Proceedings of the International Conference on Industrial Engineering and Operations Management
Bogota, Colombia, October 25-26, 2017

Detect Multiple Choice Exam Cheating Pattern by Applying Multivariate Statistics

Mason Chen
Stanford Online High School, Palo Alto, CA, USA
Mason.chen.training@gmail.com

Abstract

This aim of this project is to apply a series of pattern detection Data Mining algorithms to accurately identify during classroom test exams. To detect if a pattern could be identified on the answer keys between students not attributable to chance alone, multivariate statistics tools were used to determine whether there was any association pattern among the students. Hierarchical Clustering and Dendrogram Tree were used to identify the grouping affinity behavior related to exam cheating pattern. Authors also used Heat Map to identify and recognize patterns in exam scores using visual analysis. The authors also selected the top 20% of questions considered the most difficult ones in order to increase the detection power. The probability of picking the same wrong answers on the difficult questions are even more unlikely by chance alone as compared to picking the right answers for the easy questions. It is statistically even more improbable that students would unintentionally select the same wrong answers on difficult questions, and therefore provides very evidence of cheating. Principle component analysis was also used to identify pairs of students who cheated, with. The predictive model approach using Data Mining tools was very powerful for analysis of the complex exam cheating patterns.

Keywords

Data Mining, Heat Map, Clustering Analysis, Dendrogram Tree, Principle Component Analysis, JMP

1. Introduction

For each instructor, designing an effective assessment exam is a critical job[1,2]. While a written exam of comprehensive questions and free-form answers may demonstrate critical thinking and depth/breadth of knowledge, this exam takes as significant amount of time to grade, and grading may often be subjective. Multiple-choice exams are more common due to their ease of quick and objective assessment by graders, despite limitations in demonstrating breadth of knowledge. Unfortunately, students may try to cheat by copying or checking answers each other during these exams, especially if they are seated very close to each other (as is often the case due to space limitations, such as in public schools). The inevitable possibility of cheating under these circumstances, presents a dilemma for most instructors, challenging them on resourceful design their exams to minimize cheating risk. This paper uses Data Mining tools and techniques (using JMP 12 Software) to detect patterns in multiple-choice responses among pairs of students that are indicative of cheating. The demonstrated effectiveness of this 'cheating detection' approach warns students proactively, by cautioning them to avoid attempting cheating before taking any multiple-choice exam.

In this case study, there were 75 students who sat in 25 different small tables (with 3 students per table) in a very limited classroom space. The instructor modified the original exam into three different orders (versions A, B, C). Three students from the same table would each take different versions (one student per version, per table). Students could not use cell phones or laptops during the assessment exam to prevent communication with one another. However, students were still smart enough to attempt to synchronize the questions in each of the versions, thusly providing evidence of cheating as shown in the analysis that follows. The objective of this paper is to implement a data mining algorithm to detect any cheating pattern from students taking the exam at the same table.

2. Data Collection and Multivariate Correlation Analysis

The raw data includes each student's ID, Exam Version, Answers, and Table Number. In order to reduce the computing time and also improve data quality (signal-to-noise ratio), the lowest 25 multiple-choice exam scores were excluded from the analysis. Also, it is highly unlikely that we can locate any evidence of cheating from the worst-performing students. Seating location was randomly

assigned for each student (per table). Therefore, it was statistically unlikely that most of the worst performers were sitting at the same table during the exam. In addition, there is little to know behavioral incentive for low-performing students to attempt to cheat off of their low-performing peers.

2.1 Multivariate Correlation Analysis

Firstly, JMP 12 Multivariate Correlation Analysis[3] was used to study the presence of correlation (as determined by calculated pairwise correlation coefficients) between the top 50 students scores, with results presented per Table 1. JMP's Multivariate platform was used to explore how many students' scores relate to each other. The word multivariate simply means involving many variables (each Student Scores here) instead of analysis of only one (univariate) or two (bivariate) variables. From the Multivariate report, you can:

- Summarize the strength of the linear relationships on Exam Score between each pair of Student IDs, using the Correlations table
- Identify dependencies, outliers, and clusters using the Scatterplot Matrix
- Use other techniques to examine multiple variables, such as: partial, inverse, and pairwise correlations, covariance matrices, and principal components.

Table 1. Multivariate Correlation Analysis

Correlations

	SID 1	SID 2	SID 3	SID 4	SID 5	SID 6	SID 7	SID 8	SID 9	SID 10	SID 11	SID 12	SID 13	SID 14	SID 15	SID 16	SID 17	SID 18	SID19
SID 1	1.0000	0.1317	0.0722	0.3094	0.5046	0.0346	-0.0341	-0.0201	0.0683	0.2237	0.4059	0.5854	0.3615	0.1723	0.2730	0.4670	0.4318	-0.0545	0.1531
SID 2	0.1317	1.0000	0.6838	0.2372	0.0994	0.3807	-0.0644	-0.0924	0.1905	0.2613	-0.1676	0.0366	-0.1688	0.4298	-0.2409	-0.0304	0.2846	0.4206	0.2509
SID 3	0.0722	0.6838	1.0000	0.3461	-0.0682	0.3333	0.0162	0.0703	0.3875	0.3541	-0.2509	-0.0076	-0.0659	0.5857	0.1518	0.0626	0.4833	0.7771	0.2305
SID 4	0.3094	0.2372	0.3461	1.0000	0.3335	0.4200	-0.1649	0.3603	-0.1576	0.2870	0.4522	0.3797	0.5021	0.5937	0.1135	0.5085	0.5347	0.2473	0.4994
SID 5	0.5046	0.0994	-0.0662	0.3335	1.0000	0.2555	-0.0432	-0.2693	-0.1897	0.3375	0.3247	0.3890	0.3583	-0.0074	-0.0490	0.5342	0.1484	-0.0690	0.0925
SID 6	0.0346	0.3807	0.3333	0.4200	0.2555	1.0000	0.0941	-0.1529	-0.0148	0.7471	0.2331	0.0538	0.0721	0.2284	0.0690	0.4589	0.5915	0.1369	0.5557
SID 7	-0.0341	-0.0644	0.0162	-0.1649	-0.0432	0.0941	1.0000	-0.2284	-0.0585	0.2635	-0.0956	-0.1426	-0.3954	0.0135	0.5452	-0.0923	0.0923	0.0279	0.1091
SID 8	-0.0201	-0.0924	0.0703	0.3603	-0.2693	-0.1529	-0.2284	1.0000	0.0039	-0.1338	0.3379	-0.0956	0.4808	0.5856	-0.0834	-0.0681	0.1603	0.1948	-0.0642
SID 9	0.0683	0.1905	0.3875	-0.1576	-0.1897	-0.0148	-0.0585	0.0039	1.0000	-0.0433	-0.4976	0.0481	0.0970	0.0787	0.2832	-0.0295	-0.0227	0.0452	0.1717
SID 10	0.2237	0.2613	0.3541	0.2870	0.3375	0.7471	0.2635	-0.1338	-0.0433	1.0000	0.3723	0.2030	0.1090	0.2408	0.1149	0.6042	0.4727	0.2434	0.5320
SID 11	0.4059	-0.1676	-0.2509	0.4522	0.3247	0.2331	-0.0956	0.3379	-0.4976	0.3723	1.0000	0.1939	0.4574	0.1759	-0.0423	0.4077	0.3514	-0.0062	0.2229
SID 12	0.5854	0.0366	-0.0076	0.3797	0.3890	0.0538	-0.1426	-0.0956	0.0481	0.2030	0.1939	1.0000	0.0612	0.0900	0.2064	0.6298	0.0351	-0.1645	0.3474
SID 13	0.3615	-0.1688	-0.0659	0.5021	0.3583	0.0721	-0.3954	0.4808	0.0970	0.1090	0.4574	0.0613	1.0000	0.2202	-0.0343	0.4119	0.3854	-0.0390	0.2360
SID 14	0.1723	0.4298	0.5857	0.5937	-0.0074	0.2284	-0.0135	0.5856	0.0787	0.2408	0.1750	0.0900	0.2202	1.0000	0.1078	0.1134	0.4722	0.6290	0.2185
SID 15	0.2730	-0.2409	0.1518	0.1135	-0.0490	0.0690	0.5452	-0.0834	0.2832	0.1149	-0.0423	0.2064	-0.0343	0.1078	1.0000	0.3658	0.3264	-0.0350	0.1608
SID 16	0.4670	-0.0304	0.0626	0.5085	0.5342	0.4589	-0.0923	-0.0681	-0.0295	0.6042	0.4077	0.6298	0.4119	0.1134	0.3658	1.0000	0.4509	-0.1489	0.5248
SID 17	0.4318	0.2846	0.4833	0.5347	0.1484	0.5915	0.0933	0.1603	-0.0227	0.4727	0.3514	0.0351	0.3854	0.4722	0.3264	0.4509	1.0000	0.3631	0.2309
SID 18	-0.0545	0.4206	0.7771	0.2473	-0.0690	0.1369	0.0279	0.1948	0.0452	0.2434	-0.0062	-0.1645	-0.0390	0.6290	-0.0350	-0.1489	0.3631	1.0000	-0.0215
SID19	0.1531	0.2509	0.2305	0.4994	0.0925	0.5557	-0.1091	-0.0642	0.1717	0.5320	0.2229	0.3474	0.2360	0.2185	0.1608	0.5346	0.2309	-0.0215	1.0000
SID 20	0.1170	0.1006	0.2199	0.3322	0.4241	0.2287	0.2234	-0.0018	-0.0789	0.1840	0.1712	-0.0585	0.2821	0.3764	0.1835	0.1644	0.4007	0.3449	-0.0135
SID 21	0.4178	0.1256	0.1195	0.6784	0.2864	0.0803	-0.0995	0.4840	-0.0870	0.1885	0.4344	0.4181	0.3390	0.6315	0.0872	0.4166	0.2324	0.0847	0.2468
SID 22	0.2893	0.2923	0.2738	0.2525	0.2093	0.5980	-0.1234	0.0000	0.0480	0.7074	0.2874	0.3494	0.1294	0.2938	0.1959	0.6919	0.4448	0.0788	0.5280
SID 23	0.2571	-0.1267	0.1612	0.0851	0.1323	0.0092	0.4324	-0.0136	-0.0424	-0.0227	-0.1170	0.0333	-0.0606	0.1317	0.5463	0.1031	0.2618	0.0601	-0.0318
SID 24	0.0994	0.4828	0.5752	0.4112	0.0325	0.4549	0.3653	0.2745	0.0521	0.5633	0.1672	-0.1508	0.0535	0.6362	0.1500	0.1457	0.4715	0.3956	0.3174
SID 25	0.3516	0.1024	0.2763	0.4220	0.0838	-0.0967	-0.2408	0.3668	-0.1417	0.1510	0.5260	0.1639	-0.0637	0.4019	-0.0637	0.2053	0.2498	0.3863	0.2114
SID 26	0.1968	0.3912	0.4592	0.5603	0.1410	0.4195	-0.2088	0.4588	0.2263	0.3517	0.2915	0.1687	0.3662	0.7017	0.0572	0.3264	0.3885	0.3450	0.5831
SID 27	0.4382	-0.0026	0.2120	0.3394	0.0661	0.1266	0.5622	-0.0019	-0.0433	0.3149	0.1167	0.2483	0.1090	0.2916	0.6230	0.3478	0.4727	0.1208	0.2580
SID 28	0.2238	0.4038	0.3061	0.3856	0.4291	0.3624	-0.1429	-0.1404	0.0310	0.4668	0.3285	0.2523	0.1464	0.3462	-0.2097	0.2376	0.1298	0.3819	0.3875
SID 29	0.2301	-0.0413	-0.0625	0.4440	0.2912	0.2541	-0.4270	0.2715	0.2481	0.2376	0.2848	0.2065	0.7870	0.0909	-0.0411	0.5787	0.2249	-0.2680	0.5113
SID 30	0.2913	0.1194	0.0190	0.2695	-0.1331	-0.3511	0.1372	0.1273	-0.0640	-0.2400	0.1490	0.3588	-0.0126	0.2716	0.3390	0.0309	0.0165	0.0049	0.1284
SID 31	0.1731	0.5350	0.5752	0.3662	0.1724	0.3318	-0.1117	0.0477	0.0521	0.3530	0.0573	0.1609	0.2074	0.4178	-0.1120	0.1457	0.4715	0.6064	0.1996
SID 32	0.3806	0.1309	0.0733	0.2773	0.4309	0.2876	0.0712	-0.0286	0.0973	0.0897	0.1527	0.1126	0.1239	0.2296	0.2462	0.3132	0.4071	-0.1032	-0.0242
SID 33	0.1902	0.4116	0.6769	0.4748	-0.0906	0.3022	0.3673	0.2601	0.2020	0.3915	0.0630	0.0155	0.0848	0.6955	0.3680	0.0408	0.4511	0.5981	0.3687
SID 34	0.2499	-0.0830	0.1839	0.4163	0.3714	0.0590	-0.2883	0.0291	0.0022	0.2225	0.2125	0.6822	0.1580	0.2534	0.2108	0.6904	0.1426	0.1930	0.2418
SID 35	0.2521	0.4117	0.4833	0.5347	0.1484	0.4415	-0.2197	0.3907	0.2382	0.4214	0.3067	0.1301	0.3854	0.6852	0.1134	0.3972	0.4626	0.3202	0.6136
SID 36	0.4768	0.2846	0.4336	0.2602	0.1484	0.4415	-0.0408	0.2525	0.2904	0.4727	0.2621	0.2726	0.1509	0.5787	0.3264	0.3972	0.5164	0.3202	0.3744
SID 37	0.0612	0.2523	0.5630	0.3773	0.2047	0.3535	0.1097	0.1524	0.0271	0.3008	0.0536	-0.1693	0.0927	0.6196	0.0552	0.0871	0.3937	0.5566	0.1954
SID 38	0.4361	0.2440	0.2945	0.3258	0.0417	0.1310	-0.0765	0.2979	0.1530	0.1718	0.2042	0.2056	0.1948	0.4999	0.3122	0.1688	0.3058	0.1644	0.4709
SID 39	0.3918	0.3159	0.2584	0.3655	0.0314	0.1736	0.0441	-0.0218	0.0741	-0.0364	-0.0458	0.0749	0.0518	0.3318	0.2478	0.0191	0.3222	0.0962	0.2963
SID 40	0.1374	0.3457	0.4799	0.1718	-0.1052	0.2952	0.0143	0.2130	0.1299	0.2297	-0.0762	-0.1233	-0.0967	0.5622	0.0038	-0.0764	0.3840	0.3380	0.0085
SID 41	0.4956	0.1999	0.5451	0.3655	0.1201	0.0194	-0.0308	0.2362	0.1281	0.1657	0.2673	0.0982	0.2909	0.4243	0.3142	0.3125	0.6459	0.4741	0.0185
SID 42	0.2666	0.1228	0.0673	0.4597	-0.0617	-0.1628	0.0030	0.4843	-0.0991	-0.1912	0.2992	0.1755	-0.2273	0.6173	0.2960	0.0164	0.2250	0.1161	0.0691
SID 43	0.3056	0.1740	0.3399	0.1585	0.1617	0.4782	0.0834	0.3371	0.4648	0.1755	0.3703	-0.0121	0.3879	0.4243	0.4861	0.3125	0.1623	0.4081	
SID 44	0.1968	0.3912	0.4592	0.5603	0.1410	0.4195	-0.2088	0.4588	0.2263	0.3517	0.2915	0.1687	0.3662	0.7017	0.0572	0.3264	0.3885	0.3450	0.5831
SID 45	0.2613	0.3039	0.2907	0.9745	0.3308	0.3771	-0.1907	0.3117	-0.1798	0.2408	0.3962	0.3723	0.4525	0.5780	0.0023	0.4329	0.4190	0.2048	0.5029
SID 46	0.4182	-0.2151	-0.0840	0.2322	0.3850	-0.0423	0.1513	-0.0390	0.0883	-0.0434	0.1123	0.3616	0.1984	0.0991	0.5856	0.3637	0.0455	-0.2174	0.2429
SID 47	0.5368	0.1777	0.2752	0.4184	0.0308	0.1331	0.0565	0.2722	0.2117	0.0555	0.2396	0.1156	0.3912	0.4513	0.4873	0.3083	0.6978	0.1107	0.1769
SID 48	0.0760	-0.0538	0.1024	0.6503	0.0962	0.1270	0.1513	0.4289	-0.3531	0.0868	0.4533	0.0894	0.2777	0.4504	0.0000	0.1361	0.2728	0.3623	0.1514
SID 49	0.7182	0.1369	0.0043	0.5487	0.5146	0.2672	-0.1638	0.2362	-0.1304	0.2673	0.5328	0.4276	0.4303	0.4771	-0.0023	0.4722	0.4861	0.0074	0.1608
SID 50	0.7182	0.1369	0.0043	0.5487	0.5146	0.2672	-0.1638	0.2362	-0.1304	0.2673	0.5328	0.4276	0.4303	0.4771	-0.0023	0.4722	0.4861	0.0074	0.1608
Reference	0.4061	0.3766	0.4588	0.7731	0.2941	0.0505	-0.2252	0.1840	-0.0702	0.0841	0.2927	0.4850	0.3035	0.4006	-0.0403	0.3456	0.3298	0.3366	0.2795

From the Multivariate Correlation Analysis, there are Combination of (50, 2)= 1,225 correlation coefficients (where the notation (A,B) denotes "A choose B", where choose denotes a linear combination) between any two students (A and B). This massive correlation table is a good start to visualize any correlation pattern, but not too effectively to draw any inference on systematic patterns, due to lack of concise summarized information. A better analysis than the Multivariate Correlation Analysis is needed for a deeper investigation.

2.2 Sort Students' Score

To further detect any cheating pattern from any table, the authors then sorted students' scores (reference column) from top to bottom, as presented per Table 2. The sorted data shows that for some scenarios of students sitting at the same – Table No.1, No.15, No.17, and No.4 – not less than 2 of the three students received the same or a very similar score. There is a fair chance that students sitting at the same table have similar scores by random chance. And so, the same total score by student does not definitively prove that cheating occurred, and stronger evidence would be given by finding that suspected students have a similar pattern of answers on the majority of questions! Therefore, we cannot just conclude cheating based on the analysis so far. A better analytical tool to reliably uncover any patterns which would indicate cheating is still needed.

Table 2. Sort Score vs. Table Information

	Reference	Table			
Reference	100%				
SID 4	77%	2	SID 5	29%	17
SID 45	76%	25	SID 11	29%	23
SID 25	59%	16	SID 33	28%	10
SID 34	53%	7	SID19	28%	21
SID 41	53%	11	SID 47	26%	17
SID 21	50%	9	SID 29	25%	8
SID 12	49%	24	SID 24	21%	11
SID 48	48%	19	SID 39	20%	13
SID 3	46%	13	SID 37	20%	11
SID 28	45%	7	SID 8	18%	19
SID 31	42%	18	SID 20	16%	9
SID 1	41%	2	SID 38	15%	12
SID 14	40%	25	SID 27	13%	12
SID 2	38%	8	SID 32	10%	9
SID 49	37%	15	SID 22	10%	10
SID 50	37%	15	SID 46	9%	16
SID 26	36%	1	SID 10	8%	22
SID 44	36%	1	SID 36	7%	4
SID 30	36%	14	SID 6	5%	15
SID 16	35%	3	SID 40	2%	14
SID 18	34%	5	SID 43	-1%	4
SID 17	33%	6	SID 15	-4%	2
SID 35	33%	1	SID 23	-6%	6
SID 13	30%	3	SID 9	-7%	20
SID 42	30%	16	SID 7	-33%	18

3. Data Mining Algorithms and Results

The authors explored more powerful Data Mining algorithms to detect the patterns which would suggest cheating with greater reliability and confidence. JMP 12 Hierarchical Clustering Dendrogram, Heat Map, and Principle Component Analysis were used to detect any cheating pattern.

3.1 Hierarchical Clustering Dendrogram Analysis

Hierarchical Clustering Analysis (HCA)[4] was used to further analyze and uncover evidence of cheating. In data mining and statistics, hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types[5]:

- Agglomerative: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

In the general case, the computing time of the Agglomerative approach is faster than the Divisive approach. Optimal efficient agglomerative methods have been developed to significantly improve the computing algorithm for large data sets[6,7]. The main objective of this analysis was to search for the degree of similarity among exam answers, and to search for patterns (and trends) of similarity, among the students. The Agglomerative approach can identify a clustering pattern faster and more accurately. The Divisive approach may not split the student's scores which are more concentrated on the bottom level efficiently.

Therefore, the authors chose the Agglomerative approach. This approach builds the hierarchy from the individual elements by progressively merging clusters based on a defined distance metric (Euclidean distance). The distance is calculated by the answering discrepancy of each question. This HCA approach can pair the students with similar exam answering patterns and use clustering to isolate those students who cheated from the other students. While Correlation analysis is limited in that it only compares the total exam score per student, Clustering analysis goes a step further since it considers the pattern(s) in which specific questions were answered between students.

JMP 12 was used to calculate the closest distance (the affinity) among all 1,225 pairs, and grouped the first pair, at the strongest affinity (based on their similar answering pattern; see Figure 1 Dendrogram Tree). The linkage criterion determines the distance between sets of observations as a function of the pairwise distances between observations[8,9,and 10]. After grouping the first pair, JMP 12 software calculated the center of the new formed group and found the next strongest affinity pair until the pairs were broken down as shown in the Dendrogram[11] (Figure 1). Four groups [(49, 50), (4, 45), (26, 44, 35), (36, 43)] suspected of cheating were identified. These results are very similar to the previous analyses, using Correlation (Table 1) and Sorting (Table 2), respectively. The result of this analysis – combined with the learnings from the previous –provides very convincing evidence that cheating occurred; it illustrates that the instances where students obtained similar or the same scores occurred where these students were sitting next to each other (at the same table).

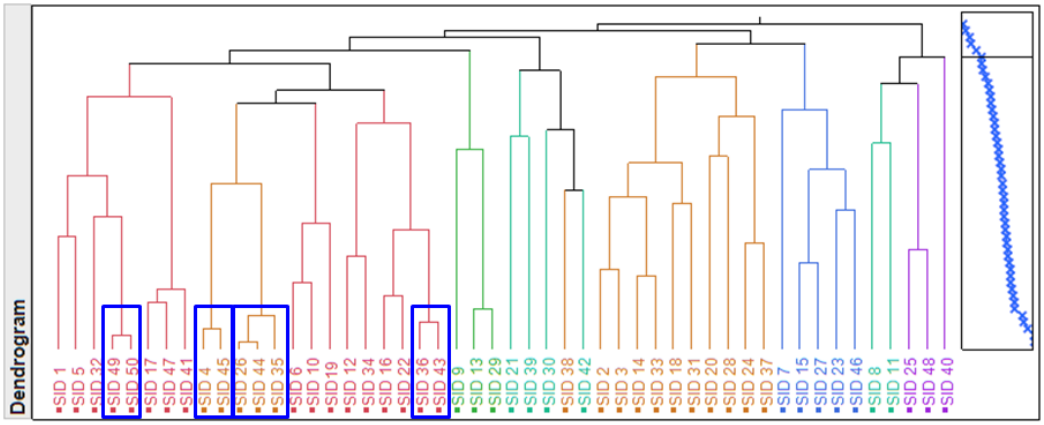

Figure 1. Hierarchical Dendrogram Tree

To further answer the question, authors conducted the JMP Clustering History Analysis (Table 3). Based on the distance metric, the first five clusters were significantly shorter than the following ones. The authors ran the distance outlier test and identified that the lowest five distance numbers were statistically significantly less than the other 44 distance numbers. There is a significant difference in magnitude separation from that of the first five clusters and the remaining ones (bimodal distribution). To ensure the cutoff point (between the first five and the rest) was correct, the authors checked the next five clusters (44-40) in Correlation Analysis (Tables 1 and 2). A weak correlation on their scores was observed, and further their tables were also far away each other. Therefore, we can limit and focus further clustering analysis on the top five clusters. The author's added Exam Table information to verify the hypothesis that cheating occurred in these particular groups. Based on the clustering history, 4 out of the 5 pairings correspond to students that sat at the same table. The 2nd pairing – from two students who sat at different tables – came from the two students with the top exam scores. These two students were sitting at Table 2 and Table 25 (far away each other). The authors made an assumption that the probability of cheating is zero among pairs of students sitting at different tables (and did not incorporate seating distance as a factor in this analysis). Further, we do not find it surprising that the two top-scorers had similar patterns of answers given that they both scored highly on the exam and therefore selected most of the 'correct' answers. The data per Table 2 also showed no objective evidence that students from different tables had high correlation between scores or between answering patterns (except the top two students already identified). Hierarchical clustering analysis yielded very strong evidence of cheating where patterns existed, as evidenced by the significantly lower pairing distance between groups indicated in Red vs other groups (Table 3). Students from Table 1, 4, and 15 have been identified with answer patterns indicative of cheating on the exam. Table 17 (identified in the correlation analysis) is gone in the Clustering analysis. Therefore, clustering analysis is more reliable than correlation analysis.

Table 3. JMP Clustering History Analysis

Clustering History

Number of Clusters	Distance	Leader	Joiner	
49	1.352128704	SID 35	SID 44	Table 1
48	1.425846800	SID 4	SID 45	Different Tables
47	1.496796009	SID 49	SID 50	Table 15
46	1.585511282	SID 36	SID 43	Table 4
45	1.867906951	SID 26	SID 35	Table 1
44	2.409003240	SID 17	SID 47	
43	2.569975255	SID 20	SID 37	
42	2.592763820	SID 13	SID 29	
41	2.625652250	SID 3	SID 31	
40	2.690502326	SID 16	SID 22	
39	2.948493073	SID 6	SID 10	

3.2 Enhanced Hierarchical Clustering Dendrogram Analysis

In order to improve the model accuracy (avoid misjudgement), authors have identified the six most difficult questions shown in Figure 2.

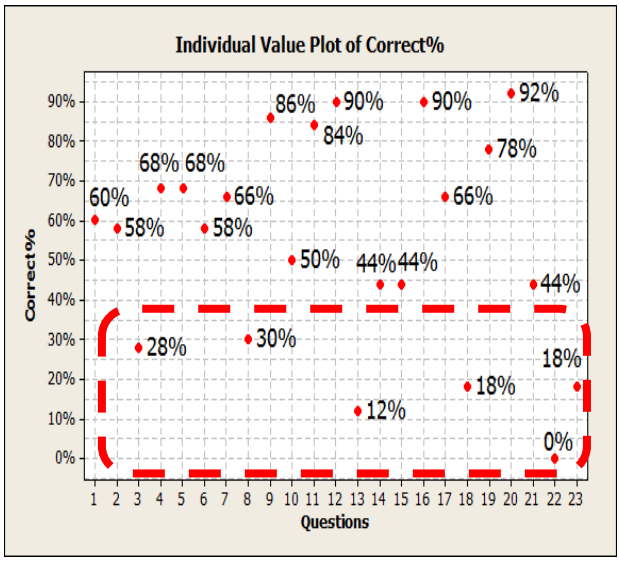

Figure 2. The six most difficult questions.

Students more likely picked the same "correct answer" if question is very easy; less likely picked the same "wrong" answer if question is very difficult. Authors have redone the clustering analysis based on these six most difficult questions. As shown in Figure 3 Clustering Analysis, Tables 1,4, and 15 were identified as cheating tables and their clustering joint distance = 0.000, which means the students from these tables have the identical wrong answers on all six questions. The chance for randomly picking up the same wrong answer= (1/5)*(1/5)*4= 16% chance. The probability of picking the same wrong answers on all six difficult questions is (16%)^(6) < 0.002%. The enhanced clustering model can defend the cheating detection pattern with more than 99.998% confidence. Therefore, students could not defend their cheating pattern in front of this enhanced clustering analysis.

Clustering History

Number of Clusters	Distance	Leader	Joiner	
49	0.000000000	SID 4	SID 34	Top Four Students
48	0.000000000	SID 26	SID 35	Table 1
47	0.000000000	SID 36	SID 43	Table 4
46	0.000000000	SID 26	SID 44	Table 1
45	0.000000000	SID 4	SID 45	Top Four Students
44	0.000000000	SID 49	SID 50	Table 15
43	0.505761336	SID 16	SID 29	
42	0.505761336	SID 8	SID 41	

Figure 3. The clustering analysis of six most difficult questions.

3.3 Heat Map Analysis

JMP Heat Map analysis was conducted to visualize the cheating pattern among the students identified in previous Dendrogram analysis. The easiest way to understand a heat map is to think of a cross table or spreadsheet which contains colors instead of numbers. The default color gradient sets the lowest value in the heat map to dark blue, the highest value to a bright red, and mid-range values to light gray, with a corresponding transition (or gradient) between these extremes. Heat maps are well-suited for visualizing large amounts of multi-dimensional data and can be used to identify clusters of rows with similar values, as these are displayed as areas of similar color[12,13,and 14]. It's very clear from the Heat Map, SID (26, 35, and 44), SID (36, 43) and SID (49, 50) have similar heat map color patterns. This graphical analysis could provide a simpler way of showing objective evidence of answer pattern that indicate cheating among these students.

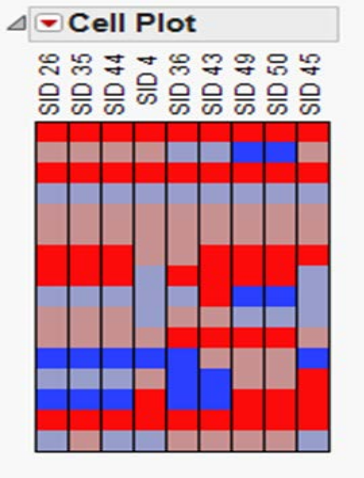

Figure 4. Heat Map Analysis

3.4. Principal Component Analysis (PCA)

Lastly, the authors conducted Principle Component Analysis using JMP 12, with results shown in Figure 3. Principal Component Analysis (PCA) is the general name for a technique which uses sophisticated underlying mathematical principles to transform a number of possibly correlated variables into a smaller number of variables called principal components.

The origins of PCA lie in multivariate data analysis based on Matrix Eigenvalue and Eigenvector algorithms used to derive the two strongest principle components in a linear combination of all the answering variable dimensions[15,16,17,18]. PCA is a very powerful tool for reducing variables' dimensions in larger data sets, in order to reduce the amount of computation and to make the analysis output easier to interpret. The authors used JMPs PCA algorithm to verify the previous clustering patterns observed, as shown by a map of the top two principle components (eigenvectors; Figure 3).

PCA analysis has identified the same four clusters as those indicated by Hierarchical Clustering Analysis. Students SID (26, 35, and 44), SID (36, 43) and SID (49, 50) were assigned in the same region based on the top two principle components (in X-Y). Even the mathematical calculation is different between Hierarchical Clustering (Euclidean Distance) and Principal Component Analysis (Linear Algebra Matrix Eigenvector), but the practical results – which convincingly show those same students that have same or similar answer patterns – are identical. This a good practice to cross-validate three different Data Mining algorithms or tools on reaching the same result and in making the same decision point. At this moment, it would be difficult for students to argue in defense of cheating based on the degree of similarity in the answer patterns identified. Data mining analytical tools (Dendrogram, Heat Map, and Principal Components Map) are significantly more powerful for discovering complicated patterns of association than traditional Analytical Tools (such as Correlation Analysis). PCA also indicated some inter-table cross cheating pattern as shown in Figure 5. SID 35 from Table 1 and SID 36 from Table 4 are in the overlapping area of two PCA clusters. There is a significant chance that these two students may involve in any cross-table cheating activity. Based on the table location, Table 4 is just right next to Table 1 in a short distance. PCA model is more powerful than clustering algorithm to detect such cross-table cheating pattern. This observation may indicate the power of linear combination (eigen vectors) may be more powerful than hierarchical clustering analysis.

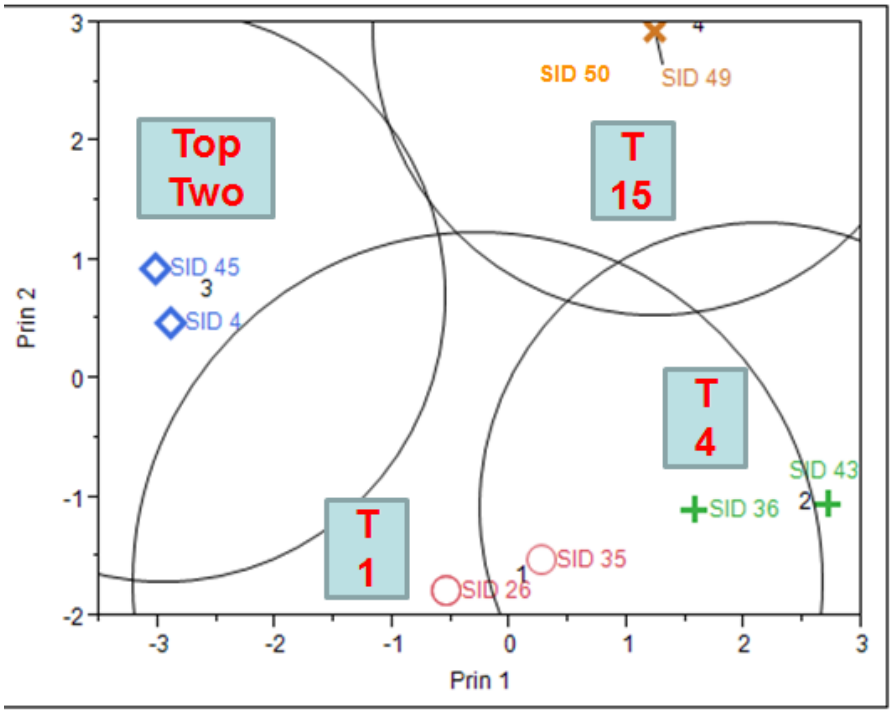

Figure 5. Principle Component Analysis Map

4 Results

Results from all analyses presented previously are summarized per Table 4. Three cheating tables were identified in after taking into account each of the four analyses used. Table 17 was identified as suspect of cheating in the Correlation Analysis, but not in the other three analyses . Correlation Analysis is only based on the accumulated score, not on the correspondence between patterns of answers between individuals. Therefore, with Correlation Analysis there may be a chance of wrong detection (Alpha risk) of cheating, since two students can have the same or similar scores when seated at same table, but their pattern of answers by question can differ significantly. PCA model is more powerful than clustering algorithm to detect such cross-table cheating pattern. This observation may indicate the power of linear combination (eigen vectors) may be more powerful than hierarchical clustering analysis.

The above results have demonstrated the powerful prediction accuracy of detecting similar patterns of answers between students sitting at the same table when taking a multiple-choice exam. The methods and analyses used herein have been shared with other faculty and students in graduate school both to discourage cheating among students and to stimulate students' learning by providing a practical example of real-world statistical techniques used in relation to their daily life.

Table 4. Summary of Data Mining Results

	Correlation Analysis (Table 2)	Clustering Analysis (Table 3)	Heat Map Analysis (Figure 2)	Principal Component Analysis (Figure 3)
1	Table 15 (SID 49, SID 50)	Table 1 (SID 35, SID 44, SID 26)	Table 1 (SID 35, SID 44, SID 26)	Table 1 (SID 35, SID 44, SID 26)
2	Table 1 (SID 26, SID 44, SID 35)	Table 15 (SID 49, SID 50)	Table 15 (SID 49, SID 50)	Table 15 (SID 49, SID 50)
3	Table 17 (SID 5, SID 47)	Table 4 (SID 36, SID 43)	Table 4 (SID 36, SID 43)	Table 4 (SID 36, SID 43)
4	Table 4 (SID 36, SID 43)			

*Note Table 17 (highlighted) is not a likely candidate for cheating among students SID 5 and ID 47.

5 Conclusions

The authors have utilized Data Mining Algorithms such as Multivariate Correlation, Hierarchical Dendrogram Clustering, Heat Map, and Principal Component Analysis to detect patterns in responses to multiple choice exams which indicate cheating took place among students. In the world of Big Data, there are no perfect algorithms which can provide a "catch all" solution to any given problem. Using several Data Mining tools together to cross-validate study results enables the student researcher to make more extensive inferences on their data by considering the data through multiple points of view. Ultimately, this offers the possibility of a more meaningful study conclusion, but choosing the right Data Mining tools or algorithms for the problem is critical for success so as to minimize the risk of algorithm bias. The Data Mining results in this paper sever as a powerful framework to help instructors to manage exam grading for multiple choice exams. By more accurately detecting cases of cheating on these exams, the use of a comprehensive exam question format can be avoided, saving Instructors' exam preparation time and grading time. The authors have identified three tables where students were very likely to have cheated. The prediction accuracy should be very reliable since the answer choice correspondence patterns were identified using various data mining tools (Correlation, Clustering, Heat Map, and Principal Component Analysis) and achieving statistical significance. These students have a poor defense against claims of cheating, based on the extraordinary correspondence between their answers on the exam! The same Data Mining concept and algorithm choices can be applied to many other applications to uncover otherwise hidden patterns such as in: Sports Analytics, Customer Relational Management, or Biostatistics.

Acknowledgements

Special thanks to Chia Lin and Chao-Yuan Chen for supporting our research. Thanks to the Graduate Students from International Technology University (ITU) in Business Analytics, for their efforts in gathering multiple-choice exam raw data. Thanks to Katherine Lim and Timothy Liu for meaningful discussions and feedback on our Data Mining approach. Thanks to the technical support staff at JMP for providing us the Data Mining techniques used herein.

References

[1] Kevin Yee and Patricia MacKown, p. 141-147, Center for Academic Integrity, Rutland Institute for Ethics, Clemson University. Available at http://www.academicintegrity.org

[2] Finding cheaters using multiple-choice comparisons http://jd-mathbio.blogspot.com/2015/02/finding-cheaters-using-multiple-choice.html
[3] JMP Multi-Correlation Analysis http://www.jmp.com/support/help/Correlations_and_Multivariate_Techniques.shtml
[4] Michael R. Anderberg. Cluster analysis for applications. Academic Press, New York, 1973. ISBN 0120576503.
[5] Rokach, Lior, and Oded Maimon. "Clustering methods." Data mining and knowledge discovery handbook. Springer US, 2005. 321-352
[6] R. Sibson. "Slink: an optimally efficient algorithm for the single-link cluster method". The Computer Journal, British Computer Society. 16 (1): 30-34, 1973.
[7] D. Defays. "An efficient algorithm for a complete-link method". The Computer Journal, British Computer Society. 20 (4): 364-366.
[8] Szekely, G.J. and Rizzo, M.L. "Hierarchical Clustering vis Joint Between-Within Distances: Extending Ward's Minimum Variance Method". Journal of Classification 22, 151-183, 2005.
[9] Ward, Joe H. "Hierarchical Grouping to Optimize an Objective Function". Journal of the American Statistical Association. 58(301): 236-244.
[10] Ma, et al. "Segmentation of multivariate mixed data via lossy data coding and compression." IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(9) (2007): 1546-1562.
[11] Daniel Mullnier. Modern hierarchical, agglomerative clustering algorithms. Arxiv: 1109.2378V1 (stat.ML), 2011.
[12] Tibco. https://docs.tibco.com/pub/spotfire/6.5.2/doc/.../heat/heat_what_is_a_heat_map.htm
[13] Wilkinson, Leland; Friendly, Michael (May 2009). "The History of the Cluster Heat Map". The American Statistician. **63** (2): 179–184
[14] Perrot, A.; Bourqui, R.; Hanusse, N.; Lalanne, F.; Auber, D (2015). "Large interactive visualization of density functions on big data infrastructure". IEEE 5th Symposium on Large Data Analysis and Visualization (LDAV), 2015: 99–106
[15] Soren Hojsgaard. Examples of multivariate analysis Principal component analysis (PCA). Statistics and Decision Theory Research Unit, Danish Institute of Agricultural Sciences
[16] Pearson, K. (1901). "On Lines and Planes of Closest Fit to Systems of Points in Space". Philosophical Magazine. 2 (11): 559–572
[17] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24, 417–441, and 498–520
[18] Abdi. H., & Williams, L.J. (2010). "Principal component analysis". Wiley Interdisciplinary Reviews: Computational Statistics. 2 (4): 433–459

Biography

Mason Chen is student in Stanford On-Line High School Program. Mason has certified Lean Six Sigma Black Belt through IASSC (International Associate of Six Sigma Certification), and also certified IBM SPSS/Modeler Statistics and Data Mining Certificates. Mr. Chen has been invited to several conferences like IEOM, ASQ, AQI, ASA, JMP/SAS and local ASQ SV statistics group to present his STEM Projects. His STEM projects have drawn interest in Robotics/EV3, JAVA Science, Poker Probability, Powerball Lottery, Sports Analytics, Biostatistics and Healthcare Statistics… Mason is familiar with Lean Six Sigma DMAIC, DFSS, and Minitab 17, JMP 13, SPSS 24, and Modeler 18 Statistics Software. Mason has also been learning Data Mining and Big Data Analytics through several STEM Projects. As a Stanford High School Student, he has published several Conference Proceeding Papers in IEOM, ISF, IWSM, FSDM conferences.