of which has one data record.
2. Merge the two closest clusters to form a new cluster that replaces the two original clusters and contains data records from the two original clusters.
3. Repeat Step 2 until there is only one cluster left that contains all the data records.

The next section gives several methods of determining the two closest clusters in Step 2[23].

4.4. Make Density Based Clustering Algorithm

The key idea of the DBSCAN method is that, for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points, that is, the density in the neighborhood has to exceed some predefined threshold. This method needs three input parameters [26]:

· k, the neighbor list size;
· ϵ , the radius that delimitate the neighborhood area of a point (ϵ-neighborhood);
· MinPts, the minimum number of points that must exist in the ϵ-neighborhood.

The clustering process is based on the classification of the points in the dataset as core points, border points and noise points, and on the use of density relations between points (directly density-reachable, density-reachable, density-connected) to form the clusters.

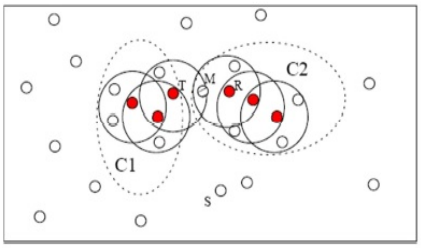

Figure 2: Two clusters discovered by Density Based

In figure (2), two clusters, C1 and C2, are discovered by the DBSCAN algorithm. Let min pts = 3. Here, core objects are represented by solid points while the non-core objects are represented by hollow points. Data objects that belongs to C1 or C2 all lies within the ϵ-neighborhood at least one core object from C1 or C2and there are no two core objects such that they lie within the ϵ-neighborhood of each other and yet belong to different clusters. A non-core object like M lies within the ϵ-neighborhood of T and R which are core objects from C1 and C2 respectively. It thus can be assigned to either C1 or C2. Finally, the object S is seemed to be noise because it is not in the ϵ-neighborhood of any core object [27].

5. Principal component analysis (PCA)

Principal component analysis (PCA) is a statistical technique of representing high-dimensional data in a low-dimensional space. PCA is usually used to reduce the dimensionality of data so that the data can be further visualized or analyzed in a low-dimensional space. For example, we may use PCA to represent data records with 100 attribute variables by data records with only
2or 3 variables [23].

Principal component analysis (PCA) in many ways forms the basis for multivalued data analysis.

PCA provides an approximation of a data table, a data matrix, X, in terms of the product of two small matrices T and P'. These matrices, T and P', capture the essential data patterns of X.

Plotting the columns of T gives a picture of the dominant "object patterns" of X and, analogously, plotting the rows of P' shows the complementary " variable patterns"[28].

PCA can be generalized as correspondence analysis (CA) in order to handle qualitative variables and as multiple factor analysis (MFA) in order to handle heterogeneous sets of variables. Mathematically, PCA depends upon the eigen-decomposition of positive semidefinite matrices and upon the singular value decomposition (SVD) of rectangular matrices [29].

Both PCA and factor analysis aim to reduce the dimensionality of a set of data, but the approaches used to do so are different for the two techniques. Principal Component analysis has been extensively used as part of factor analysis, but this involves 'bending the rules' which govern factor analysis, and there is much confusion in the literature over the similarities and differences between the techniques [30].

6. Model

As shown in Figure (3) below, the model construction process passes through four steps starting with data preprocessing and ending with results evaluation. Firstly, the process of building questionnaire considered as a part of data preprocessing step. Data preprocessing step includes all processes of preparing data set for evaluation, cleaning data, converting data ranges and creating derived column (Failed) based on column (Number of Failed Courses). The column (Failed) created based on a simple condition:

If (Number of Failed Courses>0) then Failed='F'
Else Failed='P';

where F is an abbreviation of Failed, P is an abbreviation of Passed;

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The column "Failed" is considered as the goal class of the model.

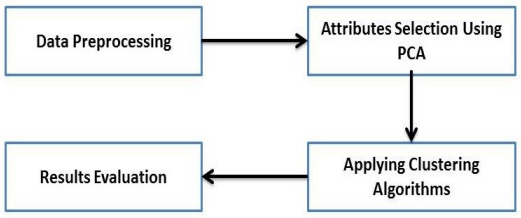

Figure 3: Model Diagram

6.1. Data Preprocessing

Data preprocessing involves the following steps:

A. Data Collection: The questionnaires were built based on Google form and open source application (LimeSurvey) to collect students' answers of (CSIT) College of Computer Science and Information Technology, University of Basrah. The first questionnaire (based on limesurvey) is constructed to collect answers locally from the college building of CSIT while Google form used to collect the answers over Internet. The total number of students' answers is 161 after combining the result "csv" (Comma Separated Values) files from Google form and LimeSurvey questionnaire. The research sample (161 answers) represents an acceptable sample of the population of CSIT with 10% as the percentage of error for the result of this study [31].

Table (1) below shows both questions' description and questions' answers' range of the questionnaire. The description of all questions in this table is shortened so it can be seen in tree nodes and can be understood easily. Questions' ranges also shortened and converted from nominal to numeric type for ease of use and understand.

Since the questionnaire holds more than 60 questions, so there is a need to shorten the questions. Questions' descriptions were shortened in order to be used by Bayes algorithms. Table (2) lists some of question's description which were shortened in the previous table.

The first step in data preprocessing is preparing data for processing by removing rows with empty values and converting data so it can be evaluated and processed. The number of rows with empty values in one or more columns is 6 rows. After removing these rows the total number of answers became 156 answers. The second step is converting rows values in order to process them in weka 3.8 tool with its built in classifiers.

B:Reliability: Reliability is used to describe the overall consistency of a measure. A measure is said to have a high reliability if it produces similar results under consist conditions. For example, measurements of people height and weight are often extremely reliable [32]. In statistics, the coefficient alpha is the most frequently used method for the calculation of internal consistency that is used as a measure of reliability for the dependent variable of the study. With Cronbach's alpha 0.7, it indicates satisfactory internal consistency in reliability [33]. Table (3) shows that the coefficient alpha is 0.85 for the scaled variables which contain 62 items and 161 respondents.

Table 3: Questionnaire Reliability

Cronbach's alpha	No. of items	No. of respondent	% of respondent
0.85	62	161	100%

6.2. Attribute Selection using Principle Component Analysis (PCA)

The second step in the model is attributes selection. Since there are 61 attributes in the questionnaire each one represent a question, so it is important to find the most correlated questions to the final class. In the model, Principle Component Analysis (PCA) with search method (Ranker) is proposed as an attribute selector to the final class (Failed or Pass) with evaluation mode (10 fold cross validation) to ensure high accuracy result. The MaximumeAttributeName property in the attribute evaluator is set to 1 to list only one attribute per row.

Table 4: Attributes Ranking Based on PCA

Variance	Question Rank	Question Abbreviation	Question Number
0.7952	1	OptimToAchvGoals	24
0.7307	2	YearsOfStudy	17
0.6764	3	EnoughBudget	57
0.6327	4	ResponAbtMyEdu	48
0.5936	5	LiveWithParent	8
0.5584	6	RequestHelpFromOthers	58
0.526	7	LDegNotMakeMeFail	22
0.4961	8	LiveWithParent	8
0.4684	9	MakeFriendship	30
0.4426	10	CalmDurExam	21
0.4181	11	Address	5
0.3948	12	Credits	14
0.3724	13	Address	5
0.3517	14	IHvSkillToSchvAcadmWork	60
0.3331	15	CalmDurExam	21
0.3152	16	MotherWork	11
0.2984	17	ParentAlive	9
0.282	18	Work	7
0.2667	19	Dep	1
0.2523	20	FatherWork	10
0.2388	21	ListImporPoints	18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

0.2259	22	CnUseLaptpToAchvSucc	36
0.2139	23	PlanToNotReadAgian	39
0.202	24	AbsenceDays	13
0.1905	25	ListImporPoints	18
0.1795	26	EduIsLiveJob	46
0.169	27	ExiToMater	26
0.1592	28	CalmDurExam	21
0.1498	29	RequestHelpFromOthers	58
0.1407	30	PlanToNotReadAgian	39
0.1318	31	DevRelationWithOthers	28
0.1232	32	ClrIdeaABoutPlans	51
0.115	33	PlanToDoFunThing	40
0.107	34	ClrIdeaAbtMyBudget	42
0.0995	35	WorkedRecently	52
0.0924	36	PractRegular	33
0.0855	37	GPA	15
0.0789	38	PlanDaily	38
0.0728	39	WriteNotes	19
0.0671	40	Address	5
0.0616	41	CalmDurExam	21
0.0564	42	LDegNotMakeMeFail	22
0.0513	43	ResponAbtMyEdu	48
0.0468	44	RelationWithOthers	56

The table above the correlation between questions and the final class based on PCA attribute selector. The first column represents average number of variance of each question while the second column is the rank of each question arranging ascendingly. The column of question abbreviation represents the shortcut phrase of the question and the final column represents question number. The table helps us to find the questions with high correlation average and keep the top correlated question in the model. The total number of questions with high variance is 29 questions.

Based on PCA result the questions with high average variance will be chosen as base attributes of the model. The remaining questions will be removed to increase result accuracy. The questions with less correlation to the final class based on the PCA results are (2, 3, 4, 6, 12, 16, 20, 23, 25, 27, 29, 31, 32, 34, 35, 37, 41, 43, 44, 45, 47, 49, 50, 53, 54, 55, 59, and 61). These questions and the corresponding answers will be removed in order to get accurate result after applying clustering algorithms.

6.3. Applying Clustering Algorithms

Weka provides different clustering algorithms which can be used to get different results with different accuracy performance criteria's. Four different algorithms will be used in this stage (EM, Hierarchical Clustering, Density Based and K-Means) and all of them are applied after removing the less correlated questions. Removing attributes process is very helpful to discover the effectiveness of these attributes on the performance and how it can increase or decrease the accuracy.

Table 5: Clustering Algorithms Performance after Removing All Uncorrelated Attributes

Clustering Algorithms Via PCA	Avg. of TP Rate	Avg. of FP Rate	Average Precision	Average Recall
EM	0.536	0.498	0.526	0.536
Hierarchical Clustering	0.603	0.603	0.363	0.603
Make Density Based	0.551	0.494	0.549	0.551
K-Means	0.487	0.481	0.524	0.487

Table 5 shows that removing all uncorrelated attributes in one step is not granular because we want to find the attributes which affect on the result accuracy. The process of removing attributes is performed on the less correlated attributes and for one attribute in each step. After removing the first attribute, the performance criteria's (TP rate, FP rate, Precision and Recall) are collected to find the optimal clustering algorithm and to keep the attributes which enhance the result accuracy. TP Rate refers to total number of correctly classified instances, while FP Rate refers to total number falsely classified instances. Recall is also referred to as the TP Rate or sensitivity, and precision is also referred to as positive predictive value (PPV); other related measures used in classification include true negative rate and accuracy. The classifier (ClassificationViaClustering) is used in this step. This classifier provides the ability to select clustering algorithm and show performance results in easy way.

Figures (5, 6, 7, and 8) show the overall performance results (TP rate, FP rate, Precision and Recall) after applying the proposed clustering algorithms (EM, Hierarchical Clustering, Make Density Based and K-Means). Each single step in the previous charts is the result of removing the corresponding question number and applying the clustering algorithms using (ClassificationViaClustering) classifier. Based on PCA filter, there are 28 questions (table 6) detected as uncorrelated questions to the final class based on PCA. The table below lists and describes the uncorrelated questions to the final class (Failed and Success).

Table 6: Less Correlated Questions

Seq	Question Number	Abbreviation
1	2	Age
2	3	Stage
3	4	Gender
4	6	Status
5	12	FCourses

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

6	16	ComCredits
7	20	PrepStudySchedule
8	23	EasCanChosColgStudy
9	25	CnStudyEvUImpoBothMe
10	27	ClrIdeaAbtBenifit
11	29	ContrlMyAnger
12	31	OpenWithOthers
13	32	IHvEnrgyEnjoy
14	34	MyHealthHelp
15	35	FreshFood
16	37	PlanForWeek
17	41	ContrlMyBudget
18	43	CnWork
19	44	MyEduSuppoMyGoal
20	45	IHvSavPlan
21	47	ClrAbotMyLiveGoal
22	49	RespoAbtMyMyLiveQuality
23	50	RedyToFacChallng
24	53	KnowledWtBossExpectFrmMe
25	54	EduChoicesToAchivGoal
26	55	IHvEnoughMoney
27	59	TryToEnhancMySelf
28	61	IHvSkillsForSelfFeel

6.4. Results Evaluation

In this stage, the final evaluation of the clustering algorithms is performed to select the best clustering algorithm. In order to classify algorithms, the classifier ClassificationViaClustering is used in order to use each clustering algorithm alone and compare performance details of each other. Based on the results from figures (5,6,7, and 8), it can be clearly seen that EM algorithm is the best algorithm for cluster student answers based on the average of TP rate, FP Rate and time. In figure (5), the average TP rate of EM algorithm gets the highest result (0.67) with high average Precision (0.66) in figure (7). The average Recall gets also the highest rate (0.674) compared with other clustering algorithms (figure 8) while the FP rate gets the third ranked algorithm (figure 6). Since EM algorithm exceeds the other algorithms in (three of the four criteria's), so it can be selected as the best algorithm for classifying students' answers.

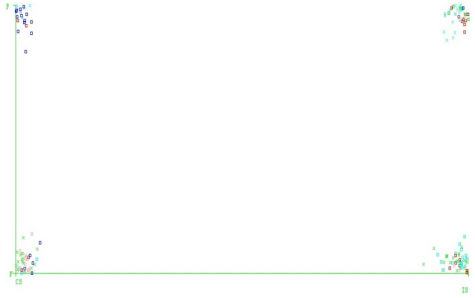

Figure 9: Clusters of EM Algorithm (Class with Departments)

Weka tool provides easy way to view cluster assignments. Figure (4) shows the final result of Make Density Based algorithm clusters with possibility to view clusters based on X axis and Y axis. The figure shows the results based on Department (X axis) and Failed (Y axis). I each colour represents a specific cluster which can be viewed by simple double click on the colour. It can be easily view the instances assigned to each cluster with each department and specific case (Pass or Failed) to find the instances which formulate the cluster.

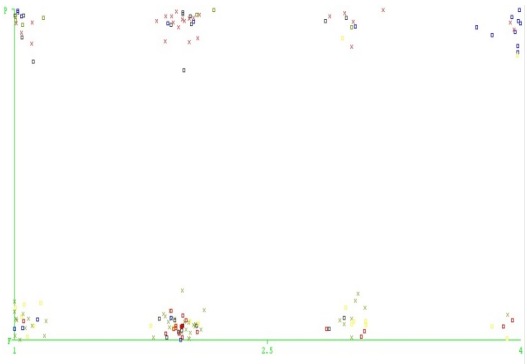

Figure 10: EM Clusters (Class to Year of study)

Figure 10 shows the clusters assignments according to year of studying and final class. According to the questionnaire, the number of studying years is four so the X axis shows the clusters assignment to four positions (first stage, second stage, third stage, and fourth stage). The important point in this figure is it can be easily classify the students according to stage and performance (success and failure). Each stage clusters can be viewed to find the assignment answers to each cluster and analyse the result. The clusters can help the lecturers and students to find the optimal way to success for each stage and avoid the points which led to failure.

Based on the result evaluation and attributes testing using PCA, it can be clearly seen that the model differs from prior works in three points:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

1. The model differs from prior works which applying PCA as attribute selector with clustering algorithms in testing each single attribute with four proposed algorithms to enhance the accuracy and finding the attributes which really affects on the accuracy.
2. The effectiveness of using PCA in attributes selection by removing single attribute per step and observing the result accuracy because PCA gives set of attributes with variances (correlation to final attribute).
3. The model discovers the attributes (questions) with less correlation to final class.

7. CONCLUSION AND FUTER WORK

The core of clustering in data mining is grouping data into similar objects in order to analyse and classify them. The objective of this work is to explore the possibility of analysing students' answers based on applying clustering algorithms on students' questionnaire answers. The questionnaire contains about 161 answers, some of them are uncompleted. The questionnaire contains also many unimportant questions which can be discovered by applying attribute filter. PCA is the proposed method to filter attributes and discover the most correlated attributes to the final class.

PCA result table can be depended to remove the uncorrelated attributes to increase the accuracy of the results. Many factors can affect the accuracy of the result such as large data set, less number of attributes and the clean date set. To get high accurate result, many steps are followed such as cleaning data set, removing the uncompleted answers and removing the less correlated questions to the final class. After discovering the uncorrelated attributes using PCA, the process of removing all uncorrelated attributes in single step led to decrease the final results accuracy. PCA gave us a list of uncorrelated attributes but not all of them are really uncorrelated. Removing one attribute per step gave us the privilege of increasing the accuracy and find the right attributes which affect or not on the accuracy.

The classification of students answers performed based on applying four clustering algorithms on the proposed questionnaire. Classification is important function in data mining which can be used for analysing data set and even for predication. A comparison is made between the proposed clustering algorithm based on many criteria's such as TP rate, FP rate, Precision and Recall. The selected clustering algorithm (EM Clustering Algorithm) is chosen based on performance criteria's as the proposed algorithm for classifying data. The model can be depended on by both students and academic staff of each department to decide which questions/answers will enhance the academic performance and then improve the institution success.

Since PCA gives only variance of attributes set to the final class, so the next step as a future work, an enhanced PCA can be implemented to list the correlation between attributes and final attribute with possibility of listing the percentage of attributes effectiveness to the overall accuracy. This kind of PCA can be used to give more accurate model and can be implemented with different data mining functions.

REFRENCES:

[1] Jayabal, Yogalakshmi, and Chandrashekar Ramanathan. "Clustering Students Based on Student's Performance-A Partial Least Squares Path Modeling (PLS-PM) Study." International Workshop on Machine Learning and Data Mining in Pattern Recognition. Springer, Cham, 2014.

[2] Sansgiry, Sujit S., Monali Bhosle, and Kavita Sail. "Factors that affect academic performance among pharmacy students." American journal of pharmaceutical education 70.5 (2006): 104.

[3] Patel, Jyotirmay, and Ramjeet Singh Yadav. "Applications of Clustering Algorithms in Academic Performance Evaluation." Open Access Library Journal 2.08 (2015): 1.

[4] A. K. Hamoud, "Selection of Best Decision Tree Algorithm for Prediction and Classification of Students´ Action," American International Journal of Research in Science, Technology, Engineering & Mathematics, vol. 1, no. 16, pp. 26-32, 2016/9.

[5] Alaa Khalaf Hamoud, Aqeel Majeed Humadi, Wid Akeel Awadh and Ali Salah Hashim. Students' Success Prediction based on Bayes Algorithms. International Journal of Computer Applications 178(7):6-12, November 2017.

[6] Mishra, Tripti, Dharminder Kumar, and Sangeeta Gupta. "Mining students' data for performance prediction." Fourth International Conference on Advanced Computing and Communication Technologies. 2014.

[7] Saa, Amjad Abu. "Educational Data Mining & Students' Performance Prediction." International Journal of Advanced Computer Science & Applications 1 (2016): 212-220.

[8] Mahesh Singh, Anita Rani, Ritu Shama. "An Optimised Approach For Student's Academic Performance By K-Means Clustering Algorithm Using Weka Interface " International Journal of Advanced Computational Engineering and Networking, Volume-2, Issue-7, July-2014

[9] Moucary, C. E., Marie Khair, and Walid Zakhem. "Improving student's performance using data clustering and neural networks in foreign-language

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

based higher education." The Research Bulletin of Jordan ACM 2.3 (2011): 27-34.

[10] R. Sasi regha, Dr R. Uma rani, "An Efficient Clustering Based Feature Selection for Predicting Student Performance", International Journal of Engineering and Technology (IJET), Vol 9 No 2 Apr-May 2017.

[11] Hillol Kargupta, Weiyun Huang, Krishnamoorthy Sivakumar, Erik Johnson," Distributed clustering using collective principal component analysis." Knowledge and Information Systems 3.4 (2001): 422-448.

[12] Arora, Rakesh Kumar, and Dharmendra Badal. "Evaluating Student's Performance Using k-Means Clustering." International Journal of Computer Science And Technology (IJCST) (2013).

[13] J. James Manoharan, S. Hari Ganesh, M. Lovelin Ponn Felciah, A.K. Shafreen Banu, "Discovering Students' Academic Performance Based on GPA Using K-Means Clustering Algorithm." Computing and Communication Technologies (WCCCT), 2014 World Congress on. IEEE, 2014.

[14] Singh, Ishwank, A. Sai Sabitha, and Abhay Bansal. "Student performance analysis using clustering algorithm." Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference. IEEE, 2016.

[15] Romero, C. and Ventura, S. (2013), Data mining in education. WIREs Data Mining Knowl Discov, 3: 12–27.

[16] Romero, Cristóbal, and Sebastián Ventura. "Educational data mining: a review of the state of the art." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 40.6 (2010): 601-618.

[17] Baker, Ryan SJD, and Kalina Yacef. "The state of educational data mining in 2009: A review and future visions." JEDM| Journal of Educational Data Mining 1.1 (2009): 3-17.

[18] Romero, Cristobal, and Sebastian Ventura. "Educational data mining: A survey from 1995 to 2005." Expert systems with applications 33.1 (2007): 135-146.

[19] Baker, Ryan Shaun, and Paul Salvador Inventado. "Educational data mining and learning analytics." Learning analytics. Springer New York, 2014. 61-75.

[20] Peña-Ayala, Alejandro. "Educational data mining: A survey and a data mining-based analysis of recent works." Expert systems with applications 41.4 (2014): 1432-1462.

[21] Witten, Ian H., et al. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

[22] Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2012.

[23] Ye, Nong. Data mining: theories, algorithms, and examples. CRC press, 2013.

[24] Moon, Todd K. "The expectation-maximization algorithm." IEEE Signal processing magazine 13.6 (1996): 47-60.

[25] North, Ben, and Andrew Blake. "Learning dynamical models using expectation-maximisation." Computer Vision, 1998. Sixth International Conference on. IEEE, 1998.

[26] Berkhin, Pavel. "A survey of clustering data mining techniques." Grouping multidimensional data 25 (2006): 71.

[27] Neha R. Gameti, 2 Prof. Ketan J. Sarvakar. "Density Based Methods to Discover Clusters with Arbitrary shape in weka", IJRIT International Journal of Research in Information Technology, Volume 1, Issue 5, May 2013, Pg. 280-286.

[28] Wold, Svante, Kim Esbensen, and Paul Geladi. "Principal component analysis." Chemometrics and intelligent laboratory systems 2.1-3 (1987): 37-52.

[29] Abdi, H. and Williams, L. J. (2010), Principal component analysis. WIREs Comp Stat, 2: 433–459. doi:10.1002/wics.101

[30] Jolliffe, Ian T. "Principal Component Analysis and Factor Analysis." Principal component analysis. Springer New York, 1986. 115-128.

[31] Israel, Glenn D. "Determining Sample Size. University of Florida IFAS extension." (2009).

[32] Carson B. "The transformative power of action learning." Retrieved 2017 from http://wial.org/executive-board/bea-carson-executive-board/.

[33] Sekaran, Uma, and Roger Bougie. Research methods for business: A skill building approach. John Wiley & Sons, 2016.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: Questionnaire Questions' Descriptions

Question	Description	Range	Question	Description	Range
Q1	Dep	IS,CS	Q32	IHvEnrgyEnjoy	1,2,3,4,5
Q2	Age	1,2,3,4	Q33	PractRegular	1,2,3,4,5
Q3	Stage	1,2,3,4	Q34	MyHealthHelp	1,2,3,4,5
Q4	Gender	F,M	Q35	FreshFood	1,2,3,4,5
Q5	Address	IN,OUT	Q36	CnUseLaptpToAchvSucc	1,2,3,4,5
Q6	Status	M,S	Q37	PlanForWeek	1,2,3,4,5
Q7	Work	YES,NO	Q38	PlanDaily	1,2,3,4,5
Q8	LiveWithParent	YES,NO	Q39	PlanToNotReadAgian	1,2,3,4,5
Q9	ParentAlive	0,1,2,3	Q40	PlanToDoFunThing	1,2,3,4,5
Q10	FatherWork	0,1,2	Q41	ContrlMyBudget	1,2,3,4,5
Q11	MotherWork	1,2	Q42	ClrIdeaAbtMyBudget	1,2,3,4,5
Q12	FCourses	0,1,2,3	Q43	CnWork	1,2,3,4,5
Q13	AbsenceDays	0,1,2	Q44	MyEduSuppoMyGoal	1,2,3,4,5
Q14	Credits	0,1,2	Q45	IHvSavPlan	1,2,3,4,5
Q15	GPA	1,2,3,4	Q46	EduIsLiveJob	1,2,3,4,5
Q16	ComCredits	1,2,3,4	Q47	ClrAbotMyLiveGoal	1,2,3,4,5
Q17	YearsOfStudy	1,2,3,4	Q48	ResponAbtMyEdu	1,2,3,4,5
Q18	ListImporPoints	1,2,3,4,5	Q49	RespoAbtMyMyLiveQuality	1,2,3,4,5
Q19	WriteNotes	1,2,3,4,5	Q50	RedyToFacChallng	1,2,3,4,5
Q20	PrepStudySchedule	1,2,3,4,5	Q51	ClrIdeaABoutPlans	1,2,3,4,5
Q21	CalmDurExam	1,2,3,4,5	Q52	WorkedRecently	1,2,3,4,5
Q22	LDegNotMakeMeFail	1,2,3,4,5	Q53	KnowledWtBossExpectFrmMe	1,2,3,4,5
Q23	EasCanChosColgStudy	1,2,3,4,5	Q54	EduChoicesToAchivGoal	1,2,3,4,5
Q24	OptimToAchvGoals	1,2,3,4,5	Q55	IHvEnoughMoney	1,2,3,4,5
Q25	CnStudyEvUImpoBothMe	1,2,3,4,5	Q56	RelationWithOthers	1,2,3,4,5
Q26	ExiToMater	1,2,3,4,5	Q57	EnoughBudget	1,2,3,4,5
Q27	ClrIdeaAbtBenifit	1,2,3,4,5	Q58	RequestHelpFromOthers	1,2,3,4,5
Q28	DevRelationWithOthers	1,2,3,4,5	Q59	TryToEnhancMySelf	1,2,3,4,5
Q29	ContrlMyAnger	1,2,3,4,5	Q60	IHvSkillToSchvAcadmWork	1,2,3,4,5
Q30	MakeFriendship	1,2,3,4,5	Q61	IHvSkillsForSelfFeel	1,2,3,4,5
Q31	OpenWithOthers	1,2,3,4,5			

Journal of Theoretical and Applied Information Technology
15th April 2018. Vol.96. No 7

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2: Questions' Abbreviations and Descriptions

Question	Abbreviation	Description
Q1	Dep	Your Department
Q2	Age	Your age
Q3	Stage	Your stage
Q4	Gender	Your gender
Q5	Address	Where do you live?
Q6	Status	Your status
Q7	Work	Are you working now?
Q8	LiveWithParent	Do you live with your parents?
Q9	ParentAlive	Are your parents alive?
Q10	FatherWork	What is your father's work scope?
Q11	MotherWork	What is your mother's work scope?
Q12	FCourses	Number of courses did you fail in per semester
Q13	AbsenceDays	Absence days per semester
Q14	Credits	Number or registered credits per semester
Q15	GPA	GPA
Q16	ComCredits	Number of completed credits
Q17	YearsOfStudy	Number of academic years till now
Q18	ListImporPoints	I can write down the important points during reading the material?
Q19	WriteNotes	During lecture, I can write notes and use them for exam preparation?
Q20	PrepStudySchedule	I prepare time schedule for studying?
Q21	CalmDurExam	During exam, I stay calm and coherent
Q22	LDegNotMakeMeFail	Getting low grades doesn't make me feel failure?

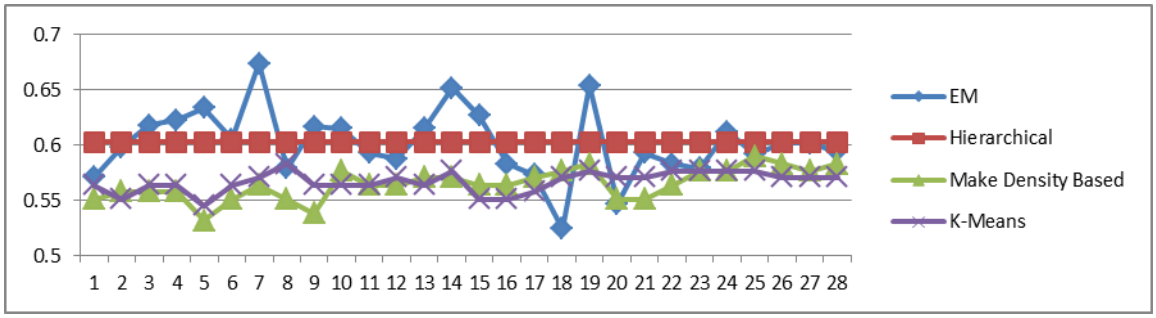

Figure 4:TP Rate for Clustering Algorithms

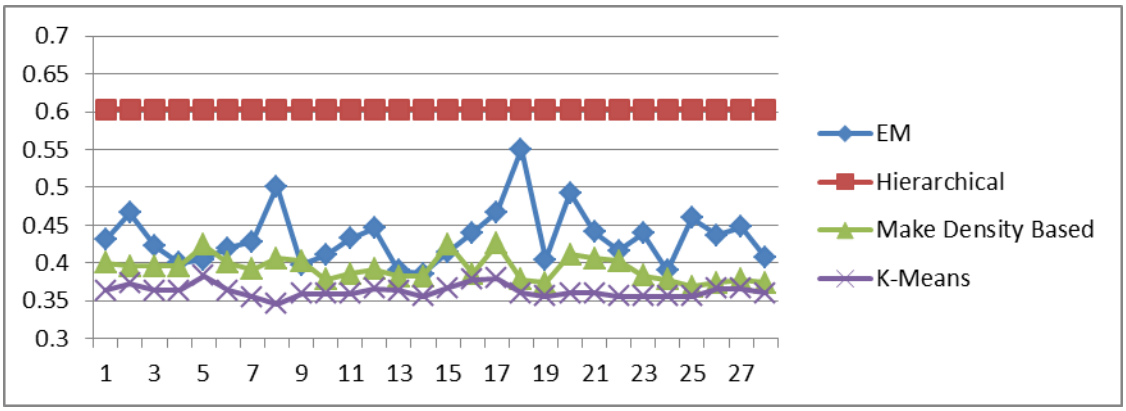

Figure 5: FP Rate for Clustering Algorithms

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

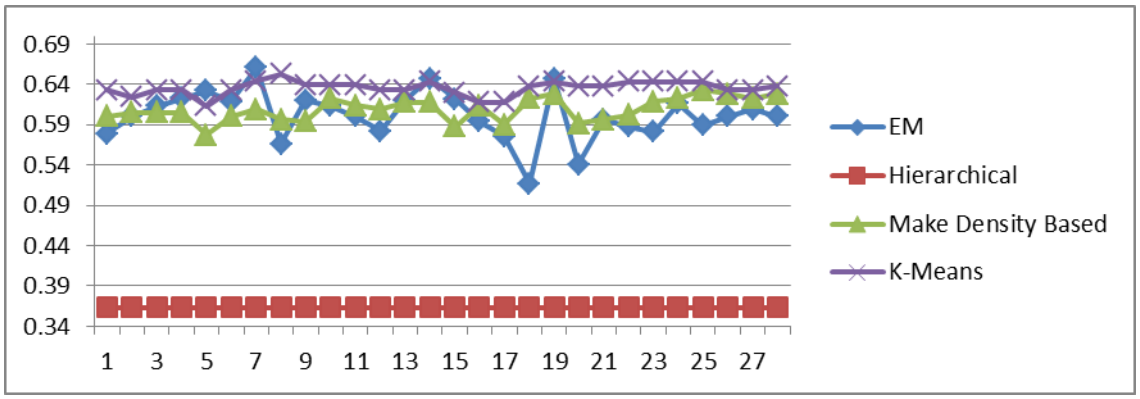

Figure 6: Precision of Clustering Algorithms

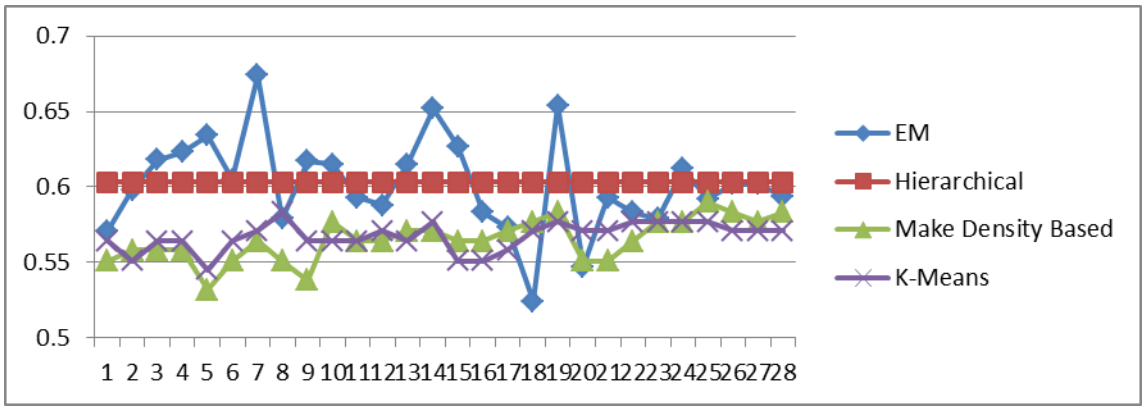

Figure 7: Recall of Clustering Algorithms