Investigating Performance of Students: a Longitudinal Study

Raheela Asif
Dept. of Computer Science & I.T.
N.E.D University of Engineering & Technology, Karachi, 75270, Pakistan
92-21-99261261
engr_raheela@yahoo.com

Agathe Merceron
Dept. of Computer Science & Media
Beuth University of Applied Sciences, Berlin, 13353, Germany
49-30-45045105
merceron@beuth-hochschule.de

Mahmood Khan Pathan
Dept. of Computer Science & I.T.
N.E.D University of Engineering & Technology, Karachi, 75270, Pakistan
92-21-99261261
mkpathan@hotmail.com

ABSTRACT
This paper, investigates how academic performance of students evolves over the years in their studies during a study programme. To determine typical progression patterns over the years, students are described by a 4 tuple (e.g. x1, x2, x3, x4), these being the clusters' mean to which a student belongs to in each year of the degree. For this purpose, two consecutive cohorts have been analyzed using X-means clustering. Interestingly the patterns found in both cohorts show that a substantial number of students stay in the same kind of groups during their studies.

Categories and Subject Descriptors
I.5.2 [**Pattern Recognition**]: Design Methodology – *Pattern Analysis.* J.1. [**Administrative Data Processing**]: *Education*.

General Terms
Measurement, Performance, Design, Human Factors.

Keywords
Students' performance, progression, X-means clustering

1. INTRODUCTION
While there are many studies predicting performance of students, few works have investigated how performance of students evolves during their studies. It would be interesting to answer some questions that academics confront as they witness different patterns of students performance during the course of their studies. For instance, do students with a poor performance in first year progress steadily, or on the contrary, keep earning low marks all the way through their studies? Do students with a high performance in first year retain this high performance till the end of their studies? Teachers do encounter these kinds of patterns. Are these generalizable to an entire cohort? How scattered is the performance of students' in first year? Does it become more heterogeneous or more homogeneous over time? Can these trends be calculated in a way that can be easily explained to stakeholders and intuitively visualized? This paper presents a case study in that area taking marks of students of two follower cohorts of a four-year degree in information technology discipline. The patterns noted during analysis indicate that there is some stability in both cohorts and that a substantial number of students stay in the same kind of group during their studies.

A first investigation on the evolution of performance of students has been presented in [1]. The focus in [1] is on comparing the performance of first year with the performance of students in the subsequent years to study if it stays the same or show some decrease or increase. The progression of a student is defined and used for clustering. Students in the same cluster have a similar progression. For example students in the 'stable' cluster have a performance that always stays in the same interval throughout their studies. The 'stable' cluster can contain top students or poor students with former always having their performance in a high interval, and the latter showing a low performance throughout their studies. The present study complements this work by exploring the change in the performance of students in comparison with other students: students in the same cluster have similar marks during their studies.

An analysis that bears strong similarities with the present work is [3] that uses all K-12 marks in all topics to cluster school students using hierarchical clustering. Dendrograms are combined with heatmaps to provide an intuitive visualization of each student and of the groups in the same picture. However, the present study does not cluster the students collectively taking all their marks in all subjects as performed in [3]. Instead, students are clustered each year and then the clusters are combined. This allows to exhibit small groups like the group of students who have low marks in first year but progress and finish with high marks in last year, or vice versa.

The paper is organized in five sections. After introducing the study in the first section, data and analysis methods and results are presented in sections 2 and 3 respectively. Discussions are presented in section 4 followed by conclusions in the last section that also reflects on the present work and discusses future research.

2. DATA AND METHODOLOGY
Students' marks of a four-year IT bachelor degree of the NED University of Engineering & Technology (NEDUET), Karachi, have been used in this study. Two subsequent cohorts have been analysed to better identify the recurrent trends. Cohort 1 has 105 students who graduated in 2012, and cohort 2 comprises 104 students who graduated in 2013. Table 1 and Table 2 provide an

LAK '15, March 16 - 20, 2015, Poughkeepsie, NY, USA
Copyright 2015 ACM 978-1-4503-3417-4/15/03†$15.00
http://dx.doi.org/10.1145/2723576.2723579

overview of the distribution of marks of the students over the four years for the two cohorts. The average marks of each student for each year have been calculated as a number from 0 (worst mark) to 100 (best mark). 50 is the minimum required mark to pass. Due to the strict placement process at the university followed by rigorous academic procedures, there is almost no trend of students failing the degree or dropping out.

Table 1. Distribution of the marks over the four years cohort 1

	90-100	80-90	70-80	60-70	50-60
year1	0	9	56	31	9
year2	0	13	55	25	12
year3	1	47	37	16	4
year4	6	62	26	11	0

It may be noticed from the data presented in tables 1&2 that marks over time tend to become better, though for cohort 2 in year 2 there is a shift from the interval 70-80 towards 60-70 and 50-60. It is also observed that cohort 2 has slightly better marks than cohort 1 in first year only. In subsequent years cohort 1 has better marks.

Table 2. Distribution of the marks over the four years cohort 2

	90-100	80-90	70-80	60-70	50-60
year1	0	14	55	29	6
year2	0	13	46	34	11
year3	0	30	48	22	4
year4	0	31	54	18	1

To discover typical performance patterns over the years, students have been clustered each year taking their final marks of each course in each year. The list of all courses over the four years is presented in Table 3. Except for a few exceptions, the number associated to the name of a course indicates the year in which the course is offered. For example CT-153 is taught in first year, while CS-251 is taught in second year and CT-354 in third year.

Clustering has been done using the X-Means and the K-means algorithms of the tool RapidMiner taking the Euclidean distance for both algorithms. The X-Means algorithm is a modification of K-means that includes an automatic estimation of the optimal number of clusters [6]. It is well known that the cohesion of each cluster increases with K, the number of clusters. However a big value of K gives a model or clustering which is more complex. X-means relies on the Bayesian Information Criterion (BIC) to find a trade-off between cohesion and complexity. It turns out that the sets of clusters obtained applying K-means and SSE (Sum of Squared Errors) to determine an optimal number of clusters on one hand, and applying X-means on the other hand, while not exactly the same, are not fundamentally different: they show the same trends in the data. Both algorithms provide clusters of students with low marks in all courses, intermediate marks in all courses and high marks in all courses. In this paper we give the results obtained with X-means.

Each cluster is characterized by a set of centroids, the mean of the marks obtained by the students in the cluster for each course. As will be explained with more details in the next section, the mean of all centroids is calculated and is used to describe each student by a 4-tuple, the cluster s/he belongs to in each year of the degree.

3. RESULTS

This section presents the results of X-means clustering that has been applied to the examination data of the students belonging to Cohort 1 in section 3.1, and Cohort 2 in section 3.2.

3.1 Cohort1

Tables 5 to 8 present the observed clusters that are depicted through their centroids. One can notice that the number of clusters found is relatively small, 3 clusters each year, except year 3 with 4 clusters. One notices also that each clustering in each year can be qualitatively described in exactly the same way: lowest marks in all courses, intermediate marks in all courses and highest marks in all courses. There is no cluster that represents low marks in course c1 and c2, and high marks in course c3, c4 and c5 for example.

Table 3. Attributes used for clustering

Name	Description
CT-153	Programming Languages
CT-157	Data Structures Algorithms and Applications
CT-158	Fundamentals of Information Technology
CT-161	Computing Lab
EE-115	Electrical Technology Fundamentals
EL-134	Basic Electronics
HS-102	English
HS-105/127	Pakistan Studies
MS-171	Differential & Integral Calculus
HS-205/206	Islamic Studies or Ethical Behavior
MS-121	Applied Physics
MS-172	Discrete Structures
CS-251	Logic Design and Switching Theory
CS-252	Computer Architecture and Organization
CT-251	Object Oriented Programming
CT-254	System Analysis and Design
CT-255	Assembly Language Programming
CT-257	Data Base Management System
EL-238	Digital Electronics
HS-208	Business Communication & Ethics
MS-271	Ordinary Differential Equation & Complex Variable
MS-272	Linear Algebra & Geometry
HS-207	Financial Accounting and Management
CT-352	Computer Graphics
CT-353	Operating Systems
CT-354	Software Engineering
CT-360	Visual Programming
CT-361	Artificial Intelligence & Expert System
CT-362	Web Engineering
CS-351	Computer Communication Networking
CS-352	Digital Communication Systems
CS-353	Microprocessor & their Applications
MS-331	Probability & Statistics
CT-452	Modeling & Simulation
CT-455	Distributed Database Client Server Programming
CT-456	Data Warehouse Methods
CT-460	Network & Information Security
MS-471	Applied Numerical Methods
CS-451	Parallel Processing
CT-454	Compiler Design
CT-461	E-Commerce
CT-481	Wireless Network & Mobile Computing

CT-483	System Administration

Table 4 represents the total number of students in each cluster for all the four years. It indicates that most of the students belong to the intermediate cluster in all the years except year 2 where the "High" cluster is particularly large.

Table 4. No. of students cluster wise in four years for Cohort 1

Cluster	First Year	Second Year	Third Year	Fourth Year
Low	32	15	14	23
Intermediate	47	25	27 (Intermediate - Low)	49
			32 (Intermediate - High)	
High	27	66	33	34

Table 5. Cluster centroids of first year for Cohort 1

Attribute	Low	Intermediate	High
CT-153	61.714	75.636	87.259
CT-157	66.257	80.636	86.259
CT-158	76.543	82.114	82.963
CT-162	58.371	71.614	82.741
EE-115	49.571	62.273	76.259
EL-134	51.686	68.977	76.778
HS-102	53.714	56.409	64.185
HS-105/127	60.114	61.955	68.444
HS-205/206	58.743	64.500	69.296
MS-121	60.400	73.136	82.778
MS-171	60.686	78.341	82.926

Table 6. Cluster centroids of second year for Cohort 1

Attribute	Low	Intermediate	High
CS-251	41.8	48.136	62.957
CS-252	46.6	59.909	72.493
CT-251	48.6	54.273	66.058
CT-254	65.133	73.455	77.130
CT-255	49.133	61.545	77.145
CT-257	60	72.818	80.333
EL-238	43.867	61.636	72.899
HS-207	59.067	65.500	77.638
HS-208	50.933	57.545	66.507
MS-271	47.133	51.455	69.957
MS-272	56.8	74.955	84.493

Table 7. Cluster centroids of third year for Cohort 1

Attribute	Low	Intermediate-Low	Intermediate-High	High
CS-351	56.857	71.667	78.094	84
CS-352	62	78.444	82.844	90.667
CS-353	47.071	59.704	68.031	83.182
CT-352	57.000	75.037	81.625	86.879
CT-353	53.857	65.778	74.969	81.091
CT-354	60.214	72.074	76.156	78.697
CT-360	50.929	55	67.469	78.364
CT-361	57.071	73.593	79.188	83.333
CT-362	52.571	59	73.969	85.394
MS-331	43.214	54.111	67.906	74.485

Table 8. Cluster centroids of fourth year for Cohort 1.

Attribute	Low	Intermediate	High
CS-451	61.440	76.383	86.353
CT-452	63.840	80.234	86.647
CT-454	69.320	77.830	87.559
CT-455	64.360	76.298	83.588
CT-456	62.640	75.064	79.765
CT-460	63.360	77.638	85.206
CT-461	69.520	78.489	87.676
CT-481/CT-483	76.760	79.766	83.118
MS-471	52.480	73.723	83.559

As mentioned earlier, the mean of all centroids of each cluster in each year is calculated and rounded as an integer. For example consider Table 5: The mean of cluster 'Low' is 60, the mean of cluster 'Intermediate' is 71 and the mean of cluster 'High' is 78. To obtain an intuitive overview of how the performance of students globally evolves over the four years qualitatively, each student is described by a 4-tuple whose elements are the means of the centroids of the clusters the student belongs to. For example a student belonging to the cluster with the lowest mark in first year and second year, with intermediate-low marks in third year and intermediate marks in fourth year will be represented by the tuple (60, 52, 66, 77) while a student belonging to the cluster with high marks in all four years will be described by the tuple: (78, 73, 83, 85).

Figure 1 aggregates all the tuples of all students. The height of the bar represents the number of students characterized by the same tuple. The diagram is ordered from right to left: low values on the right, and high values on the left.

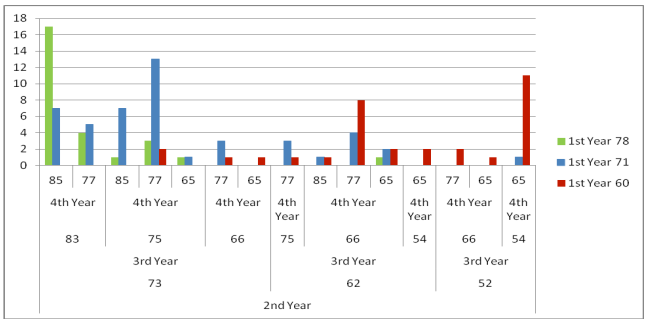

Figure 1. Tuples summary of Cohort1.

The color indicates the first year clusters. The 2nd year is depicted at the bottom of the diagram and divided into three parts: 52, 62 and 73 corresponding to the means of the clusters Low, Intermediate and High. Each of these three parts is divided by the clusters of 3rd year. Note for example, that part 52 of 2nd year is divided only by the parts 66 and 54 of 3rd year as there are no student with low marks in 2nd year and intermediate-high or high marks in 3rd year. Finally the highest level of the hierarchy divides further each part with the clusters of the 4th year.

The left most bar indicates the number of students, here 17, who have high marks in all four years and described by the tuple (78, 73, 83, 85). The right most bar gives the number of students, here 11, with low marks in all years and described by the tuple (60, 52, 54, 65). The second highest bar indicates that 13 students are characterized by the tuple (71, 73, 75, 77): except in second year these students were always in a group with intermediate marks.

The number of bars, here 28, gives the number of tuples that describes all students. Though the number of clusters each year is quite small, the number of combinations present in the data is not small anymore.

3.2 Cohort2

The clusters for cohort 2 are illustrated through their centroids in Tables 9 to 12. As observed for cohort 1, the number of clusters is also relatively small, 3 clusters each year, but for year 1, there are only 2 clusters. Its analysis show that the clustering follows the same trend as for cohort 1 i.e. low marks in all courses, intermediate marks in all courses and high mark in all courses.

Table 9. Cluster centroids of first year for Cohort 2

Attribute	Low	High
CT-153	58.143	74.945
CT-157	66.143	79.636
CT-158	71.041	81.855
CT-162	66.918	79.291
EE-115	50.306	69.945
EL-134	73.061	85.291
HS-102	63.041	67.091
HS-105/127	54.531	62.164
HS-205/206	61.122	72.236
MS-121	53.878	69.382
MS-171	48.776	64.764

Table 10. Cluster centroids of second year for Cohort 2.

Attribute	Low	Intermediate	High
CS-251	64	79.366	85.964
CS252	45.222	54.951	67.857
CT-251	62.267	69.78	75
CT-254	74.511	84.366	89.429
CT-255	49.378	58.707	77.036
CT-257	60.711	69.195	77.649
EL-238	49.178	67.854	78.429
HS-207	66.467	74.659	81.750
HS-208	62.533	67.78	71.786
MS-271	48.444	67.537	74.5
MS-272	57.578	73.976	82.750

Table 11. Cluster centroids of third year for Cohort 2

Attribute	Low	Intermediate	High
CS-351	60.082	72.265	83.548
CS-352	74.673	81.412	88.290
CS-353	48.857	57.912	72.677
CT-352	61.837	74.118	79.968
CT-353	59.571	66.059	75.387
CT-354	58.694	71.059	75.968
CT-360	59.245	68.176	79.258
CT-361	63.898	71.412	79.194
CT-362	60.224	69.588	80.387
MS-331	55.327	70.176	75.548

Table 12. Cluster centroids of fourth year for Cohort 2

Attribute	Low	Intermediate	High
CS-451	55.289	65.256	76.269
CT-452	70.267	82.349	89.808
CT-454	61.356	72.326	83.692
CT-455	69.578	75.465	82.423
CT-456	67.022	78.721	81.423
CT-460	63.6	77.791	86.885
CT-461	73.444	83.395	88.962
CT-481/CT-483	66.422	75.953	80.346
MS-471	57.489	75.116	85.192

Table 13 gives the total number of students in each cluster for all four years. Like cohort 1, most of the students belong to the intermediate cluster in all years except in first year; remember that first year does not have an intermediate cluster. This matches well Table 2, which has the highest number of students in the interval 70-80 in all four years.

Table 13. No. of students cluster wise in four years for Cohort 2

Cluster	First Year	Second Year	Third Year	Fourth Year
Low	49	20	18	34
Intermediate	-	41	43	44
High	55	43	43	26

As for cohort 1, Figure 10 shows all the tuples of all students of cohort 2.

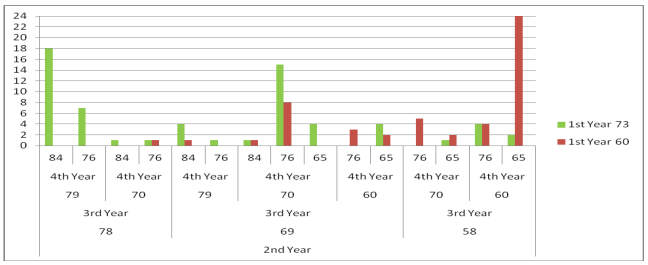

Figure 2. Tuples summary of Cohort2.

The left most bar indicates the number of students, here 18, who have high marks in all four years and described by the tuple (73, 78, 79, 84). The right most bar gives the number of students, here 24, with low marks in all years and described by the tuple (60, 58, 60, 65). The third highest bar indicates that 15 students are characterized by the tuple (73, 69, 70, 76): except in first year these students are always in a group with intermediate marks.

There are 23 bars; all students are described by 23 different tuples.

4. DISCUSSION

The clustering results of cohort 1 and cohort 2 confirm the shift towards better marks over the four years as observed in Tables 1 and 2. The data reveal that the mean of all clusters tends to increase. Consider the mean of the clusters with low marks. It takes the values 60, 52, 54 and 65 for cohort 1 and the values 60, 58, 60 and 65 for cohort 2. The smaller numbers 52 and 54 in cohort 1 have to be placed in relation with the small size of the clusters, as shown in Table 4. Similarly, the means of the clusters with high marks take the values 78, 73, 83, 85 for cohort 1 and 73, 78, 79, 84 for cohort 2. Figure 1 and 2 suggest that the scattering of performance in each year for the two cohorts stays stable as the differences between the mean of the cluster with low marks and the mean of the cluster with high marks do not change much over the years. This is further confirmed by the standard deviation (SD) of all courses in each year as the values of SD do not change much between the years.

The tuple summaries show two important groups of students: the top students, bar on the left, and the weak students, bar on the right. Top students have high marks in all four years and this group is equally important in both cohorts. Weak students have low marks in all four years and this group is slightly bigger for cohort 2 than for cohort 1.

Other significant groups for both cohorts are those constituted of students who have intermediate marks in all years except for one year. The second and fourth highest bars of Figure 1 show such groups for cohort 1: students with intermediate marks in all years but in 2nd year where they have high marks make up the second highest bar and students with intermediate marks but in 1st year, where they have low marks, make up the fourth highest bar. The third and fourth highest bars of Figure 2 are other examples of students having mainly intermediate marks for cohort 2.

More generally, the two diagrams Figure 1 and 2 shows that a significant number of students remain stable during their studies which means they tend to remain in the same kind of group.

A glaring observation is presence of more high bars towards the left of Figure 1, as compared to Figure 2. This indicates that there are more students with higher marks in year 2, and high or intermediate marks in year 3 and 4 in cohort 1 than in cohort 2, which visualizes the trend given by Tables 1 and 2.

Figure 1 and 2 also show that there are atypical students that are described as weak students in 1st year who progress in 2nd year and finish with intermediate or high marks in 4th year. They are found looking for red bars towards the left of the diagram. Another interesting but relatively small group is constituted by students who have high marks in 1st year but low marks in all subsequent years. They are represented by spotting the green or blue bars in the right of the diagram. There are 2 of them in cohort 2 and 1 in cohort 1. Because of the shift towards good marks, the progression of atypical students is more spectacular than the regression of the last 3 students.

5. CONCLUSION AND FUTURE WORK

This paper presents a work that shows how the performance of the students evolves over the years in their studies. Two consecutive cohorts have been analyzed by using X-means clustering algorithm.

As presented in the preceding discussion, a sizable number of students tend to stay in the same kind of group throughout four years in the two cohorts. There are very few atypical students: students who begin with low marks in 1st year and finish with high marks in 4th year, or the other way round: students who begin with high marks but from 2nd year earn only low marks. In both cohorts one observes a shift towards better marks over the years, though this shift is more visible for cohort 1 than for cohort 2. This shift towards better marks could be explained in relation with the university policy for calculating the final marks at the end of the degree with year-wise weighting as follows: 10% of first year examination marks, 20% of second year examination marks, 30% of third year examination marks and 40% of fourth year examination marks.

The work reported in [2] shows that it is possible to use one cohort to predict the performance of the subsequent cohort. Indeed, taking cohort 1 as training set, performance of cohort 2 has been predicted with an accuracy varying from 55.77% to 83.65% and a κ varying from 0.352 to 0.727 using different classifiers. These results are comparable to the ones obtained in other works like [4, 5, 7] that use cross validation, which means one cohort only, to validate prediction. The present research complements this work in showing that the two cohorts exhibit the same trends regarding performance evolution over the four years.

This study provides the direction for a future work to study the relationship between the results of progression with those of prediction of performance. Another future work may be to explore other visualizations to capture at a glance the progression of the performance of students over the years better. Finally an important future work is to take action. The findings exhibit a group of students with low marks all the way through their studies. They could be identified in first year already and receive special attention. Thus these findings have to be discussed with teachers and heads of department.

6. REFERENCES

[1] Asif, R., Merceron, A. and Pathan, M. 2014. Investigating Performances' Progress of Students. In Proceedings of the Workshop Learning Analytics, 12th e_Learning Conference of the German Computer Society (DeLFI 2014), Freiburg, Germany, September 15, 2014, 116-123.

[2] Asif, R., Merceron, A. and Pathan, M. 2014. Predicting student academic performance at degree level: a case study. In *International Journal of Intelligent Systems and Applications (IJISA)*, Vol. 7(1), 49-61. DOI: 10.5815/ijisa.2015.01.05.

[3] Bower, A.J. 2010. Analyzing the Longitudinal K-12 Grading Histories of Entire Cohorts of Students: Grades, Data Driven Decision Making, Dropping Out and Hierarchical Cluster Analysis. In *Practical Assessment, Research & Evaluation*, Vol. 15 (7), 1-18.

[4] Golding P., Donaldson O. 2006. Predicting Academic Performance. In Proceedings of 36th ASEE /IEEE Frontiers in Education Conference. DOI: 10.1109/FIE.2006.322661

[5] Kabakchieva D.2013. Predicting Student Performance by Using Data Mining Methods for Classification. In Cybernetics and Information Technologies, Vol. 13(1), 61-72.

[6] Pelleg D, Morre A. 2000. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In Proceedings of the Seventeenth International Conference on Machine Learning, San Francisco, CA, USA, 727-734.

[7] Zimmermann J., Brodersen K. H., Pellet J. P., August E., Buhmann J. M. 2011. Predicting graduate-level performance from undergraduate achievements. In Proceedings of the 4th International Conference on Educational Data Mining, Eindhoven, the Netherlands, 357-358.