Group 7

Applying Data Mining Techniques with Bayesian Networks and Cluster Analysis on Decision of Traditional General Education Course Digitalization

105034539 ROSS LEE
105034533 JACKSON CHU
105034572 DOLLAR HUANG

Outline

- Introduction
- Literature Review
- Methodology
- Case Study
- Research Conclusion

Introduction

Background

- The term MOOC was coined in 2008 by Dave Cormier from the University of Prince Edward Island in Canada.
- Basically, videos are recorded by school professors, experts and researches and put on the internet as online course videos. These videos are integrated and developed as MOOCs.
- However, in spite of the benefits MOOCs provide, the effectiveness is widely questioned. One of the primary reasons is its low participation rate which consists of online discussions, posting etc.

Motivation

- Followed by the rapid development of the Internet and its use for educational purposes, a great deal of MOOC platforms appeared within a decade and the boundary between online and traditional course becomes blurrier than before.
- Compared to the traditional classes, MOOCs equip with several advantages including the flexibility to control speed of videos, watching videos without location limitations, etc. Therefore, a few schools begin to replace traditional courses with online courses and consider these credits verified credits. This situation forms a vague boundary between online courses and traditional courses.

Research Purpose

- In order to help school determine whether to include credits from MOOCs, students' perspectives are collected. By pointing out the key pros and cons that causally lead to the result of course digitalization and comparing it to the satisfaction of taking traditional courses gathered from students, this study provides suggestions about course digitalization decisions.

Literature Review

MOOCs

- MOOCs opens the course to the online users and they can participate in the learning process.

- Main problem： Anoush Margaryan, Manuela Bianco, Allison Littlejohn (2015) proposed Instructional quality of Massive Open Online Courses (MOOCs). Student and teacher is difficult to have interaction in online course is a main problem.

- The challenge：Anthony G. Picciano(2002) considered the interaction between people is the challenge in the future.

Cluster Analysis

- Cluster analysis is a good tool used to find the patterns and information hidden in data.

- K-means Algorithm: Split the observation into k groups.

- Application: The voting patterns、data-mining、and machine learning…

Bayesian Network

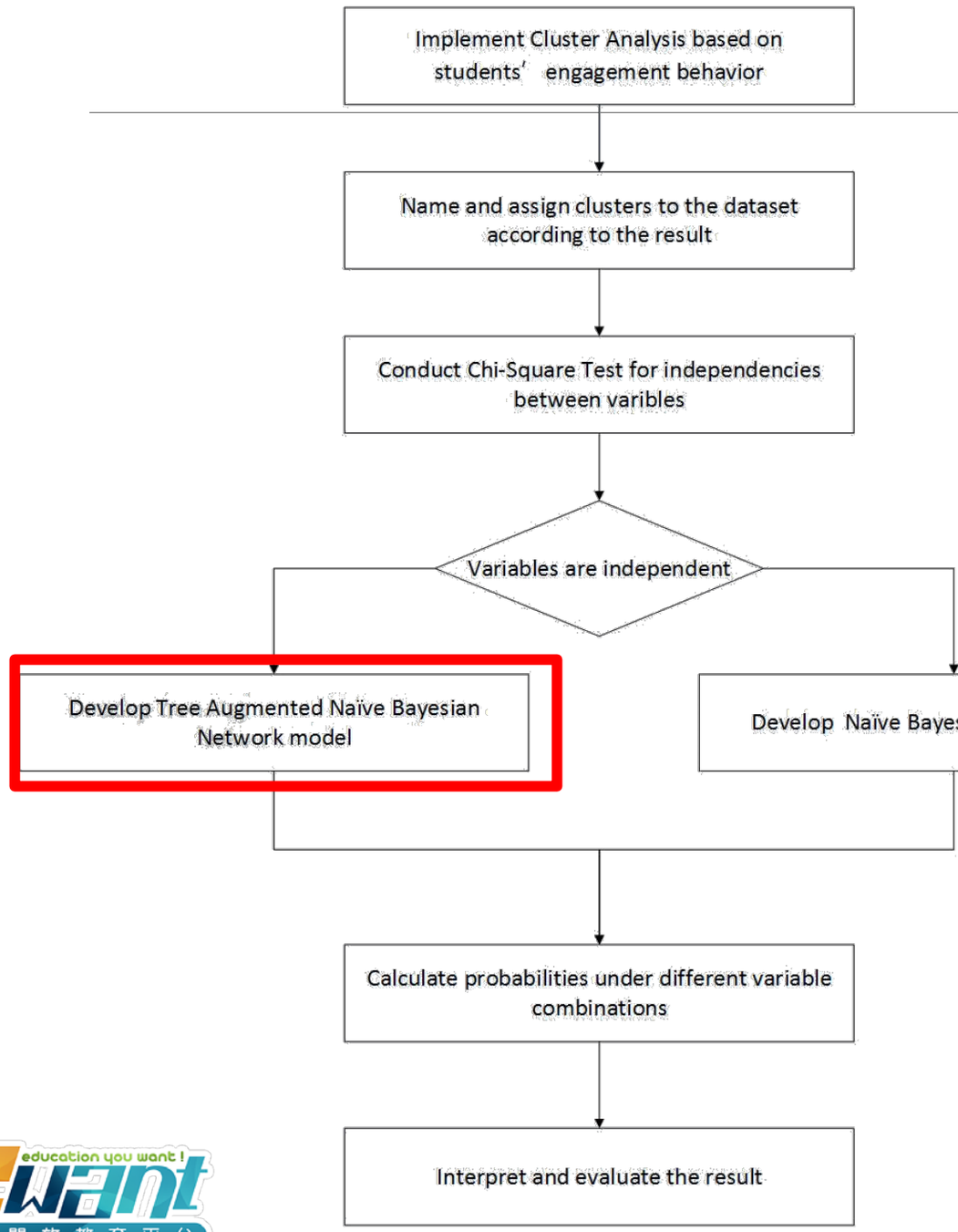
- Bayesian network is a graphical representation of uncertain quantities, which can reveal the probabilistic dependencies between a set of variables, that is, causal relationships.

- Huang (2009) proposed a three-layer Bayesian network method to evaluate network-course learning effect. Xiao (2009) presented Bayesian network model to predict students' accomplishments and provide advice in terms of the arrangements of university courses.

- The Tree Augmented Naive Bayes(TAN): Friedman, Geiger, & Goldszmidt (1997) improve the weakness of original Bayesian Network which has to assume variable independencies and computing workload by considering maximum two prior events for each node.

Methodology

Research Procedures



- Step1

  Use cluster to classify student data.

- Step2

  Name and assign clusters to dataset.

- Step3

  Use Chi-Square Test to confirm independence.

- Step4

  Calculate probabilities under different variable combinations.

- Step5

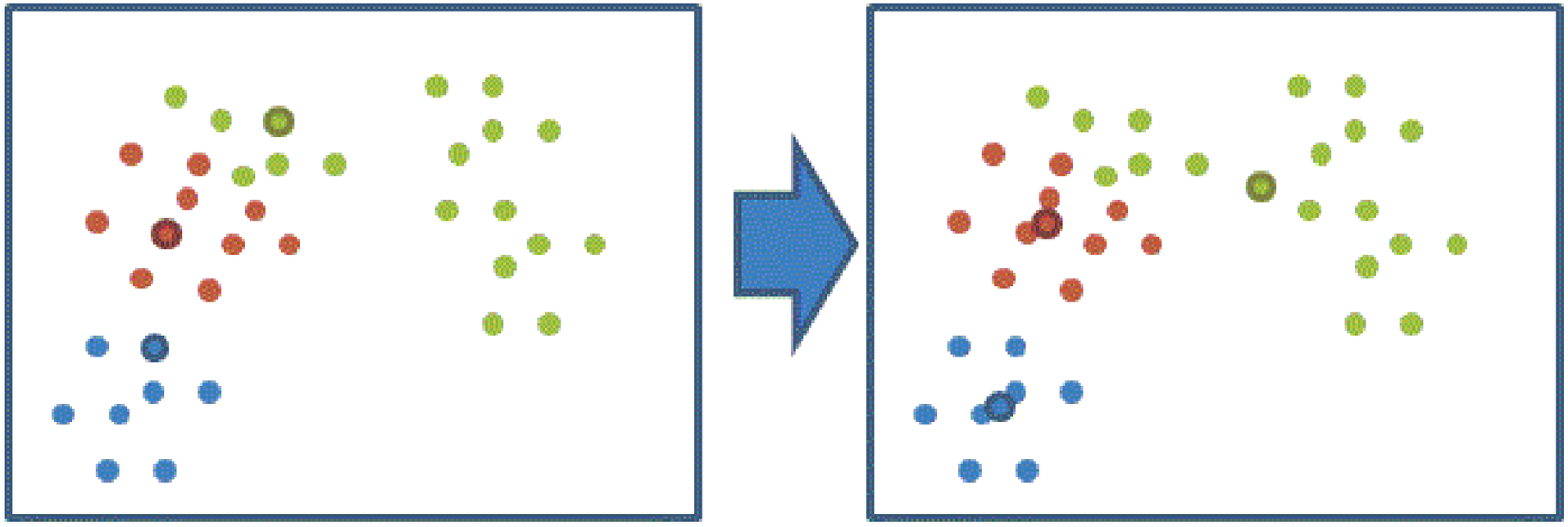
  Interpret and evaluate the result

Cluster Analysis

K-means algorithm

in other word is “Birds of a feather flock together.”

$\min \sum_{i=1}^{k} \sum (x_j - \mu_j)^2$



Until convergence, the moving distance will become very small

Bayesian Network

- **Bayes' theorem** is stated mathematically as the following equation:

$$P(A|B) = \frac{P(B|A)\cdot P(A)}{P(B)} \qquad (1)$$

where A and B are events and $P(B) \neq 0$

- **Bayesian Network** joint probability function is:

$$P(A|B,C) = P(A|B,C) \cdot P(B|C) \cdot \mathrm{P(C)} \qquad (2)$$

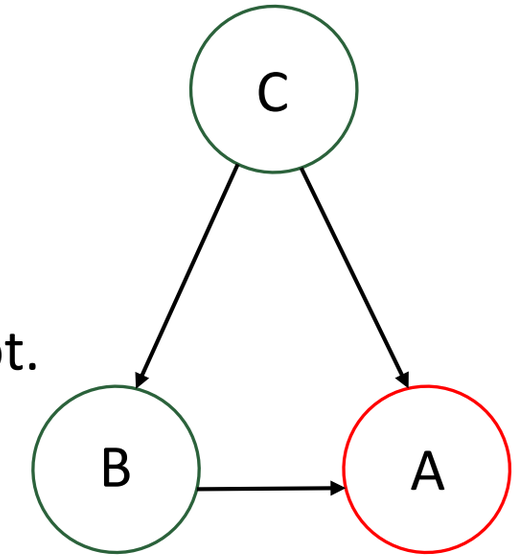
Where B and C are two events which can cause A event to happen.

Bayesian Network

- Take a simple example to illustrate Bayesian Network:

	A	B	C
Levels	1, 0	1, 0	1, 0

Assume A, B, C all have two levels
1 means occurring and 0 means not.



Event A is cause by event C & B in the meantime. Stated as: P(A|B,C)

- When C = 0 and B = 0, we can refer the conditional probability of event A = 0.82432
- Given A = 0, B = 1, what's the probability event C occurring?

		A	
		1	0
C=0	B=0	0.82432	0.60753
	B=1	0.17568	0.39247
C=1	B=0	0.36792	0.3549
	B=1	0.63208	0.6451

Case Study

Data preparation

NCTU has developed online general courses actively

A questionnaire of general courses for college students, expect to figure out students' learning situation and opinion such as

- ✓Engagement ⟶ **Cluster Analysis**
- ✓Satisfaction for traditional courses
- ✓Pros and cons of online courses
- ✓Agree or Do not agree for online general courses ⟶ **Bayesian Network**

Stratified sampling in order to align the proportion of colleges

Implement Cluster Analysis

Engagement part of questionnaire

- Attendance
- Distraction
- Investing Time for HW
- Preparing Time for Exam
- Extracurricular

Cluster Analysis - K-means

- In order to distinguish different student's course engagement behavior, K-means cluster analysis is applied.
- Each cluster's center is calculated after several iterations where two initial centers are randomly assigned.
- The table below shows the results of k-means algorithm.

Cluster /Attr.	Attendance	Distraction	InvestTime_HW	PreparingTime_Exam	Extrac urricular
1	4.469799	3.604027	1.765101	1.832215	1.080537
2	4.831858	2.274336	3.150442	2.327434	1.327434

However, two attributes seem insensitive.

Cluster Analysis - K-means

- With respect to the attributes, absolute difference between cluster centers are computed.

Attendence	Distraction	InvestTime_HW	PreparingTime_Exam	Extracurricular
0.3486303	1.0434438	1.5475505	0.5662824	0.2428334

- We've found attributes "Attendance" & "Extracurricular" are relatively insensitive after several iterations.

- Therefore, three attributes left are considered to be the discriminant variables in cluster analysis.

Cluster /Attr.	Distraction	InvestTime_HW	PreparingTime_Exam
1	3.532051	1.769231	1.807692
2	2.292453	3.235849	2.396226

Cluster Analysis

- Two personal characteristics of students are defined based on the result of cluster analysis.

Cluster /Attr.	Distraction	InvestTime_HW	PreparingTime_Exam
1	3.532051	1.769231	1.807692
2	2.292453	3.235849	2.396226

Name and assign clusters

- Students in the first group doesn’t concentrate in class. They also spend little time on homework and exams. This group is named as “Slack Students”.
- Students in the second group are focused in class. They also take more time for homework and exams. However, they don’t diligent extremely so this group is named as “Ordinary Students”.

Chi-Square Test

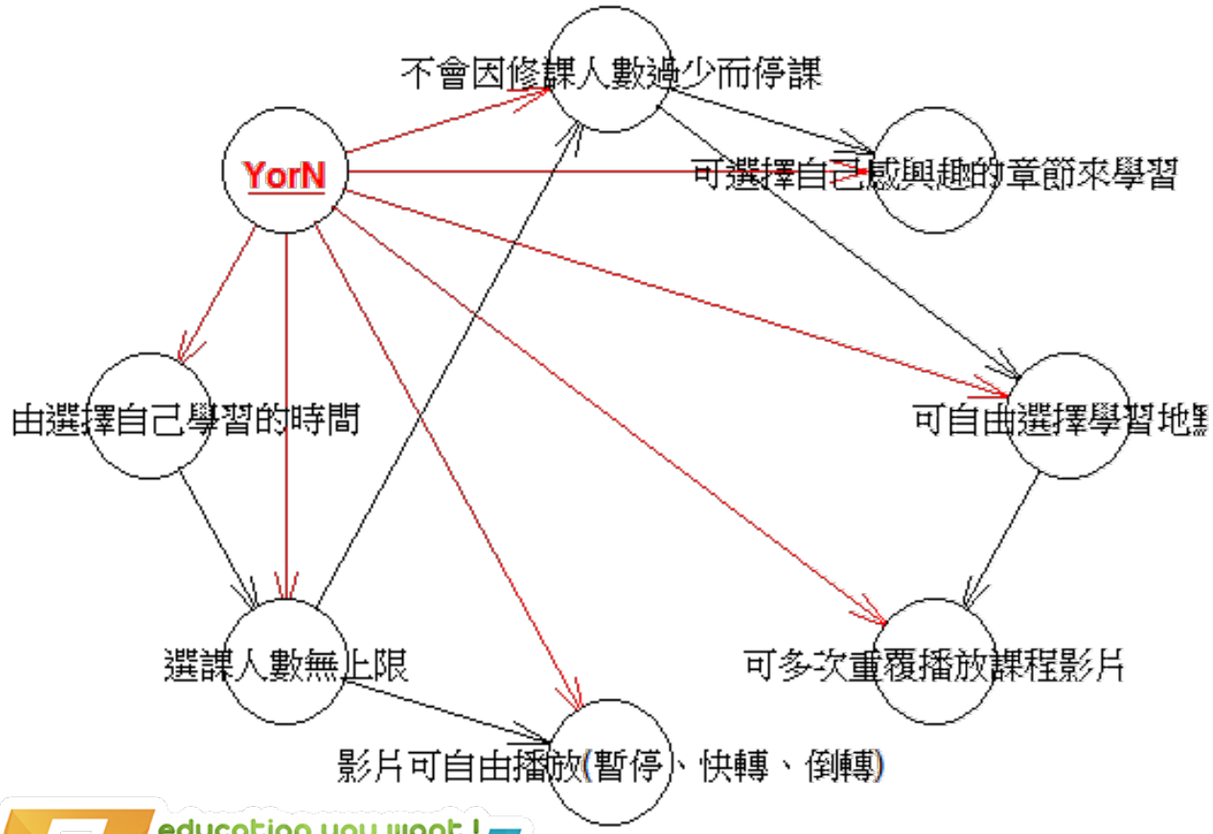
- We also conduct Chi-Square Test for independencies between variables so that the appropriate model can be selected.

	Pro Group 1	Pro Group 2	Cons Group 1	Cons Group 2
p-value	2.2E-16	2.982E-13	2.2E-16	1.98E-11

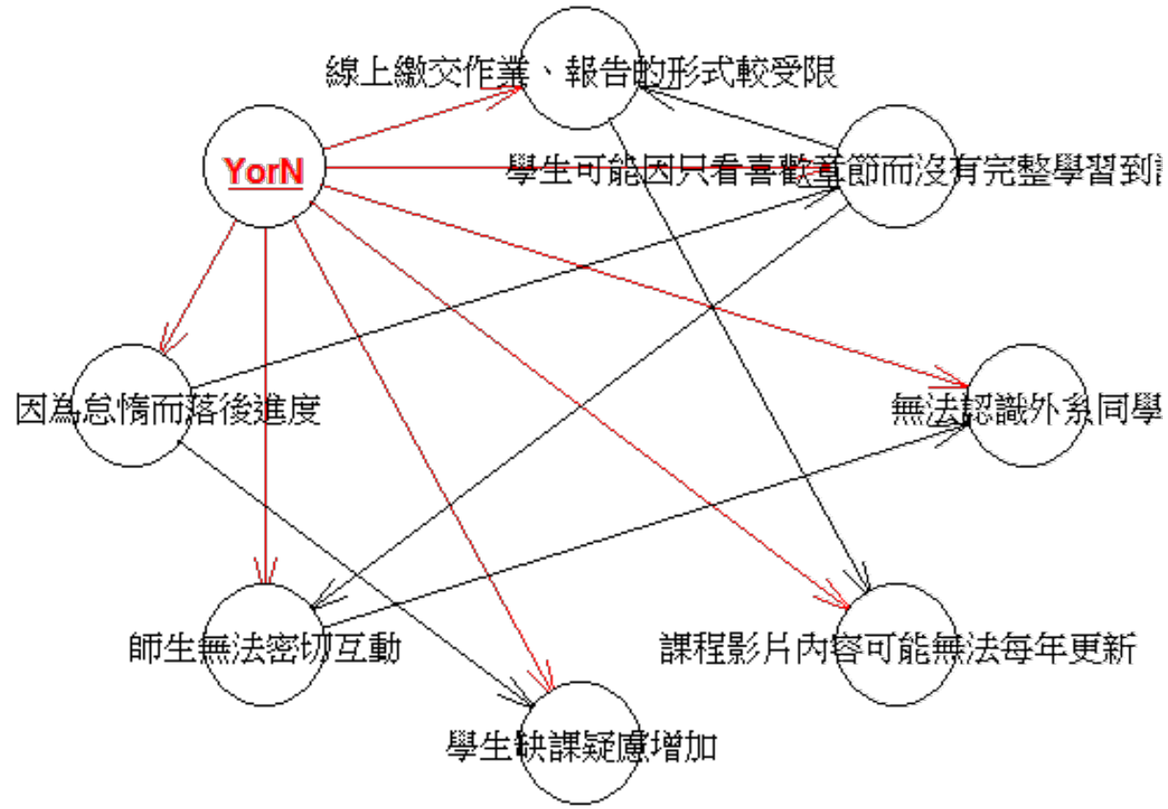
- The figure shows that all variables are dependent so that this study selects **Tree Augmented Naïve Bayesian Network** as the model to demonstrate following analysis.

Develop Tree Augmented Naïve Bayesian Network model

Pros



Cons



Calculate probabilities under different variable combinations

- This table shows the conditional probability of taking MOOCs courses given student choosing one of seven pros of MOOCs.

Pros							
Group1							
P1	P2	P3	P4	P5	P6	P7	Probability
1	0	0	0	0	0	0	0.340329
0	1	0	0	0	0	0	0.2911554
0	0	1	0	0	0	0	0.7211238
0	0	0	1	0	0	0	0.6553876
0	0	0	0	1	0	0	0.4139952
0	0	0	0	0	1	0	0.681185
0	0	0	0	0	0	1	0.2699275

可自由選擇自己學習的時間	P1
選課人數無上限	P2
影片可自由播放(暫停、快轉、倒轉)	P3
可多次重覆播放課程影片	P4
可自由選擇學習地點	P5
可選擇自己感興趣的章節來學習	P6
不會因修課人數過少而停課	P7

Assume variables are independent.

“可多次重覆播放課程影片” is a key advantage

Calculate probabilities under different variable combinations

Pros							
Group1							
P1	P2	P3	P4	P5	P6	P7	**Probability**
1	1	0	0	0	0	0	0.2656145
1	0	1	0	0	0	0	0.6777647
1	0	0	1	0	0	0	0.6126834
1	0	0	0	1	0	0	0.3645932
1	0	0	0	0	1	0	0.6394895
1	0	0	0	0	0	1	0.2340619
0	1	1	0	0	0	0	0.6279483
0	1	0	1	0	0	0	0.5492715
0	1	0	0	1	0	0	0.3127613
0	1	0	0	0	1	0	0.5868769
0	1	0	0	0	0	1	0.5075209
0	0	1	1	0	0	0	0.6620172
0	0	1	0	1	0	0	0.7396392
0	0	1	0	0	1	0	0.9050481
0	0	1	0	0	0	1	0.6085408
0	0	0	1	1	0	0	0.6838208
0	0	0	1	0	1	0	0.6772972
0	0	0	1	0	0	1	0.5425509
0	0	0	0	1	1	0	0.6399233
0	0	0	0	1	0	1	0.2803038
0	0	0	0	0	1	1	0.5708229

可自由選擇自己學習的時間	P1
選課人數無上限	P2
影片可自由播放(暫停、快轉、倒轉)	P3
可多次重覆播放課程影片	P4
可自由選擇學習地點	P5
可選擇自己感興趣的章節來學習	P6
不會因修課人數過少而停課	P7

P4	P5	P6	P7	**Probability**
0	0	0	0	0.340329
0	0	0	0	0.2911554
0	0	0	0	0.7211238
1	0	0	0	0.6553876
0	1	0	0	0.4139952

Actually, variables exist dependencies!

Interpret and evaluate the result(1/2)

Approach – compare with satisfactions of traditional courses.

- After obtaining the **conditional probabilities** of students' intention of taking MOOCs courses and **key pros and cons** of MOOCs, we compare it to the satisfactions of traditional general education courses.

學生自主彈性

- These are satisfactions of traditional general education courses.
- Each row has an advantage or disadvantage to refer.

教師教學方式
（包含速度、方法）

Avg. Satisfaction	Level
3.125954	Low
2.816794	Low
2.477099	Low
3.217557	Low
3.01145	Low
3.232824	High
3.526718	High
3.301527	High
3.80916	High
2.942748	Low
3.122137	Low
3.477099	High
3.610687	High
3.259542	High
2.881679	Low
3.232824	High

		Probability
P3	P5	0.7396392
P3	P6	0.9050481
P4	P5	0.6838208

Intentions of taking MOOCs.

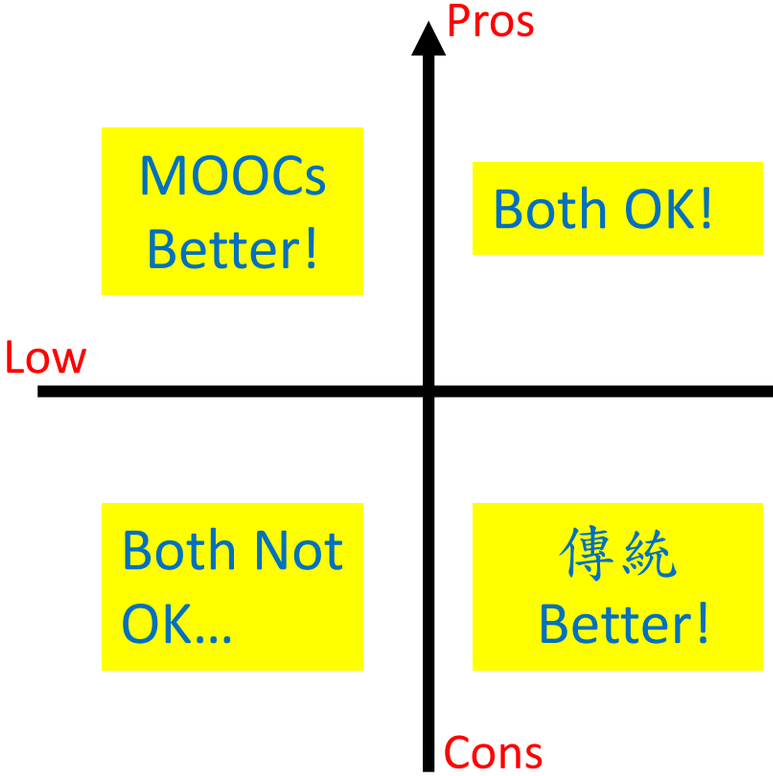
可自由選擇自己學習的時間	P1
選課人數無上限	P2
影片可自由播放(暫停、快轉、倒轉)	P3
可多次重覆播放課程影片	P4
可自由選擇學習地點	P5
可選擇自己感興趣的章節來學習	P6
不會因修課人數過少而停課	P7

- Students are already satisfied with traditional general education courses.

Interpret and evaluate the result(2/2)

- Apply this approach to each cluster and results are shown below:

Key pros and cons



Highlight the super key pros and cons.

Pros Group 1			
		Probability	Quadrant
P3	P6	0.9050481	Both OK!
P3	P5	0.7396392	Both OK!
P4	P5	0.6838208	MOOCs better!

Pros Group 2			
		Probability	Quadrant
P6	P5	0.96525216	MOOCs better!
P1	P5	0.94704203	MOOCs better!
P3	P5	0.9088385	Both OK!

Cons Group 1			
		Probability	Quadrant
C1	C7	0.68216833	Both Not OK…
C6	C7	0.67702809	Both Not OK…
C4	C6	0.58906771	傳統 Better!

Cons Group 2			
		Probability	Quadrant
C5	C6	0.7488222	傳統 Better!
C5	C7	0.71224449	Both Not OK…
C5	C1	0.64920255	Both Not OK…

It's OK for MOOCs in campus!

Research Conclusion

Conclusion & Recommendations

Conclusion

- Through Bayesian Network model, key advantages and disadvantages of MOOCs can be identified, which are based on students' perspective.
- In case study, this analysis also provides college related faculties with practical suggestions to support their decision whether to digitalize traditional general education courses and include credits.

Recommendations

- All the analysis in this study is based on existed survey data. In the future, different perspectives can be included in the analysis to obtain more complete results.

Thank you for listening

Back up

◆ 通識課程實際認知問卷

項目	非常同意 1	同意 2	普通 3	不同意 4	非常不同意 5	總計	算術平均數
1. 您認為學校通識課程是否培養學生獨立思考的能力	13	61	95	66	27	262	3.1259542
2. 您認為學校通識課程是否建立學生關懷當代世界文明的情懷	25	45	106	71	15	262	3.0229008
3. 您認為學校通識課程是否培養學生參與公民社會的意識	23	36	116	70	17	262	3.0839695
4. 您認為學校通識課程是否培養學生溝通及領導才能	19	62	91	66	24	262	3.0534351
課程制度規劃							
5. 您認為學校通識課程的安排是否平均考量各個向度的能力發展	23	54	71	95	19	262	3.1259542
6. 您認為學校通識各向度中課程的種類豐富程度是否滿足學生欲培養的能力	37	73	70	65	17	262	2.8167939
(舉例：假設有一文學向度，若夠豐富應包含古典文學、現代文學及外國文學)							
7. 您認為學校通識各課程人數上限是否夠高，能夠滿足學生需求	50	101	59	40	12	262	2.4770992
8. 您認為學校通識課程是否開設多元時段，能夠滿足全校學生的時間安排	34	57	47	66	58	262	3.2175573
9. 您認為學校通識課程必修學分規定是否考量學生的自主彈性	34	64	62	69	33	262	3.0114504
師資及教學							
10.您認為學校通識課程是否在選課前提供完整的課程資訊	14	64	64	87	33	262	3.2328244
11.您認為學校通識課程的評分是否適當分配作業、考試、報告、出席率的比例	8	18	85	130	21	262	3.5267176
12.您認為學校通識課程教師的教學方式是否符合學生對課程的需求	8	24	126	89	15	262	3.3015267
13.您認為學校通識課程的師資是否具有足夠的專業能力	3	6	67	148	38	262	3.8091603
學習成效							
14.您認為學校通識課程的考試是否能準確衡量學生的學習成果	24	44	125	61	8	262	2.9427481
15.您認為學校通識課程的考試是否能督促學生學習該課程	19	51	84	95	13	262	3.1221374
16.您認為學校通識課程的作業及報告是否有助於該課程的能力培養	10	33	64	132	23	262	3.4770992
17.您認為學校通識課程的課堂討論是否有助於該課程的能力培養	10	16	74	128	34	262	3.610687
18.您認為學校通識課程的授課內容是否有助於學生培養該課程期望達到的能力	11	31	110	99	11	262	3.259542
學習環境							
19.您認為學校通識課程的教室設備(如投影機、桌椅)是否考量學生的學習舒適度	29	73	77	66	17	262	2.8816794
20.您認為學校通識課程的教室空間大小是否考量學生人數	12	56	75	97	22	262	3.2328244