Clustering Student Learning Activity Data

Haiyun Bian
hbian@mscd.edu
Department of Mathematical & Computer Sciences
Metropolitan State College of Denver

Abstract. We show a variety of ways to cluster student activity datasets using different clustering and subspace clustering algorithms. Our results suggest that each algorithm has its own strength and weakness, and can be used to find clusters of different properties.

1 Background Introduction

Many education datasets are by nature high dimensional. Finding coherent and compact clusters becomes difficult for this type of high dimensional data. Subspace clustering was proposed as a solution to this problem [1]. Subspace clustering searches for compact clusters embedded within subsets of features, and it has proven its effectiveness in domains that have high dimensional datasets similar to educational data. In this paper, we will show that different clustering and subspace clustering algorithms produce clusters of different properties, and all these clusters help the instructor assess their course outcomes from various perspectives.

2 Clustering Student Activity Data

We assume that datasets are in the following format: each row represents one student record, and each column measures one activity that students participate in. Our test data contains 30 students with 16 activities, and 7 students failed this class. The final grade is the weighted average from the scores in all 16 activities.

2.1 Student clusters

Student clusters consist of groups of students who demonstrate similar learning curves throughout the whole course. These clusters are helpful to identify key activities that differentiate successful students from those who fail the course. We applied the SimpleKMeans from Weka [2] to the test dataset with k being set to 2. The results show that cluster1 contains 6 out of 7 students who failed the course, and cluster2 contains 24 students among whom 23 passed the course. One student who failed was clustered into cluster2, and we found out that this student’s composite final score is 58%, which lies right on the boundary of passing/failing threshold.

2.2 Activity clusters

Here we focus on finding groups of activities in which all students demonstrate similar performance. For example, we may find a group of activities where all students show worse than average performance. This suggests that the instructor may want to spend more time on these activities to cope with the difficulty. To find this type of clusters, we need to transpose the original data matrix. We applied SimpleKMeans to the transposed test dataset. We tried different k values and found out the best clustering results was obtained at k =4 by looking at the curve of within group variance as a function of k. Among these four clusters of activities, cluster4 is the most challenging group of activities because its cluster centroid is consistently lower than the other three clusters.

2.3 Subspace clusters

We only report the results from PROCLUS [3] algorithm due to limited space. PROCLUS needs to set the number of clusters (k) and the average subspace dimensionality (l).

```
SC_0: [0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 ] #13 {2 5 7 8 10 13 14 15 17 21 23 27 29 }
SC_1: [0 0 1 0 0 0 0 1 1 1 0 0 0 0 0 0 ] #11 {0 1 4 12 16 19 20 24 25 26 28 }
SC_2: [0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 ]  #4  {3 6 9 22 }
SC_3: [0 1 0 0 0 0 0 1 1 1 0 0 0 0 0 0 ]  #2  {11 18 }
```

For k=4 and l=3, the four clusters identified are shown as above. The first subspace cluster (SC_0) lies in a subspace that contains two features: activity 8 and activity 12. SC_0 contains 13 students, and they are: stu2, stu5, stu7, and etc. A simple investigation shows that SC_2 and SC_3 contain all students who failed the class. SC_2 suggest that 4 out of 6 students who failed this class had difficulty with activities 6 and 7, and SC_3 shows that the other two students who failed had difficulty with activities 2, 8, 9 and 10. We can also see that SC_1 and SC_3 are two clusters that are best contrasted by activities 6, 7 and 8. Since all students in SC_1 passed the course while SC_3 students failed the course, these three activities may be crucial for students to pass the course.

3 Conclusions

All three types of clusters provide us with different perspectives. Since not all students experience the same difficulty in all activities, subspace clustering seems to be well suited for this purpose.

References

[1] Aggarwal C. C., Wolf J. L., Yu P. S., Procopiuc C., and Park J. S. Fast Algorithms for Projected Clustering, *Proc. of the 1999 ACM SIGMOD international conference on Management of data*, 1999, p. 61-72

[2] Hall M., Frank E., Holmes G., Pfahringer G., Reutemann P., and Witten I. H. The WEKA Data Mining Software: An Update, *SIGKDD Explorations*, Volume 11, Issue 1, 2009

[3] Müller E., Günnemann S., Assent I., Seidl T. Evaluating Clustering in Subspace Projections of High Dimensional Data, *Proc. 35th International Conference on Very Large Data Base*, 2009