Do They Choose What They are Interested in: Analysis on Students' Interested Majors

Yixian Zhou
School of Information
University of Michigan
zyixian@umich.edu

December 12, 2019

Abstract

This project analyzed what engineering undergraduate students are interested in when they first come to college and what majors they graduate with. There are some interesting patterns among the distribution of students' interests in major. More than half of the students choose what they are interested in while other students change their choices. However, we cannot successfully predict their majors with only their initial interests.

1 Motivation

College is a place where students can explore what they truly passionate in with all sorts of resources. Undergraduate students in College of Engineering enroll without specifying their majors. They can take any courses that they find interesting in their first two years. Even though students are not required to declare their majors until their junior or even senior year, the school are still curious about their initial interested majors and how these preferences related to their final choices in majors.

This project analyzed students' interested majors when they first come to college and Data Description and their declared majors to answer the following research questions:

- Are there any patterns among students interests in engineering majors?
- Do students choose what they interested in?
- With only their major preferences, can we predict the majors they finally declared?

2 Data collection and preprocessing

The data is extracted from student survey and institutional database. There are 6274 graduated undergraduate engineering students who enrolled from 2009 to 2014 and 4576 current undergraduate students who enrolled from 2016 to 2019. Each student can choose as many interested majors as they want. The number of students' interested majors varies, but its distribution is normally distributed with a long right tail where half of the students chose three majors.

There were 15 Engineering majors in 2009. Data Science Engineering was introduced in around 2014 and Climate and Space Engineering were split into Space Science and Engineering and Climate and Meteorology. Only 1.2% of the students are majored in Data Science around half of which have a dual degree in Computer Science. 0.9% of the students have dual degrees. To keep the consistence among different datasets, we excluded students with dual degrees. Since most of the students registered after 2017 have not decided their degree, we kept data from 2009 to 2017, which resulted in the size reduce to 7020.

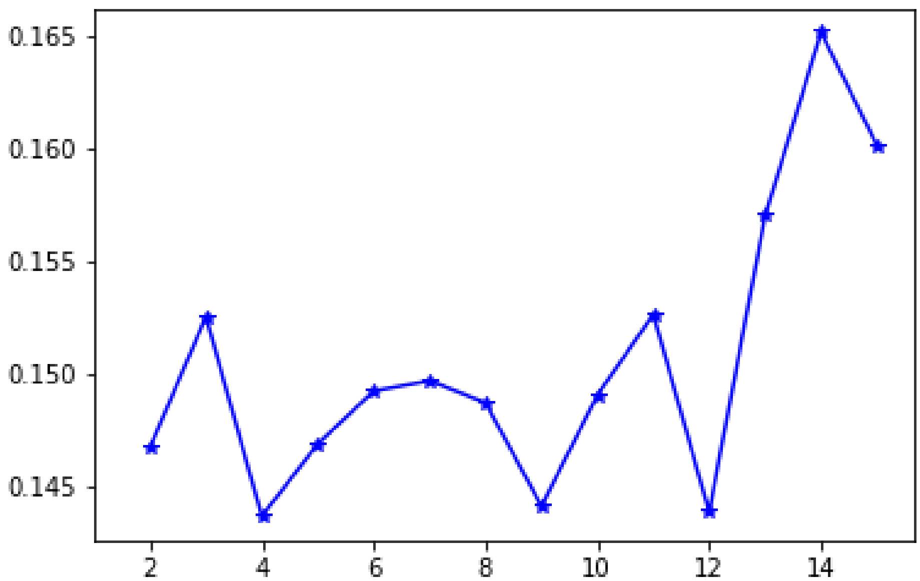

Figure 1: Silhouette scores of K-Means Clustering

3 Approaches

To answer the question on patterns of students' major preference, we implemented clustering algorithms to see if there are groups among students based on their initial interests. K-Means is a common method of clustering but requires to specify the number of clusters. The silhouette scores (Figure 1) and the within-cluster sum-of-squares criterion (Figure 2) do not suggest an appropriate amount of clusters, which indicates that K-Means might not be a good option. K-Means usually works well for data is nearly evenly distributed among clusters, especially when they are well-separated. The students being analyzed are related to each other since they all study in the realm of Engineering, so the groups are likely close to each other. The structure of the college suggests some hierarchy - majors are grouped into departments, and the college is built upon departments. Therefore, we finally choose Agglomerative Clustering. The result shows it can detect patterns among students' interest, which will be discussed later in the report.

The second question of whether they choose their initial interested major can be answered by some basic analysis and visualization methods.

Predicting students' choice in majors is a traditional classification problem. We first tried classifiers to predict what majors students graduate with but the accuracy score is at around 0.5. This might because the number of features is not large enough compared to the number of classes. Therfore, we switched the task to be classify whether students choose their originally interested majors, which is a binary classification problem. Since we only got around 7000 students data, Neural Network is not an ideal choice to build a classifier.

4 Experiments and Results

It is hard to quantitatively analyze the results of clustering because we do not have ground truth for unlabeled data. One way to identify the number of cluster is to investigate what majors students in each cluster are interested in. We ran Agglomerative Clustering and obtained different numbers of clusters by setting different cut-off of cluster distances. Figure 3 shows my interpretation for each clusters. The clustering results make the most sense when it has five clusters. Table 1 shows the abbreviations of majors.

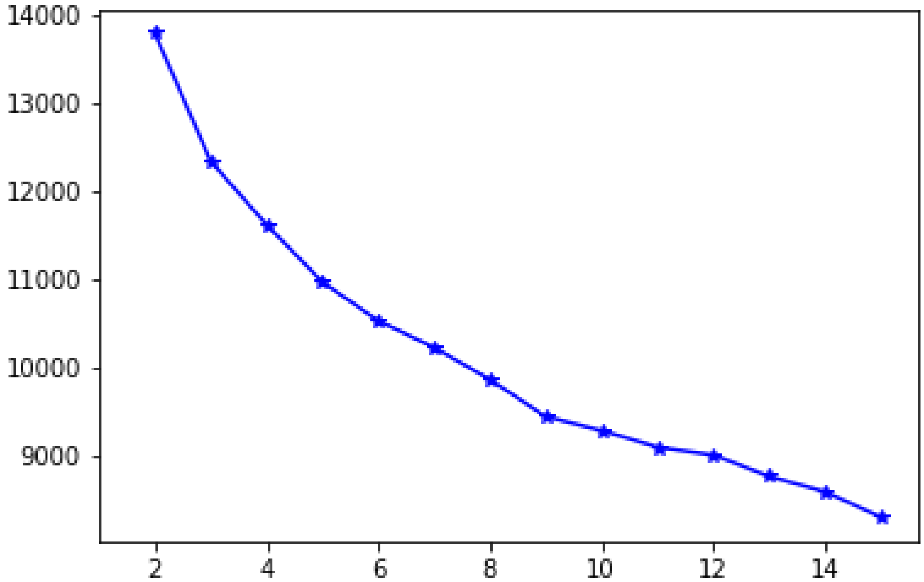

Figure 2: Within-cluster sum-of-squares of K-Means Clustering

Abbreviation	Major
AEROSP	Aerospace Engineering
BIOMEDE	Biomedical Engineering
CE	Computer Engineering
CHE	Chemical Engineering
CIVIL	Civil Engineering
CLASP	Climate and Space
CSE	Computer Science
DATASCI	Data Science
EE	Electrical Engineering
ENVIRON	Environmental Engineering
EPHYS	Engineering Physics
IOE	Industrial and Operations Engineering
MATSCIE	Materials Science and Engineering
MECHENG	Mechanical Engineering
NAME	Naval Architecture and Marine Engineering
NERS	Nuclear Engineering and Radiological Sciences

Table 1: Major abbreviations

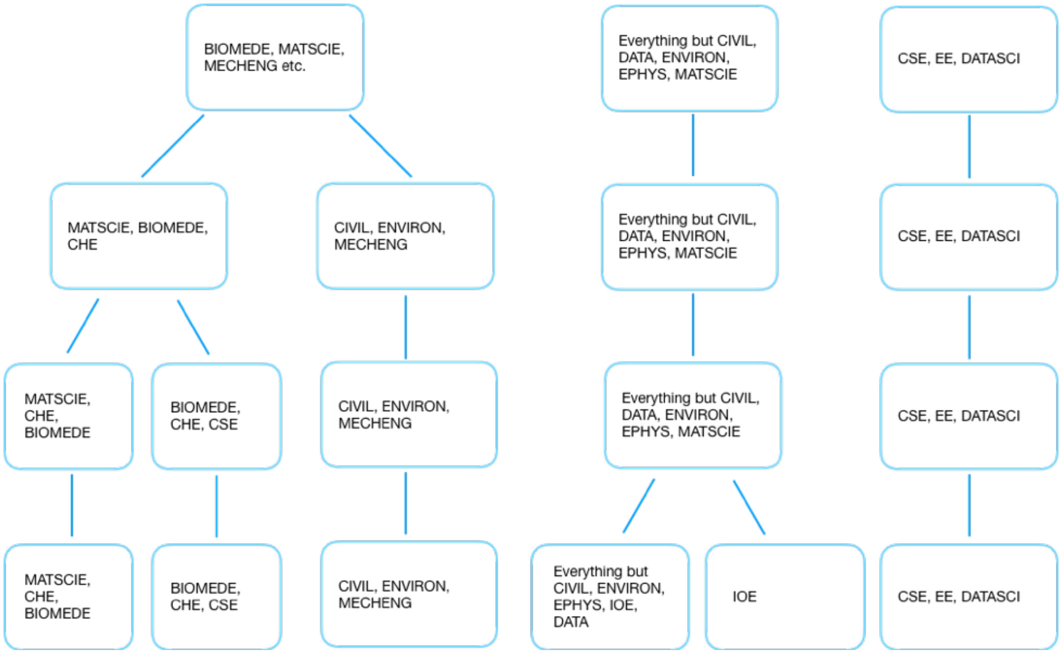

Figure 3: Results of Agglomerative clustering, we only show the results from 3 to 6 clusters here.

	precision	recall	f1-score	support
0	0.0	0.0	0.0	506
1	0.76	1.0	0.86	1600
accuracy	0.76	0.76	0.76	0
macro avg	0.38	0.5	0.43	2106
weighted avg	0.58	0.76	0.66	2106

Table 2: Classification report of Logistic Regression Classifier

For the classification task, we used a dummy classifier with stratified strategy as baseline.The weighted F1-score is 0.62. Besides students' preferences in majors, we also included the number of interested majors as we noticed that this indicated how strong students' interest in some majors are. We tried Logistic Regression Classifier and Support Vector Classifier. With grid search, we built models with best parameter. We found SVC with redial basis kernel performed better than Logistic Regression. Table 2 and 3 are classification report of these two classifiers.

5 Discussion

Figure 4 to 8 are the heatmaps of for each student clusters. Each cell is annonated by the percentage of students in the corresponding cluster. The x-axis represents majors and the y-axis represents the number of interested majors students choose. In Figure 6, the column of BIOMEDE shows that almost all of the students in cluster 4 are interested in BIOMEDE. For instance, 11.60% students who

	precision	recall	f1-score	support
0	0.64	0.1	0.17	506
1	0.78	0.98	0.87	1600
accuracy	0.77	0.77	0.77	0
macro avg	0.71	0.54	0.52	2106
weighted avg	0.74	0.77	0.7	2106

Table 3: Classification report of Support Vector Classifier

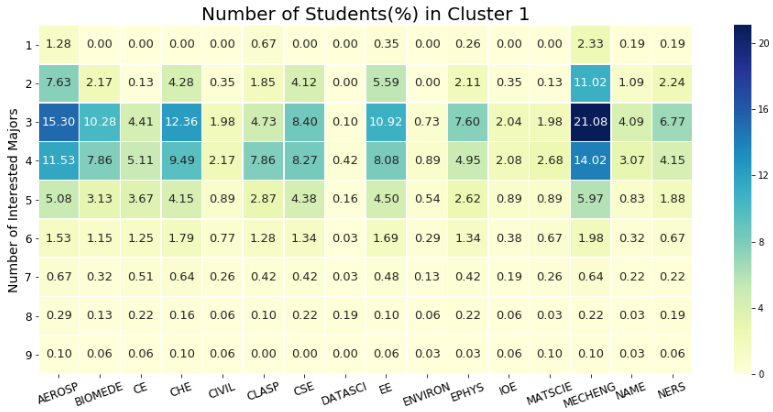

Figure 4: Cluster 1: Students interested in major in big departments

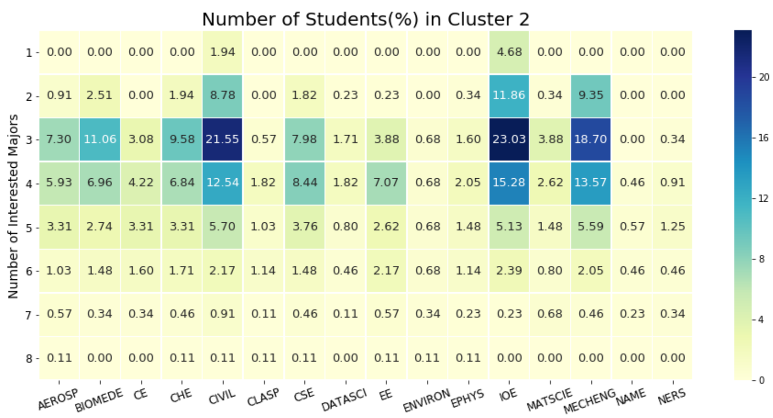

Figure 5: Cluster 2: Students interested in Civil and Environment Engineering with Industrial Operation Engineering

have one interested majors in this group have chosen BIOMEDE. As the number of interested majors increases, they pick other majors related to BIOMEDE or majors in big department such as CSE.

Figure 9 shows what students with a certain major are interested in when they start their college. Each cell represents the number of students who prefer the major in column but graduate with the major corresponded to y-axis. For instance, at the row of CE, we see 61.5% of students declare CE major with initial interest in CSE. It's interesting to see that most students had interests in Mechanical Engineering because this is a common perception about engineering. But many of them went to other majors. Also, students graduated with Industrial and Operations Engineering major had various interest when they first got to college.

The weighted F1-score of Logistic Regression Classifier is 0.66 while the one of SVC is 0.70. Both beat the baseline using dummy classifier. Table 2 shows that the Logistic Regression can not detect any students who choose majors different from their original preferences. However, SVC with a non-linear model did a better job in classifying negative class as shown in table 2. Such difference indicates the two classes are not linearly separable in the feature spaces.

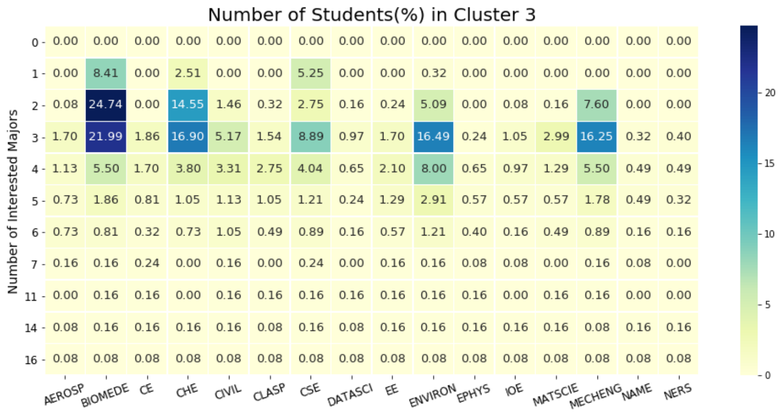

Figure 6: Cluster 3: Students interested in life science such as Biomedical Engineering

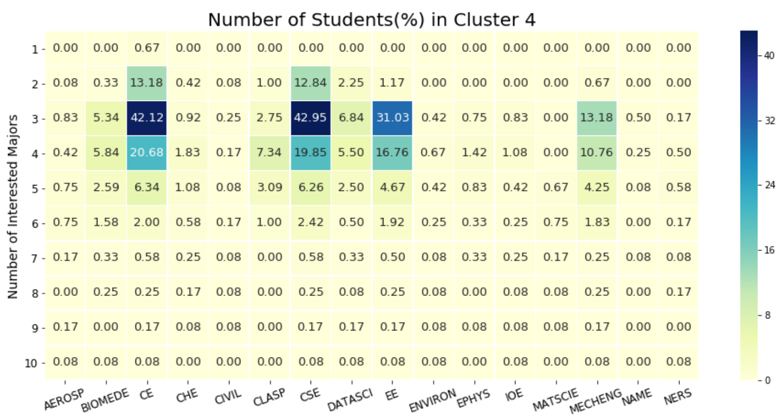

Figure 7: Cluster 4: Students interested in majors related to computer science

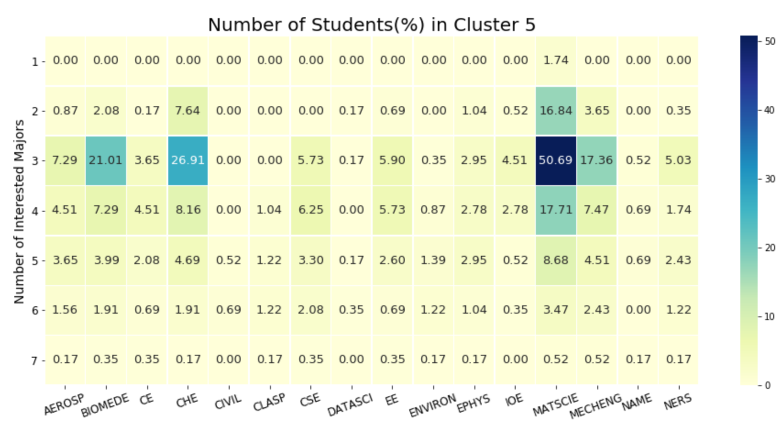

Figure 8: Cluster 5: Students interested in chemistry-related science such as Material Science Engineering

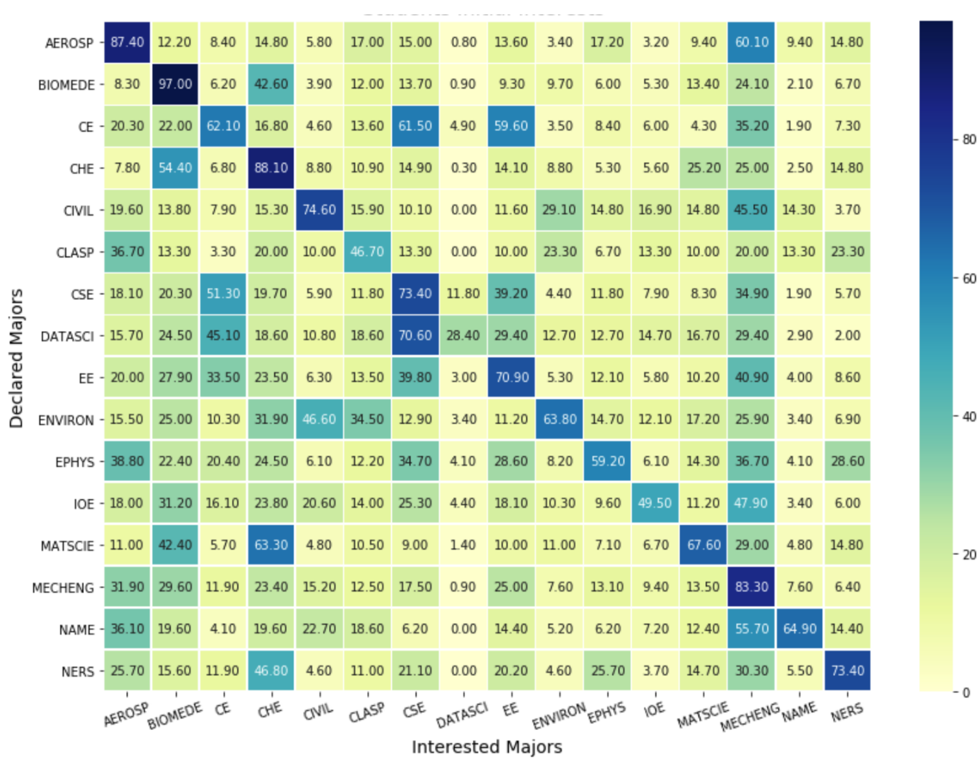

Figure 9: Declared majors vs. interested majos

6 Conclusion

This project analyzed engineering undergraduate students initial preferences of majors and what majors they declare when they graduate. Unsupervised learning can detect interesting patterns among the distribution of students' interests in major. It is common for students to change their mind and switch majors. Majors in large department such as CS and ME are always co-occur with students' interested major. We can predict whether they choose what they are interested in while we cannot successfully predict the specific major with only their initial interests.

7 Future Work

Network Analysis is another possible way to understand the changes in students' choices of majors. Noticing that some students' interests in majors remain the same while others' change to different fields, people can create a directed network with majors as nodes. Each directed edge represents students who change their choices in majors with weighting indicating the number of the students. For instance, if 10% students like Computer Science at the beginning but declare their majors as Industrial and Operations Engineering, the graph will have a directed edge from CS to IOE with weighing 0.1. People can get some insights from the analysis of such a network.

The database is still being updated every year. With more data, people can integrate the results of the clustering consider and implement semi-supervised learning to predict students declaration of majors.

Acknowledgments

We would like to thanks College of Engineering for providing the data for this project.

References

[1] Arthur, David, and Sergei Vassilvitskii. "k-means++: The advantages of careful seeding." Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, 2007.

[2] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[3] Smola, Alex J., and Bernhard Schölkopf. "A tutorial on support vector regression." Statistics and computing 14.3 (2004): 199-222.