Investigating Performances's Progress of Students

Raheela Asif[1], Agathe Merceron[2], Mahmood K. Pathan[1]

[1]NED University of Engineering and Technology, University road, Karachi 75270
[2]Beuth University of Applied Sciences, Luxemburger Strasse 10, 13353 Berlin
engr_raheela@yahoo.com
merceron@beuth-hochschule.de
mkpathan@neduet.edu.pk

Abstract: This paper investigates how performance of students progresses during their studies. Progression of a student is defined as a tuple that shows how a year average stays the same, increases or decreases compared to first year. Taking the data of two consecutive cohorts and using k-means clustering, five meaningful types of progressions are put in evidence and intuitively visualized with a deviation diagram. Interestingly, in both cohorts students globally progress or remain stable. Still, the two cohorts exhibit differences. A future work is to refine the present aggregative approach and to investigate dependencies between prediction and progression.

1 Introduction

While there are many works to predict performance of students, few works have been done to investigate how performance of students evolves during their studies. Are there a few typical progressions that give a good summary of how students evolve? Or on the contrary, are students so different and diverse that their behaviours cannot be summarized by a few typical progressions? Can progressions be calculated in a way that can be easily explained to teachers and intuitively visualized? This paper presents a preliminary case study in that area and shows that a simple methodology allows for discovering few typical progressions among the students of two follower cohorts of a four-year IT bachelor degree. While similar progression patterns can be found in the two cohorts, the distribution and the marks of students among the patterns are different. Interestingly the patterns found show that students almost stay the same or progress during their studies. There is no pattern showing students steadily regressing.

As mentioned above, the focus of many investigations is on prediction instead of on progression of performance. An analysis that bears strong similarities with the present work is [Bo10]. It 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. While this visualization does not exhibit typical progressions as we aim at doing in the present contribution, it does show distinctive groups, like students with good marks or low marks all the way through, students progressing from low marks to

better marks, or students with low marks who at some point drop off. The present work has similar findings. Though the work presented in [Ca12] analyzes the progression of students with respect to time, yet it has a connection with performance. Two curricula in higher education are considered and an ideal path is identified. Using k-means clustering, two groups are exhibited: a cluster contains students who tend to stick to the ideal path and get better graduation marks, and the other contain students who tend to study longer and get lower graduation marks. Though not limited to 2, the number of typical progressions we have found is also quite small.

The paper is organized as follows. The next section describes the data and methodology used. Section 3 presents the results and discusses them. Final remarks and future works are presented in the conclusion.

2 Data and Methodology

Students' marks of a four-year IT bachelor degree of the NED University of Technology, Karachi, have been used in this study. Two cohorts have been analysed: cohort 1 has 105 students who graduated in 2012 and cohort 2 has 104 students who graduated in 2013. Unlike at school, courses at university are different from one year or one semester to the next one and only few courses follow one another. Therefore there is no obvious way of detecting changes in performance by topics. In this exploratory study an aggregative approach is chosen that to some extend mimics academic practice: the average mark for each year is calculated and then transformed as follows: 90-100 → A mapped to 1, 80-90 → B mapped to 2, …, 50-60 → E mapped to 5. Students are thus described by four attributes, `Interval_year1, Interval_year2, Interval_year3` and `Interval-year4`. As an example consider four students s1 = (2, 2, 1, 1), s2 = (3, 3, 2, 1) , s3=(4, 4, 3, 3) and s4=(3, 4, 3, 3). Students s1 and s2 both finish with a high mark in year 4, but the second student progressed more than the first one, while s1 and s3 progressed exactly the same way but did not obtain the same marks. Fig. 1 gives an overview of the distribution of the marks over the four years for the two cohorts. One observes a shift towards better marks over the 4 years in both cohorts though the shift is more pronounced for cohort 1.

	A/1	**B/2**	**C/3**	**D/4**	**E/5**		**A/1**	**B/2**	**C/3**	**D/4**	**E/5**
Interval_year1	0	9	56	31	9		0	14	55	29	6
Interval_year2	0	13	55	25	12		0	13	46	34	11
Interval_year3	1	47	37	16	4		0	30	48	22	4
Interval_year4	6	62	26	11	0		0	31	54	18	1

Figure 1: Distribution of marks: left cohort 1, right cohort 2

To capture the progression of a student over the 4 years data are further transformed. An immediate transformation would be to simply measure the change between two consecutive years like `Interval_year(i) - Interval_year(i+1)`. We aim for a

transformation that captures better whether marks stay at the same level, decrease or increase. Therefore we define four attributes as follows: `year1 = 0, year2= Interval_year1 - Interval_year2, year3 = year2 + (Interval_year2 - Interval_year3)` and `year4 = year3 + (Interval_year3 - Interval_year4)`. It turns out that the formulas simplify to `year(i) = Interval_year1 - Interval_year(i)` for i=2 to 4. The four above students are described by the following progressions: p1=(0, 0, 1, 1), p2=(0, 0, 1, 2), p3=(0, 0, 1, 1) and p4=(0, -1, 0, 0). Progression (0, 0, 1, 1) for example shows the same average in first and second year, a progression by one interval in third year and year 4 stays in the same interval as year 3. Progression (0, -1, 0, 0) shows a student who obtains an average in the same interval all years except in year 2, where the average dropped by one interval.

A mere enumeration shows that there are more than 125 different possible progressions (and less than 625, the total number of possible combinations of yearly marks), but the data contain far less. For example the progression (0, 4, 4, 4) does not occur. This suggests that typical progressions using some clustering algorithm should be found. In this study k-means clustering has been performed with Euclidean distance using the tool RapidMiner. Because all the attributes have the same order of magnitude, data have not been normalized. Keeping the data as is makes the visualization of the clusters easier to interpret. All data have value 0 for `year1`, therefore this attribute has no influence on the clustering. It is left as it renders the visualization of the results more intuitive.

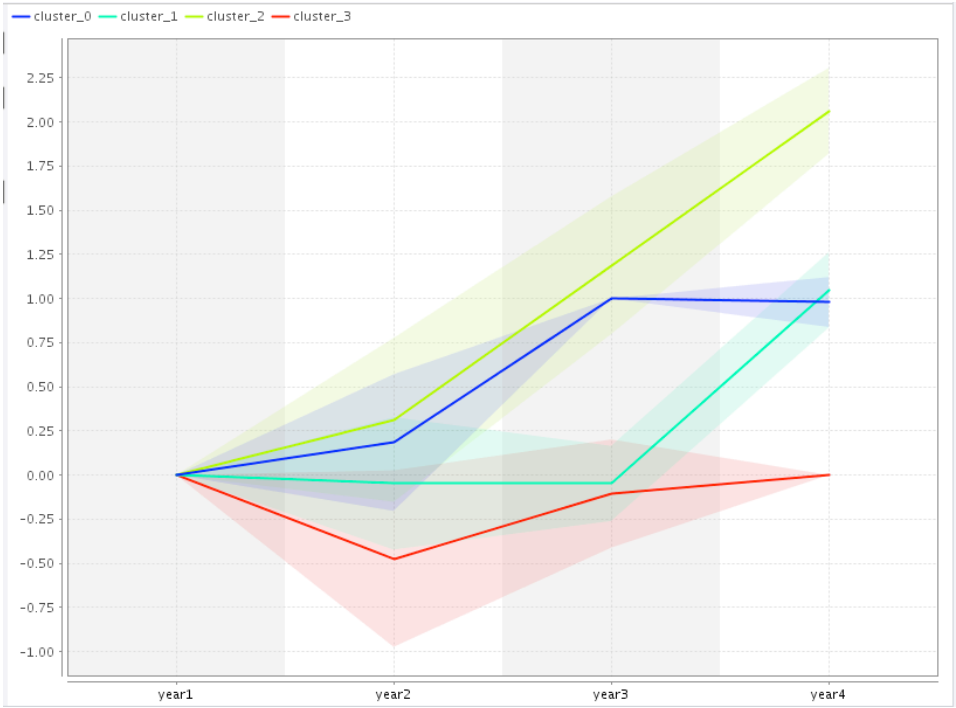

Figure 2: Progression's clusters of cohort 1 with k=4

3 Results

The value k = 4 for the number of clusters corresponds to the first sharp drop of the curve SSE (Sum of Squared Errors) against k for cohort 1 and gives interpretable clusters shown in Figure 2. The solid line shows the centroid and the shadow around shows the standard deviation of the cluster. Starting from the bottom of the diagram, cluster_3 gathers 19 students who tend to finish their degree with a mark in the same interval as in first year. We label it as "*almost stable*". Examples of progressions found in this cluster include (0, 0, 0, 0), the most common, (0, -1, 0, 0) or (0, 0, -1, 0). The next centroid line shows cluster_1 with 21 students who finish one interval higher than they began and the increase happens in year 4. We label it as "*up4*". Examples of progressions found in this cluster are (0, 0, 0, 1), the most common, (0, -1, 0, 1) or (0, 0, 0, 2). Cluster_0 has the next centroid line, contains 49 students who finish one interval higher than they began and the increase happened in year 3. It is labelled as "*up3*". Examples of progressions include (0, 0, 1, 1), the most common, (0, 1, 1, 1) or (0, 0, 1, 0). Finally the top line corresponds to cluster_2 with 16 students who progressed most. We label it as "*2-Inter-up*". Examples of progressions are (0, 0, 1, 2), the most common, (0, 1, 1, 2) or (0, 0, 2, 3). These results visualize and summarize the shift towards good marks observed in Figure 1.

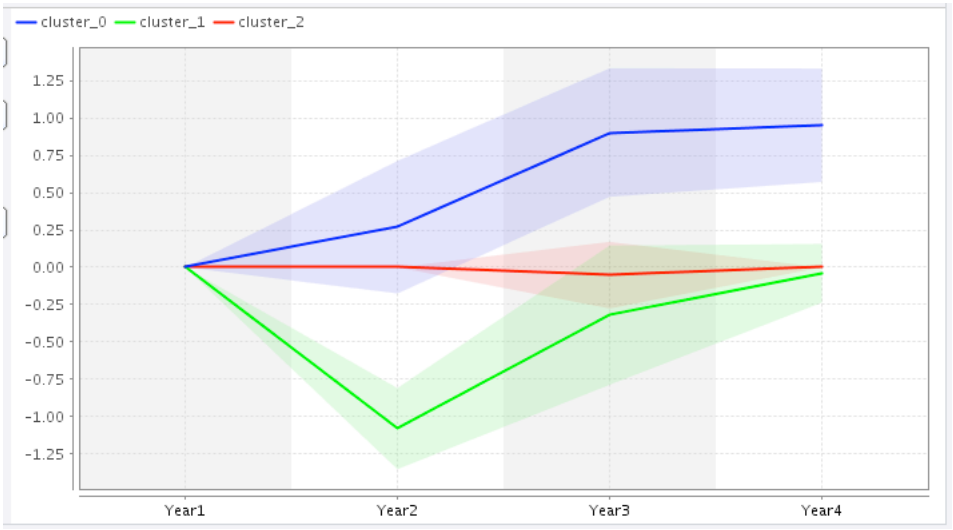

Figure 3: Progression's clusters of cohort 2 with k=3

Unlike cohort 1, k = 3 is the number of clusters that corresponds to the first sharp drop of the curve SSE (Sum of Squared Errors) against k for cohort 2 and gives interpretable clusters shown in Figure 3. As for cohort 1, the clusters do not exhibit a group that finishes with a lower mark than it started, though a few individuals do. Starting from the bottom of the diagram, cluster_1 gathers 25 students who tend to finish their degree with a mark in the same interval as in first year, from 0 to 0 but with lower marks in between.

We label it as "*down-up*". Such a cluster was not found in cohort 1 but included in the *almost stable* cluster. The next centroid line shows cluster_2 with 38 students, here the "*stable*" cluster. Cluster_0 has the next centroid line and contains 41 students who tend to finish one interval higher than they began. We label it as "*improvers*". Such a cluster corresponds to the aggregation of 3 clusters of Figure 2 for cohort 1. Comparing figures 2 and 3, one notices 19 students in cohort 1 in the *almost stable* cluster that match the 63 students of cohort 2 found in the *stable* and *down-up* clusters. Or, equivalently, 86 students of cohort 1 in the *up4, up3* and *2-Inter-up* clusters match the 38 students of the *improvers* cluster of cohort 2.

Bigger values of k simply refine the clustering shown in figures 2 and 3. They are useful to compare the two cohorts in more details. We show the clustering obtained for k=7 and cohort 1 as it is the most suitable for comparison with cohort 2. It returns 6 clusters only as shown in Figure 4. The clusters *up3* and *up4* remain exactly the same. The cluster *almost stable* of k=4 is split into 3 clusters: *stable*, 10 students, which gathers all students with progression (0, 0, 0, 0), the "*down year2*", 7 students with typical progression (0, -1, 0, 0) and the "*down year 2 and 3*", 2 students with progression (0, -1, -1, 0). The cluster *2-Inter-up* is split into two: the "*from year1 up*", 6 students with typical progression (0, 1, 1, 2), and the "*from year2 up*", 10 students with typical progression (0, 0, 1, 2).

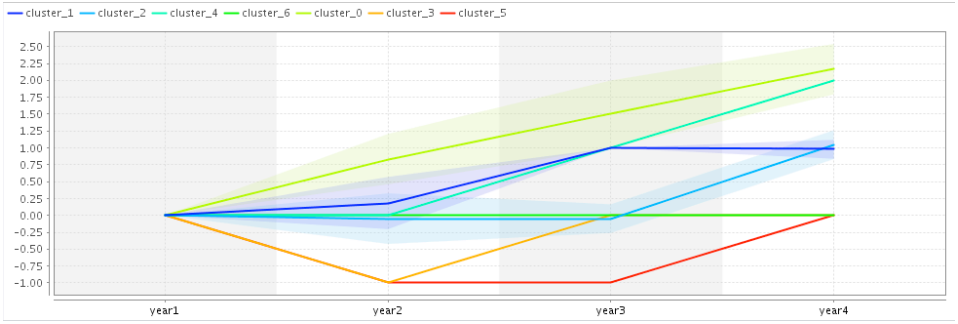

Figure 4: Progression's clusters of cohort 1 with k=7 giving 6 clusters

Proceeding in a similar way for cohort 2, a clustering with k=9 gives 6 clusters. The *stable* cluster remains the same. The *down-up* cluster, as for cohort 1, is split into two clusters: *down year2*, 17 students, and *down year 2 and 3*, 8 students. The *improvers* cluster is split into 3 clusters: *up4*, 6 students, *up3*, 24 students and the "*year2 improvers*" containing 11 students who tend to improve in year 2 already and slightly in year 3.

3.1 Progression and Performance

The refined clustering results of cohort 1 and cohort 2 show the following similarities: they both have the clusters *stable*, *up4*, *up3*, *down year2* and *down year 2 and 3*. Among the differences: cohort 1 contains a cluster *2-Inter-up* not found in cohort 2, and the *year2 improvers* cluster of cohort 2 is swallowed in the *up3* of cohort 1. Therefore, to

compare progression and performance of the two cohorts, we consider 5 clusters only: *2-Inter-up*, *year 2 improvers* and *up3* are merged as "*up2&3*", *up4*, *stable*, and *down-up*. The reason not to divide further *down-up* is the small cluster *down year 2 and 3* of 2 students for cohort 1, thus *down year 2 and 3* and *down year 2* are merged for both cohorts. The *year2 improvers* and *up3* are merged for cohort 2.

Figure 5 summarizes the different number of students per cluster in the two cohorts. Cohort 1 has more students in the clusters containing improvers: *2-Inter-up* (empty for cohort 2), *up2&3* and *up4,* which again reflects the trend of Figure 1 and the shift towards better marks. Cohort 2 has more students in the *down-up* cluster reflecting the increase and decrease of the D and E intervals of Figure 1.

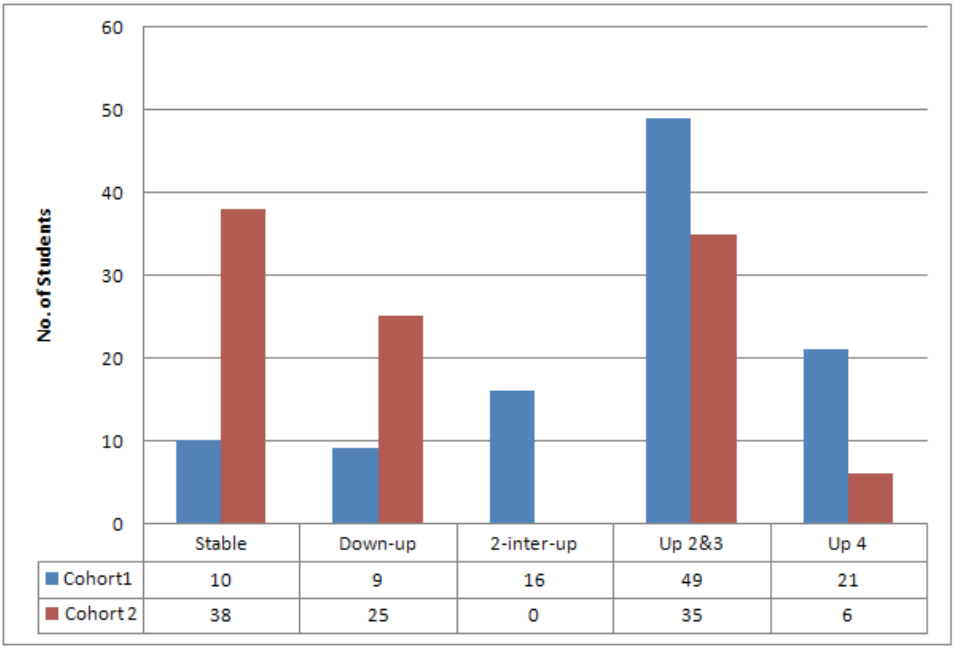

Figure 5: Number of students per cluster and cohort

Figure 6 shows how students distribute through the clusters according to their graduation mark. Altogether, there was only one student with graduation mark A, therefore it is omitted from the diagram. In each cluster, the proportion of students with graduation mark B to E is given by a pair of columns, the left column gives the frequency for cohort 1 and the right column for cohort 2. Interestingly the proportions of students with B, C or D graduation marks are quite similar for the two cohorts in the "*down-up*" cluster, though cohort 2 contains almost 3 times more students in that cluster than cohort 1.

Students of the *stable* and the *up4* clusters have the following in common: they have earned each year the same average mark, except in year 4 where students of the *up4* cluster progress by one interval. The *stable* cluster is quite small for cohort 1 and contains mainly B-students, means students with B as a graduation mark who cannot

improve that much. These students have earned good marks all the way through their studies. As a contrast, this cluster contains almost a third of the students of cohort 2, mainly C-students, who earned average marks each year during their 4 years of studies. The low achieving students of cohort 2 are all found in that cluster. Cohort 2 has few students in the *up4* cluster, while almost 20% of the students of cohort 1 are there, including all low achieving students. These two clusters show that low achieving students have been low achieving almost all the way through their studies. Taking into account the size, cluster *up4* contains mainly students with average or low marks, which is not the case for cluster *stable*. The two cohorts have a comparable number of students in the *up2&3* cluster. However the number of students of cohort 1 achieving a B graduation mark is sensibly higher than the number of those achieving a C mark, while these numbers are almost comparable for cohort 2.

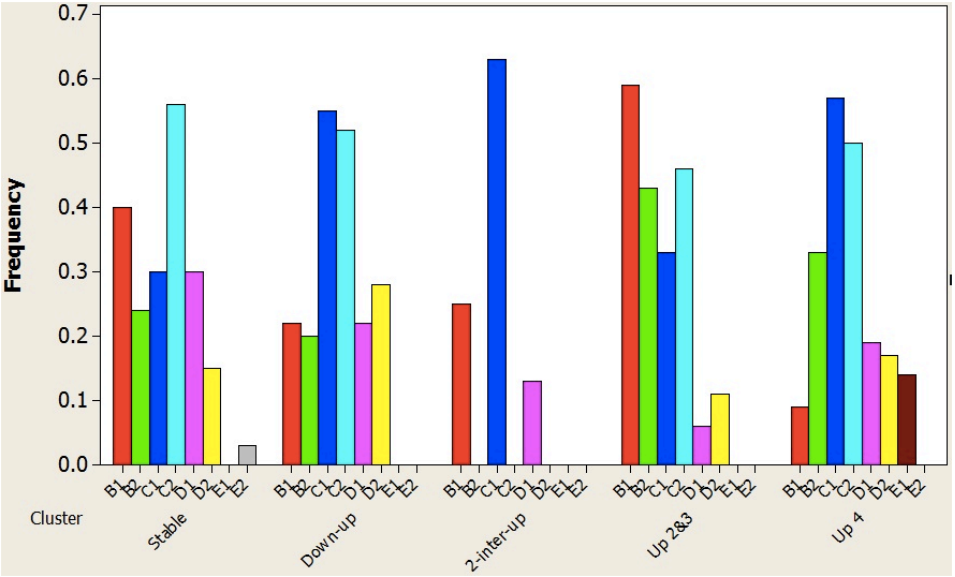

Figure 6: Clusters and graduation marks

4. Conclusion

This paper presents a first case study on performances's progression of students. Each student is represented by a 4-tuple that shows how his/her year average stays the same, increases or decreases compared to the preceding year. Using k-means clustering five typical progression are put in evidence. The two cohorts differ in the number of students and in the performance of students per type of progression: performance of students of cohort 1 increased much more than the one of cohort 2. Interestingly, these differences do not prevent of using cohort 1 to predict the performance of students of cohort 2 with a reasonable accuracy as the work in [AMP14] shows. Using only High School Certificates marks as well as 1st year and 2nd year marks, no demographical data, the interval of the final mark at the end of the 4 years degree has been predicted with

different classifiers obtaining an accuracy varying from 55.77% to 83.65% and a κ coefficient varying from 0.352 to 0.727. These results are comparable to those obtained by others using cross-validation, see for example example [GD06] or [Ka13]. It will be interesting to investigate further dependencies between prediction and progression of performance.

As already mentioned, there is no pattern showing students steadily regressing. This observation fits the way the graduation mark is calculated. Calculation is as follows: 10% average mark 1st year + 20% average mark 2nd year + 30% average mark 3rd year + 40% average mark 4th year.

Because we have chosen an aggregative approach, we obtain very synthetic progressions that give a bird-eye view. However two students having the same year average might have distinctive profiles: one may have in all courses the same mark and the other may have very good grades in some courses and low grades in others. The present approach does not allow to distinghuish them. As done in [Bo10], another approach consists in clustering students year by year taking the marks as they are. We have begun work is this direction. It shows interesting clusters that repeat each year for both cohorts: a cluster of students with low marks in all courses, a cluster with students with high marks in all courses, and clusters of students with intermediate marks whose number varies with k, the number of clusters. For cohort 1 clusters with intermediate marks tend to be ordered in the following sense: there is almost no cluster with good marks in some topics and low marks in others; students of one cluster will have better marks in all courses than students in another cluster. These patterns suggest that the present aggregative approach might be appropriate to get a general trend. Further work along these lines is in progress.

References

[AM14] Asif, A., Merceron, A., & Pathan, M.K. 2014. Predicting student academic performance at degree level: a case study. Submitted.

[Bo10] 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.

[Ca12] Campagni, R., Merlini, D., Sprugnoli, R. 2012. Analyzing paths in a student database. In Proceedings of the 5th International Conference on Educational Data Mining. (Chania, Greece, June 19-21). EDM'12. 208-209.

[GD06] Golding, P., & Donaldson, O. 2006. Predicting Academic Performance. In *Proceedings of 36th ASEE /IEEE Frontiers in Education Conference*.

[Ka13] Kabakchieva, D. 2013. Predicting Student Performance by Using Data Mining Methods for Classification. In *Cybernetics and Information Technologies*, 13(1), 61-72.