A Preliminary Study on Clustering Student Learning Data

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

Abstract

Clustering techniques have been used on educational data to find groups of students who demonstrate similar learning patterns. Many educational data are relatively small in the sense that they contain less than a thousand student records. At the same time, each student may participate in dozens of activities, and this means that these datasets are high dimensional. Finding meaningful clusters from these datasets challenges traditional clustering algorithms. In this paper, we show a variety of ways to cluster student grade sheets using various clustering and subspace clustering algorithms. Our preliminary results suggest that each algorithm has its own strength and weakness, and can be used to find clusters of different properties. We also show that subspace clustering is well suited to identify meaningful patterns embedded in such data sets.

Introduction

Traditional data mining projects deal with datasets containing thousands or millions of records. Educational data mining tends to deal with much smaller datasets, normally in the range of several hundred student records. Even though a course may be offered multiple times, it is difficult to merge all these data records because every offering of the same course may involve different set of activities. On the other hand, many educational datasets are by nature high dimensional. For example, students' learning in a course may be assessed by utilizing an aggregation of several assignments, quizzes and tests. It is typical to have more than a dozen activities that contribute to a student's final grade. The log data from online teaching courses contain even more features describing the activities participated by each individual student.

Clustering is a very useful technique that finds groups of students demonstrating similar performance patterns. Since the number of students for each dataset rarely goes beyond a thousand and the number of features tends to be comparable to the number of students, finding coherent and compact clusters becomes difficult for this type of data. It is difficult because the pair-wise distance between students using the full-dimensional space becomes indistinguishable when the number of features becomes high. This problem is described as the curse of dimensionality, and it makes traditional clustering algorithms, such as k-means, unsuitable to be directly applied to high dimensional datasets.

Subspace clustering was proposed as a solution to this problem (Agrawal *et al.* 1998). Subspace clustering algorithms search for compact clusters embedded within subsets of features, and they have shown their effectiveness in domains that have high dimensional datasets similar to educational data. One specific example is its application to microarray data analysis. Microarray datasets tend to have similar sizes as educational datasets, mostly in the range of several hundred instances (genes or students) and several hundred features (samples or activities). Subspace clustering algorithms find subsets of genes that show similar expression levels under subsets of samples (Cheng *et al.* 2000; Madeira *et al*. 2004).

In this paper, we present some preliminary results from applying a variety of different clustering techniques, including subspace clustering, to student grade sheets. We show that clustering this type of datasets can provide the instructor a tool to predict who are likely to fail the course at very early stage as well as a possible explanation why they are failing.

Related Research

Over the last decade, many data mining techniques have been applied to educational data (Bravo *et al*. 2009; Dekker *et al.* 2009; Merceron *et al.* 2009). Research has shown that some techniques are more suitable for educational data than others, mainly because of the inherent characteristics of the datasets in this domain. For example, support and confidence, the two commonly used interestingness measurements for association rules, are not suitable for pruning off association rules when applied to educational data (Merceron *et al.* 2009). Instead, the authors have found that cosine and added value (or equivalently lift) are better measurements for educational data. One possible reason is that educational data have much smaller number of instances than the market basket data. Therefore, support and confidence tend to fall short in catching the real value of a good association rule in educational context.

Subspace clustering was first introduced to cluster students skill sets in (Nugent *et al.* 2009). The authors proposed to start with a "value-hunting" scanning for each individual feature to find out all features that contain meaningful and well-separated single-dimensional clusters. Those features that contain no good clusters were disregarded from further consideration. Then using all remaining features, conventional clustering algorithms such as hierarchical clustering and k-means were applied to identify clusters

resided in higher-dimensional spaces. In their research, subspace is used to prune off uninteresting features before the actual clustering process starts, and it is very similar to a feature selection procedure.

In general, a subspace cluster is a group of similar instances within their own subset of features. After the first subspace clustering algorithm for data mining was proposed (Agrawal *et al.* 1998), many different algorithms have been proposed. These algorithms can be classified into two categories: partition based approaches (Agrawal *et al.* 1999; Agrawal *et al.* 2000) and grid based approaches (or density-based approaches) (Agrawal *et al.* 1998; Cheng *et al.* 2000; Kriegel *et al.* 2009).

Partition-based algorithms partition all instances into mutually exclusive groups. Each group, as well as the subset of features where this group of instances show the greatest similarity is reported as a subspace cluster. Similar to k-means, most algorithms in this category define an objective function to guide the search. The major difference between these algorithms and the k-means algorithm is that the objective functions of subspace clustering algorithms are related to the subspaces where each cluster resides in. Notice that in subspace clustering, the search is not only on a partition on the instance set, but also on subspaces for each instance group. For example, PROCLUS (Agrawal *et al.* 1999) is a variation of the k-medoid algorithm. In PROCLUS, the number of clusters k and the average number of dimensions of clusters are specified before the running of the algorithm. This algorithm also assumes that one instance can be assigned to at most one subspace cluster or classified as an outlier, while a feature can belong to multiple clusters. Unlike PROCLUS that finds only axis-parallel subspace clusters, ORCLUS (Agrawal *et al.* 2000) finds clusters in arbitrarily oriented subspaces.

Grid-based (density-based) algorithms consider the data matrix as a high dimensional grid, and the clustering process is a search for dense regions in the grid. In CLIQUE (Agrawal *et al.* 1998), each dimension is partitioned into intervals of equal-length, and an n-dimensional unit is the intersection of intervals from n distinct dimensions. An instance is contained in a unit if the values of all its features fall in the intervals of the unit for all dimensions of the unit. A unit is dense if the fraction of the total instances contained in it exceeds an input parameter δ. CLIQUE starts the search for dense units from single dimensions. Candidate of n-dimensional dense units are generated using the downward closure property: if a unit is dense in k dimensions, all its k-1 dimensional projection units must all be dense. This downward closure property dramatically reduces the search space. Since the number of candidate dense units grows exponentially in the highest dimensionality of the dense units, this algorithm becomes very inefficient when there are clusters in subspaces of high dimensionality. Research has been done to extend CLIQUE by using adaptive units instead of rigid grids (Kriegel *et al.* 2009), as well as to use other parameters such as entropy in addition to density to prune away uninteresting subspaces (Cheng *et al.* 2000).

Clustering Student Grade Sheets

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 the course. An example is shown in Table 1, where d_{ij} denotes the i_{th} student's performance score in the j_{th} activity. Most clustering and subspace clustering algorithms allow d_{ij} to take real values.

	Activity 1	……	Activity m
Stu 1	d_{11}	……	d_{1m}
Stu 2	d_{21}	……	d_{2m}
……	……	……	……
Stu n	d_{n1}	……	d_{nm}

Table 1 Dataset Format

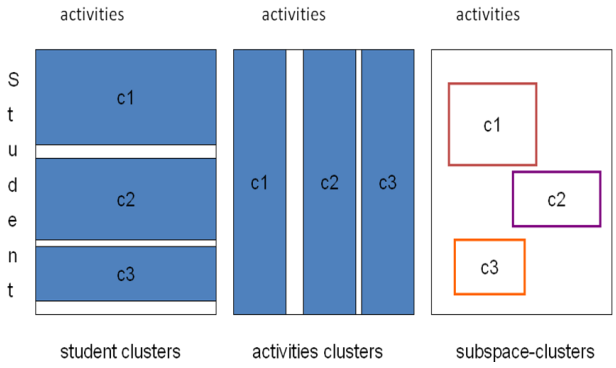

Figure 1. Clusters on Student Activity Data

Figure 1 shows three different clusters and subspace clusters that can be identified from the above data using different clustering and subspace clustering algorithms. Properties of each type of cluster as well as the process to find it will be presented in the following subsections.

The example dataset

We will use the grade sheet for a computer science service course as the example for this study. This dataset contains 30 students and 16 activities plus the final grade. The score of each activity as well as the final grade are in the range of [0, 1]. All students whose final composite grade is below .6 (60%) are marked as failing the course. In this dataset, 7 out of 30 students are marked as failed using this standard. Activities 1 through 12 are weekly in class labs in chronological order. Activities 13 and 14 are two large

projects due at mid semester and the end of the semester. Activities 15 and 16 are mid-term and the final examinations.

Each individual feature shows positive covariance with the final grade variable. Several features are highly correlated to the final grade, such as activities 6, 8, 9, 10 and 11. The least predictive features include activities 1, 13 and 15. It is not surprising for us to see that the first lab (activity 1) is not a good indicator of a student's performance in the course. But an interesting observation is that the mid-term exam (activity 15) and the mid-term project (activity 13) are both as bad as the first lab to tell whether a student will pass the course or not.

Another interesting observation is that activity 6 (lab 6) alone can predict with 100% accuracy about whether a student will pass the course or not. We later found out that the topic that was in that week was loops, which is considered challenging for most students. This suggest that if a student can grasp the concept of loop structure very well, he might as well be able to pass the course as a whole. Therefore, it would be worthwhile for the instructor to spend more time and effort on this subject matter.

Feature	Covariance	Feature	Covariance
Activity 1	.5044	**Activity 9**	**.9075**
Activity 2	.7758	**Activity 10**	**.9089**
Activity 3	.6097	**Activity 11**	**.9067**
Activity 4	.7415	Activity 12	.8137
Activity 5	.7125	Activity 13	.5283
Activity 6	**.9787**	Activity 14	.8427
Activity 7	.8670	Activity 15	.5613
Activity 8	**.9065**	Activity 16	.8046

Student clusters

Here we focus on identifying groups of students who demonstrate similar performances throughout the whole course. This type of clusters can be useful for the instructor to identify key activities that differentiate successful students from those who fail the course.

We have tried a wide variety of clustering algorithms' available from Weka (Weka URL) on the example dataset, and the results show that the simple k-means algorithm achieves at least comparable results as other more complicated algorithms in almost all cases.

Using the simple k-means algorithm, we started with k=2, that is, to find two clusters (Cluster0 and Cluster1) from this dataset. Cluster0 contains 6 out of 7 students who actually failed the course, and cluster1 contains 24 students among whom 23 are marked as passing the course. There is one failing student who is clustered into cluster1. We found out that this student's composite final score is .58, which lies right on the boundary of passing/failing threshold. This suggests that choosing 0.6 as the passing/failing threshold seems rather arbitrary.

Figure 2 shows the centroids of the two clusters. We can see that some activities are better in differentiating the two clusters than others, such as activities 6, 8, 9, 10 and 11. This result is consistent with the result from individual feature's covariance with the final grade variable, suggesting that the clusters that were identified from the algorithm may have captured some real characteristics of the dataset.

We have also tried k=3 to find three clusters from this dataset. It resulted in cutting the failing cluster (cluster0) into two even smaller clusters, leaving cluster1 remain unchanged.

Features	Full Data(30)	cluster0(6)	cluster1(24)
Activity1	0.76	0.475	0.8313
Activity2	0.7533	0.3208	0.8615
Activity3	0.7908	0.3333	0.9052
Activity4	0.785	0.3333	0.8979
Activity5	0.815	0.3792	0.924
Activity6	0.7767	0.0708	0.9531
Activity7	0.79	0.3083	0.9104
Activity8	0.7983	0	0.9979
Activity9	0.7683	0	0.9604
Activity10	0.7308	0	0.9135
Activity11	0.7833	0.1667	0.9375
Activity12	0.7667	0.1667	0.9167
Activity13	0.7647	0.5767	0.8117
Activity14	0.667	0	0.8338
Activity15	0.7693	0.7844	0.7656
Activity16	0.5674	0.2278	0.6523

Figure 2. Centroids of Student Clusters (K = 2)

Activity clusters

In this section we take a different view on the same dataset. Here we focus on finding groups of activities in which all students demonstrate similar performance patterns. For example, we may find a group of activities on which all students demonstrate consistently high performance. This suggests that these activities involve relatively easy-to-grasp concepts. We may also find a group of activities where all students show worst 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 would need to transpose the original data matrix as shown in Table 1 into Table 2.

	Student 1	……	Student n

Activity 1	d11	……	dn1
Activity 2	d12	……	dn2
……	……	……	……
Activity m	d1m	……	dnm

Table 2. Transposed Dataset

Similar as above, we applied the SimpleKMeans from Weka to the transposed example dataset. We tested five k values in the range of 2 to 6, and we tried to find the best value of k by looking at curve of within-group-variance as a function of k. The result is shown in Figure 3. As we can see, four clusters seems to be the best because the slope of the curve reduced significantly after k=4.

Among four clusters of activities, cluster3 is the most challenging group of activities because its cluster centroid is consistently lower than the other three clusters. Cluster3 contains three activities including activity 13, activity 15 and activity 16. Out of 30 students, 13 students showed significant lower than average performance on these three activities. An interesting observation is that in previous sections we have pointed out that activities 13 and 15 are also considered as insignificant in differentiating passing students from failing ones. This mean that these two activities maybe too hard to be used as criteria to predict the student overall performance of the course. On the other hand, activity 1, which is also considered as a bad feature to tell the difference between passing students from those who failed, might be too easy to be used as a criterion for that purpose.

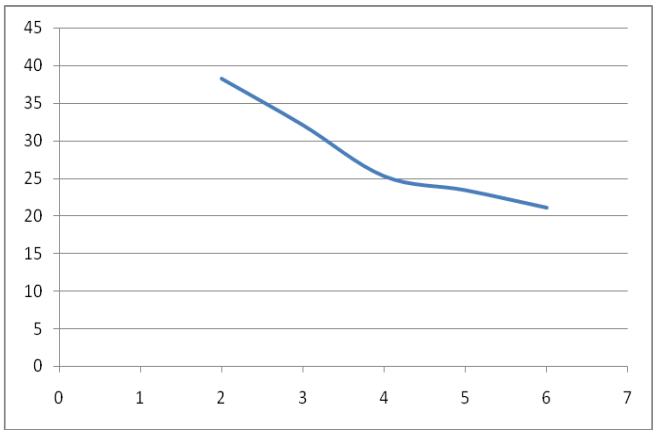

Figure 3. Within Group Variances

Subspace clusters

In this section, we show that subspace clustering algorithms can be used to find clusters embedded in subspaces. In earlier section, we have shown that student clusters contain groups of students who demonstrate similar performance throughout the whole course. Here we relax the constraint to allow any groups of students who demonstrate similar learning patterns in any subsets of activities to become candidates for clusters.

We will first show the results from partition-based subspace clustering algorithm. We chose to use PROCLUS because it reports clusters in axis-parallel subspaces, which makes the final interpretation of the clusters easier. The PROCLUS implementation is from the open source subspace clustering package (OpenCluster URL).

Similar to K-means, PROCLUS needs a pre-determined number of clusters (k) before running. In addition, it also requires knowing the average subspace dimensionality (l). We set k=4, and tried several values of l between 2 and 5, and found out that the results are highly similar for all cases. For the example dataset, PROCLUS finds the following four clusters when we set k = 4 and l = 3:

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 }

Each line describes one subspace cluster. For example, 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 failing students. In SC_2, 4 out of 6 students who fail this class have difficulty in doing activity 6 and activity 7, and SC_3 shows that the other two failing students showed difficulty in doing activity 2 and activities 8, 9 and 10. We later found out the activity 6 was a lab on loop structure and labs 8 and 9 are labs on Classes and Objects. This suggests that the majority of the students who failed this course started to fall behind when loops were introduced. The other half who failed the class failed to catch up when the concept of objected oriented programming were introduced. Therefore, the instructor may want to spend extra time to help students complete these three activities.

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.

We have also tried partition-based subspace clustering algorithm on the sample data. Grid-based algorithms produce more than a thousand subspace clusters, and the large number of reported clusters makes the interpretation of clusters very difficult. We will look into the possibility to prune off insignificant clusters based on domain knowledge. Similar research has been done in bio-medical

data analysis, where domain knowledge is used to measure the significance of each bi-cluster.

Comparisons between the three

Student clusters represent groups of students showing similar performance patterns throughout the whole course, while subspace clusters shows clusters of students who demonstrate similar performances in subsets of activities. Activity clusters is helpful in finding out difficult tasks for all students, while subspace clusters can identify subsets of activities that challenge different groups of students. Since not all students experience the same difficulty in all activities, subspace clustering seems to be well suited for this purpose. We can see from the example data that activity 6 may be a good feature to tell why some students failed this course, but it is not the only indicator. SC_3 suggests that there are some students who had no problem finishing activity 6 but still failed the course due to their unsatisfactory performance in activities 8, 9 and 10.

Conclusions and Future Work

This paper is our first attempt to adopt a rich collection of subspace clustering algorithms on educational data. Our preliminary results show that clustering and subspace clustering techniques can be used on high dimensional education data to find out interesting student learning patterns. These cluster patterns are helpful for the instructor to gain insights into the different learning behaviors and adapt the course to accommodate various students' needs. We will test and validate all presented clustering schemes on more educational data of larger size. We will also look into the possibility of applying grid-based subspace clustering algorithms to educational data guided by domain knowledge.

References

Aggarwal C. C., Wolf J. L., Yu P. S., Procopiuc C., and Park J. S. Fast Algorithms for Projected Clustering, *Proceedings of the 1999 ACM SIGMOD international conference on Management of data (SIGMOD'99),* pp. 61-72, 1999

Aggarwal C. C., and Yu P. S. Finding Generalized Projected Clusters in High Dimensional Spaces, *Proceedings of the 2000 ACM SIGMOD international conference on Management of data (SIGMOD'00),* pp. 70-81, 2000

Agrawal R., Gehrke J., Gunopulos D., and Raghavan P. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications, *Proceedings of the 1998 ACM SIGMOD international conference on Management of data (SIGMOD'98)*, pp. 94-105, 1998

Bravo J., Ortigosa A. Detecting of Low Performance Using Production Rules, *Proceedings of the Second International Conference on Educational Data Mining*, 2009. p. 31-40

Cheng C. H., Fu A. W.-C., and Zhang Y. Entropy-based Subspace Clustering for Mining Numerical Data, *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'99),* pp. 84-93, 1999

Cheng Y. and Church G. M. Biclustering of Expression Data, *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology (ISMB'00),* pp. 93-103, 2000

Dekker G. W., Pechenizkiy M., and Vleeshouwers J. M. Predicting Students Drop Out: A Case Study, *Proceedings of the Second International Conference on Educational Data Mining*, 2009. p. 41-50

Kriegel H. P., Kroger P., and Zimek A. Clustering High-dimensional data: A Survey on Subspace Clustering, *Pattern-based Clustering, and Correlation Clustering, Transactions on Knowledge Discovery from Data (TKDD),* Vol. 3, 2009

Madeira S., Oliveira A. Biclustering Algorithms for Biological Data Analysis: a survey*, IEEE Transactions on Computational Biology and Bioinformatics*, vol. 1, pp. 24-24, 2004

Merceron A., and Yacef K. Interestingness Measures for Association Rules in Educational Data, *Proceedings of the Second International Conference on Educational Data Mining*, 2009. p. 57-68

Nugent R. Ayers, E., Dean N. Conditional Subspace Clustering of Skill Mastery: Identifying Skills that Separate Students, *Proceedings of the Second International Conference on Educational Data Mining*, 2009. p. 101-110

OpenClusters:http://dme.rwthachen.de/OpenSubspace/

Weka: http:// www.cs.waikato.ac.nz/ml/weka