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

Chi-Ming Lee*, Chen-Chia Chu, Sin-Yuan Huang, Ming-Chuan Chiu
Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Taiwan
*s105034539@m105.nthu.edu.tw

Abstract. By the popularization of the computer and the developing of the Internet, more and more MOOCs (Massive Open Online Courses) platforms which is also known as distance learning and consists of taking classes via the internet are developed. However, a vague boundary starts to form between online courses and traditional courses. Most students still follow the policies of taking traditional classroom courses which require students to attend classes in person and on campus instead of the online courses. This study applies Data Mining Techniques including Cluster analysis and Bayesian Network Model with two-phase analysis frame. Connecting these critical attributes with students' satisfaction involving in MOOCs, this study finds out the key advantages and disadvantages through the causal relationships between them. In comparison of each dimension of traditional course satisfaction accordingly, some suggestions are provided for supporting decisions whether to digitalize the traditional general education courses.

Keyword: Data Mining, MOOCs, Bayesian Networks, Cluster Analysis

1. Introduction

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. Many reports have also mentioned the possible integrity of MOOCs and traditional courses. (Holotescu, Grosseck, Cretu, & Naaji, 2014) 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 courses. These online courses are integrated and developed as MOOCs. Compared to the traditional classes, MOOCs is equipped 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 obtaining solely in traditional courses before. This situation forms a vague boundary between online courses and traditional courses. 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. Moreover, Lori Breslow et al. (2013) indicates one of the most troubling aspects of MOOCs to date is their low completion rate, which averages no more than 10%. In order to 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 for colleges about decisions of course digitalization. These data can be qualitatively and quantitatively analyzed in use of data mining techniques. With two-phase analysis frame, at

first, the cluster analysis is implemented to cluster different groups of students based on engagement level of MOOCs. The second phase is followed by the demonstration of Bayesian network model, focusing on finding key indicators, that is, in this paper, students' perspectives on MOOCs. Papers that introduce multiple data mining approaches as framework to analyze such issues are rare. Through this research, further analysis is conducted and the practical advice are provided for school administrators. This study is structured as follows: Literature Review, Methodology, Case Study, Results & Analysis and Conclusion.

2. Literature Review

2.1. Massive open online course

Massive open online course (MOOCs) is said to be a new term of learning knowledge in online course. MOOC opens the course online and users can participate in the learning process. This term was used to describe an online open course 'Connectivism and Connective Knowledge', which was developed at the University of Manitoba by George Siemens and Stephen Downes and had over 2200 participants from all over the world (Margaryan, Bianco, & Littlejoh, 2015). Stanford University, Massachusetts Institute of Technology and Harvard University have also begun to try this new class since 2011. Kay, Reimann, Diebold & Kummerfeld (2013) said the quality, timing, and form of feedback is critical to effective learning. This is the topic this platform wants to overcome because it has too many uncertain factors. The online course is traditional education combine with distance education. A good online course provides flexibility to students. In this way, the students can configure their own schedule in keeping with their necessity and degree of learning (Gil, Jara, Candelas, & García, 2012). An excellent online course can follow that massive numbers of students will grab the chance to get a first-rate education for free. Picciano (2002) used the Technology Acceptance Model (TAM) to evaluate online course system. The critical success factors for online course resources can be generalized into these points: human factors pertaining to the instructors; the instructors' and students' technical competency; the instructors' and students' mindset (about learning); the level of collaboration in the course; and the level of perceived IT infrastructure and technical support (Soong, Chan, Chua & Fong, 2001). Picciano (2002) mention that an essential element for learning in a typical classroom environment is the social and communicative interactions between people, especially between teachers and students. However, rare researches take key factors from students' perspectives into account. In this study, causal relationships between these key factors are considered and corresponding suggestions are provided.

2.2. Cluster Analysis

Cluster analysis is a tool for simplifying multivariable data. When there are more than two variables, cluster analysis is often used for identifying clusters by gathering data according to their similarities. Clustering has applied in many area, such as voting patterns, data mining, and machine learning. There are several methods for calculating the distance between clusters, such as centroid method and Ward's method. Centroid method which use center of a cluster as a new point is the most popular way. K-means algorithm is the collocation algorithm which is developed depending on the combination of initial partition and reassignment rule employed. Singh, Sabitha, & Bansal (2016) used K-means algorithm, a data mining technique, to classify students into different clusters. Before that, they have to collect students' performance data during their entire term. The result of cluster analysis will help the university and technical organizations come up with strategies for improving academic performance. Cluster analysis is also a good tool used to find the patterns and information hidden in e-learner data. In an e-learning environment, it's important to find out interest of student, especially the common interest of students. Students may have a variety of learning habits, and it affect the e-learning too. Zhao, Gu, & He (2010) used cluster algorithm and transitive closure for clustering access patterns of student in an e-learning environment. A student access pattern represents a unique

surfing behavior. If a web page appears in several student access patterns, this implies that these students show common interests on this web page. Hani, Hooshmand, & Mirafzal (2013) also used cluster analysis as a data mining tool for identifying factors which affect e-learning. This study especially focuses on clustering students' learning engagement behavior and implements further analysis based on each group.

2.3. Bayesian Network

Bayes' Theorem is a process where the prior probability is adjusted as posterior according to updated knowledge. Researchers are able to infer and classify data labels by training data and establishing training model. This tool allows to investigate the relationship between the complex probability structure. Bayesian network is a probabilistic graphic model that represents probabilistic dependence among a set of random variables (Ben-Gal, 2007). Combining principles from many subjects and fields (Ben-Gal, 2007), BNs is widely applied not only due to its ability to express probabilistic relationships with graphical representations, but discover causal influences form data (Yu & Liu,2016). Xiaorong (2009) proposed a three-layer Bayesian network method to evaluate network-course learning effect. Several evaluation indexes of network-course learning are included in the model, such as program of study, pattern of manifestation, instructional design, multimedia effects and interactive function to assess the performance of network-learning. Considering certain relationships and interactions between university courses, Huang & Fang (2009) presented Bayesian network model to predict students' accomplishments and provide advice in terms of the arrangements of university courses. Itoh, Hirotaka, Nishiwaki, & Funahashi (2015) used Bayesian network to identify those students who require curriculum guidance. Also, Bayesian network is a statistical model able to find out the hidden information from data. West, et al. (2010) applied Bayesian network to extract the latent factors within the relationships between students' performance, assessment tasks and course instructions and align this information with learning progression frameworks. The traditional Bayesian network has an assumption that variables have to be inter-independent. To improve this inherent limitation, Friedman, Geiger, & Goldszmidt (1997) propose Tree Augmented Naive Bayes by combining Bayesian network(TAN) and Naive Bayes. TAN compensates the weakness from Naive Bayes which assumes that events have to be independent with each other and that usually lead to probability distortion. Furthermore, TAN doesn't consider all relationships between variables, instead, it takes maximum two variables simultaneously to avoid over complex computations. Since the premiere of online courses, the increasing online course enrollment highlights the significance for MOOCs. However, studies which link students' perceptions and perspectives of the online course to satisfactions on university general education experiences are rare. This study intends to extract key indicators of MOOCs pros and cons, and advice schools with appropriate decisions about traditional course digitalization.

3 Methodology

This research aims to provide advice for university general education division to ascertain whether to put these general education courses online by applying Data Mining techniques. This study constructs two-phase analysis framework. Phase I demonstrates the use of cluster analysis to divide learners into several groups based on their engagement behavior features on learning experience. These attributes containing participation of the course and the time taken in studying and exam preparation, etc. After the assignment of the cluster result, this study also implements the Chi-Square to test probabilistic dependence between variables.

Based on the result of Chi-Square Test, phase II presents the Bayesian network model to refine the crucial latent factors. These factors affect the intention of users taking online courses and the level of engagement. Next. this study computes the conditional probabilities of the strength and weakness of online courses from students and comes up with the conditional probabilities table. In the end, the interpretation of result and evaluation are presented.

3.1. Study Process

This study demonstrates procedures following the order of Cluster Analysis, Chi-Square Test, Bayesian Network and the interpretation and evaluation of results in Figure 1.

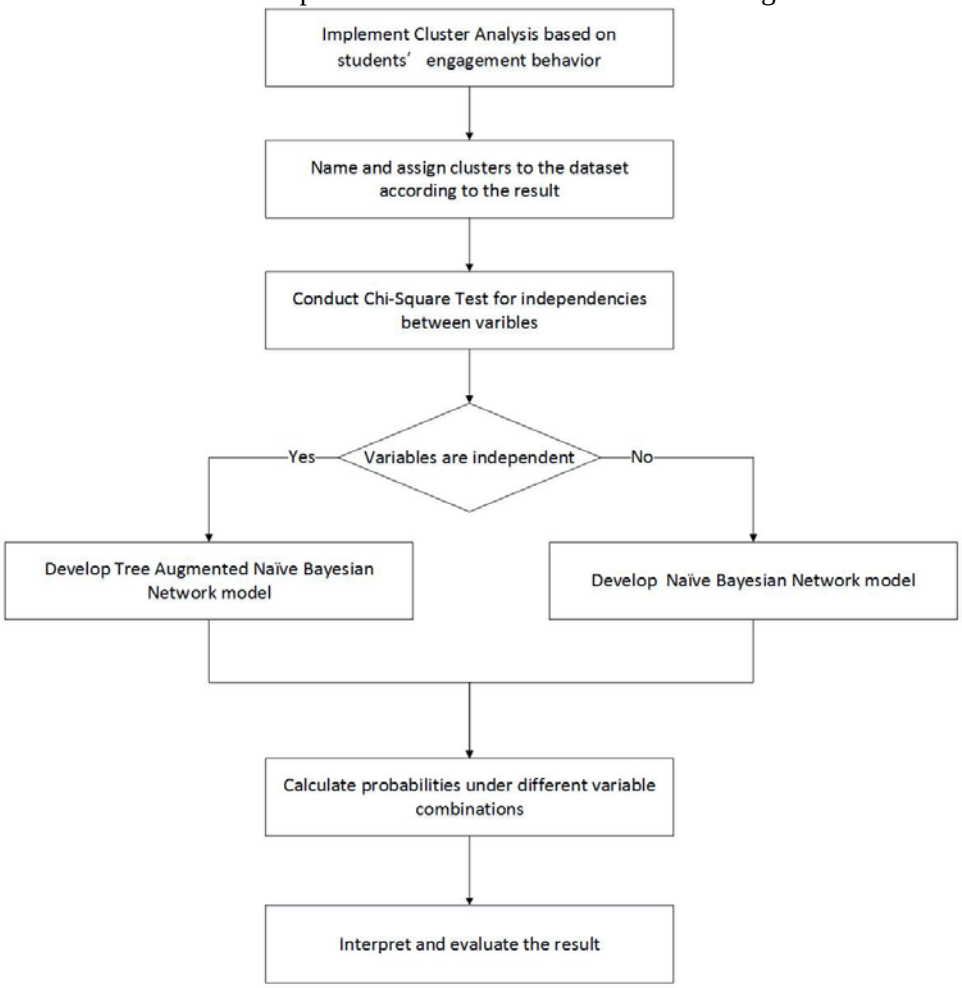

Figure 1. Study Process

- Step1. Cluster Implementation

The first step implements Cluster Analysis with students' engagement behavior which are presented in numeric format.

- Step2. Cluster Names Assignment

Groups are assigned with proper names based on the results where students possess different characteristics.

- Step3. Chi-Square Test

Decide the best model from two Bayesian networks to fit the data by conducting Chi-Square test and finding out probabilistic relationship between these variables of each group.

- Step4. Model Development

Demonstrate the appropriate model by feeding the surveyed data and extract the useful information.

- Step5. Conditional Probability Table

Calculate conditional probabilities of response under different pros and cons combinations and find out top three key variables which have influenced the response the most.

- Step6. Result Interpretation and Evaluation

Interpret and evaluate the result, providing useful and proper suggestions for schools in Taiwan.

3.2. K-means

K-Means algorithm splits the number of observation into k groups. Each group is a cluster which every observation belongs to with the nearest centroid. The performance of this algorithm largely depends on the value of k, and it should be chosen to reflect some characteristics of the data groups under assessment. It is calculated by using the following procedures. For each data i, let:
(a) i is average dissimilarity of i with all other data within the same cluster.
(b) i be the lowest average dissimilarity of i to any cluster, of which i is not a member. It is formulated as:

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \qquad (1)$$

After k value is selected, the algorithm is applied as follows:
(a) Select the initial cluster centers from the given instances randomly equal to the value of k.
(b) Now assign all the instances to the closest cluster center.
(c) Now every cluster center is updated by taking mean of the constituent instances.
(d) After assigning all objects, the position of k centroid is recalculated.
Repeat (b) and (c) steps until there is no further changes in assignment of instances to cluster.

3.3. Bayesian Networks

Bayesian network is a graphical representation of uncertain quantities, which can reveal the probabilistic relationship between a set of variables. The nodes in Bayesian network graph represent the random variables, and the arcs represent causal or probabilistic dependence between the nodes. Conditional probabilities represent likelihoods based on prior information or past experience. The equation is shown below:

$$Posterior\ probability \approx Likelihood \times Prior\ probability \qquad (2)$$

A node with no parents also has a probability table, but it consists only of prior probabilities. Given A is an evidence and B is an event of A observed, the probability of B conditioned by A is noted P(B|A). Bayes theorem describes probabilistic dependence between A and B as follows:

$$P(B|A) = \frac{P(A|B)\cdot P(B)}{P(A)} \qquad (3)$$

Bayesian Network joint probability function is shown below:

$$P(A\ /B,C) = P(A|B,C)\cdot P(B|C)\cdot P(C) \qquad (4)$$

The Tree Augmented Naïve Bayesian Network (Friedman, Geiger, & Goldszmidt, 1997) address the problem that Bayesian networks requires searching all possible nodes by approximating the interactions with maximum two attributes using a tree structure imposed on the naive Bayesian structure.

4 Case Study

For the past few years, National Chiao Tung University (NCTU) has developed online courses actively. Various sections have been established such as OpenCourseWare (OCW) and MOOCs platform (which is called “ewant”). NCTU hopes to offer more general courses on these platforms in the future so that they designed a questionnaire of general courses for college students. The school would like to figure out students’ learning behavior and satisfaction for

general courses by with surveys in order to make an informative decision on traditional general courses digitalization. Data description of the questionnaire are as follows: Amount: 262; Male: 135, Female: 127; Freshmen: 69, Sophomore: 71, Junior: 51, Senior: 71; College of Engineering: 39, College of Science: 33, College of Computer Science: 31, College of Electrical Engineering: 60, College of Biotechnology: 11, College of Management: 56, College of Humanities and Social Sciences: 12, College of Hakka Cultural: 20. Cronbach α is used as the reliability index and α=0.91 means this survey is reliable. This study also tests the survey by factor analysis, each factor loadings is greater than 0.5, proving this survey is valid. Also, feasibility of online general courses is discussed in the questionnaire.

4.1. Cluster Analysis

Table 1. Notation of pros and cons of taking online course

Pros & cons of taking online course	**Notation**
Able to learn anytime	P1
No size limit of classroom	P2
Flexible video playing control (pause, fast forward, reverse)	P3
Capable of playing videos repeatedly	P4
Able to learn anywhere	P5
Able to choose the interesting chapters	P6
Class won't be off due to insufficient registered members	P7
Will be behind the schedule due to laziness	C1
Lack of interaction between students and teachers	C2
The doubts of absence increase	C3
Course videos may not update annually	C4
No chance to get along with other students	C5
Students may not learn the whole content of class out of preferences	C6
The ways to upload the homework and report are limited	C7

The existing survey is used as the data in this section. This study divides the data into homogeneous subgroups by cluster analysis. Then uses Bayesian networks to find out the crucial latent factors. K-Means is demonstrated as a nonhierarchical cluster algorithm in this study. Setting students' engagement as the attributes, this study divides them into two groups. Each cluster's center is calculated after several iterations where two initial centers are randomly assigned. Table 2 shows the results of k-means algorithm. However, Table 3 shows attributes "Attendance" and "Extracurricular" are relatively insensitive with small mean absolute differences. Therefore, three attributes left are considered to be the discriminant variables in cluster analysis.

Table 2. Values of cluster centers

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

Table 3. Mean absolute differences of each attribute

Attendence	Distraction	InvestTime_HW	PreparingTime_Exam	Extracurricular
0.3486303	1.0434438	1.5475505	0.5662824	0.2428334

Two personal characteristics of students are defined based on the result of cluster analysis. Table 4 shows centers of three attributes for two groups. With higher average distraction scores, students in the first group doesn't seem so concentrated in class. They tend to spend less time working on homework and exams. Therefore, this group is named "Slack Students". In the second group, students secure less distraction scores in class. They also take more time on homework and exams than group one. However, they are exactly on par with other normal students who have adequate learning attitude. Therefore, this group is named "Ordinary Students" in this research.

Table 4. Final values of cluster centers

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

4.2. Chi-square Independence Test

The study applies Chi-square test for each group in order to investigate probabilistic independence between variables, referring to Table 5.

Table 5. Chi-square test for each group

	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

Table 5 shows variables are not independent of each other with significant p-values so that this study selects Tree Augmented Naïve Bayesian Network as final used model.

4.3. Bayesian network development

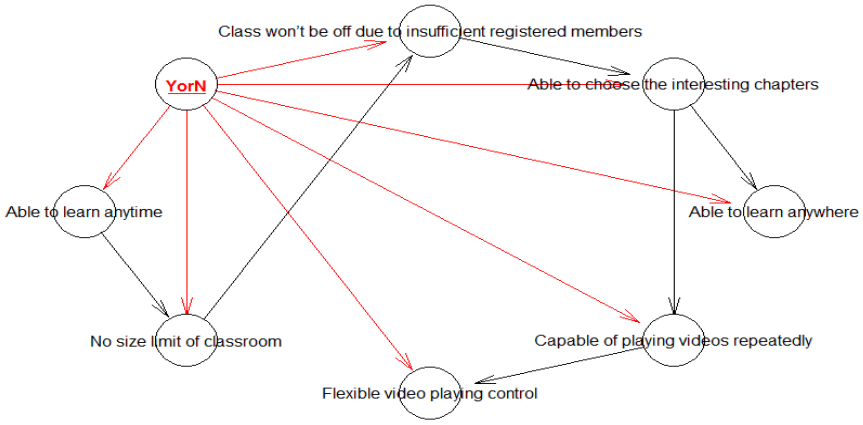

Figure 2. Bayesian network of online-course advantages

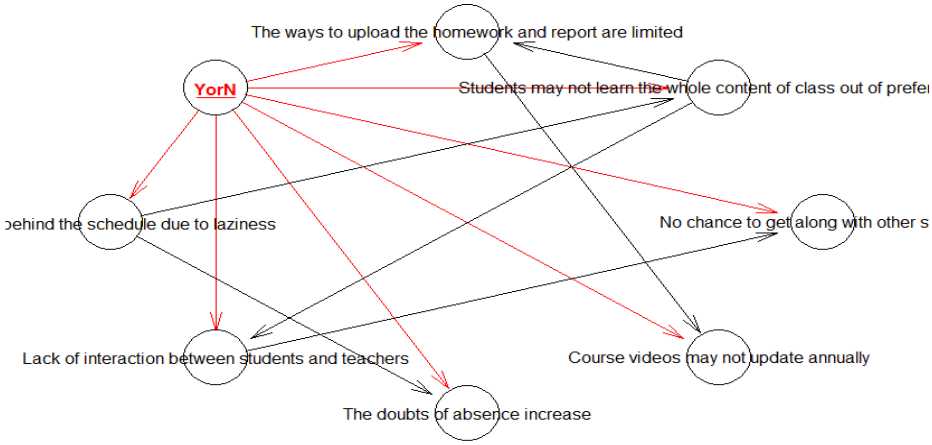

Figure 3. Bayesian network of online-course disadvantages

Figure 2 and 3 show the causal relationships with respect to advantages and disadvantages of taking online course. Every link represents that an end node which is pointed has a variable condition on a start node. For example, node “Able to learn anytime” on figure 2 has the variable condition on node “No size limit of classroom”. Every node can be connected with maximum two variable conditions and the decision variable “YorN” is the given evidence for other nodes in TAN Bayesian network model which is improved to avoid computation complexity.

Table 6. Combination Probability Table

P1	P2	P3	P4	P5	P6	P7	YorN
1	1	0	0	0	0	0	0.1628
1	0	1	0	0	0	0	0.3795
1	0	0	1	0	0	0	0.4702
1	0	0	0	1	0	0	0.9470
1	0	0	0	0	1	0	0.6202
1	0	0	0	0	0	1	0.3424
0	1	1	0	0	0	0	0.0738
0	1	0	1	0	0	0	0.0294

0	1	0	0	1	0	0	0.3943
0	1	0	0	0	1	0	0.0543
0	1	0	0	0	0	1	0.1247
0	0	1	1	0	0	0	0.3265
0	0	1	0	1	0	0	0.9088
0	0	1	0	0	1	0	0.4750
0	0	1	0	0	0	1	0.2253
0	0	0	1	1	0	0	0.7826
0	0	0	1	0	1	0	0.5732
0	0	0	1	0	0	1	0.2859
0	0	0	0	1	1	0	0.9653
0	0	0	0	1	0	1	0.7367
0	0	0	0	0	1	1	0.3034

The conditional probabilities of intentions for students to take online courses instead of traditional general education courses are shown in Table 6 which takes group one and pros for example. Each event is given maximum two evidences and the likelihood of observation evidences can be obtained. Also, this study applies TAN Bayesian network with independent relationship assumptions within variables abandoned. Selecting two pros from seven, twenty-one combinations of conditional probabilities of intentions of taking online courses are presented.

4.4. Interpretation and Evaluation

In Table 6, top 3 conditional probabilities of response "YorN" are selected and the corresponding pros combinations are identified as well. To evaluate the results of Bayesian network and provide advice about traditional general education course digitalization, this study compares corresponding satisfaction items from students for current traditional general education courses. Four possible suggestions in Table 8 are come up with. For example, based on Table 6, P1 and P5 in the fourth row cause the second highest response which means these two pros of online course make students more likely to take MOOCs. This study further compares the result to the corresponding level of scores in Table 7 and finds that the corresponding satisfaction level of traditional courses is low. In the end, refer to Table 8, the advice "MOOC is better" is selected and it also implies if the key pros P1 and P5 exist, students tend to take MOOCs instead of traditional ones.

Table 7. Table of Satisfaction Items in Traditional General Education Courses

Satisfaction Items		Avg. Satisfaction	Level
--	C2	3.125954	Low
P5	C7	2.816794	Low
P2	--	2.477099	Low
P1	C1	3.217557	Low
P5	--	3.01145	Low
--	C2	3.232824	High
--	C2	3.526718	High

P3, P6, P7,	C2, C3	3.301527	High
P6	--	3.80916	High
--	--	2.942748	Low
--	--	3.122137	Low
--	C6	3.477099	High
P6	C3, C4	3.610687	High
P6, P7	C5, C7	3.259542	High
P4	--	2.881679	Low
P2	--	3.232824	High

Table 8. Table of Possible Suggestions

Traditional/ Online	Pros	Cons
High	Both are OK	Remain traditional
Low	MOOC is better	Further discussion

Table 9. Table of Final Results

Pros Group "Slack Students"			
		Probability	**Suggestions**
P3	P6	0.905	Both are OK
P3	P5	0.739	Both are OK
P4	P5	0.683	MOOCs are better
Pros Group "Ordinary Students"			
		Probability	**Suggestions**
P6	P5	0.965	Both are OK
P1	P5	0.947	MOOCs are better
P3	P5	0.909	MOOCs are better
Cons Group "Slack Students"			
		Probability	**Suggestions**
C1	C7	0.682	Further discussion
C6	C7	0.677	Remain traditional
C4	C6	0.589	Remain traditional
Cons Group "Ordinary Students"			
		Probability	**Suggestions**
C5	C6	0.748	Remain traditional
C5	C7	0.712	Further discussion
C5	C1	0.649	Remain traditional

This study implements the approach to each cluster and summarizes the results in Table 9. It shows that P5 and C5 dominate the key pros and cons of conditional probabilities. The students who have adequate engagement of courses tend to choose MOOCs rather than traditional courses. However, students support traditional courses if these key disadvantages exist.

5 Conclusion

This study demonstrates Bayesian network for identifying key advantages and disadvantages based on dependent probabilistic relationships. By developing the comparison of satisfaction level of traditional general education courses and key indicators of taking online courses, corresponding results are presented in each group of students whose class engagement behavior are different. Eventually, these results can provide practical advice for college faculties whether

to digitalize certain traditional general education courses and make online course credits verified. In this study, all analysis is based on existed survey data. In the future, different perspectives are expected to be included for obtaining more complete results.

6 Reference

Ben-Gal, I. (2007). “Bayesian networks”. Encyclopedia of statistics in quality and reliability.

Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., & Seaton, D. T. (2013). “Studying learning in the worldwide classroom: Research into edX's first MOOC”. *Research & Practice in Assessment*, *8*.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). “Bayesian network classifiers”. *Machine learning*, 29(2-3), 131-163.

Gil, P., Jara, C. A., Candelas, F. A., & García, G. J. (2012). “Experiences with free and open courses using on-line multimedia resources”. In *Proceedings of 3th IEEE International Conference on Engineering Education* (pp. 5-10).

Hani, H., Hooshmand, H., & Mirafzal, S. (2013). “Identifying the factors affecting the success and failure of e-learning students using cluster analysis”. In *e-Commerce in Developing Countries: With Focus on e-Security (ECDC), 2013 7th Intenational Conference on* (pp. 1-12). IEEE.

Holotescu, Carmen, et al. “The Power of the Three Words and One Acronym: OER vs OER: Subtitle: I’m not an Ogre of the Enchanted Realm (of cyberspace). I’m an Omnipresent Educational Rescuer (because I use the OER!)”. *Procedia-Social and Behavioral Sciences* 191 (2015): 2531-2536.

Huang, J., & Fang, J. (2009, June). “Research on Courses Relationship Model Based on Bayesian Networks”. In *Computational Intelligence and Natural Computing, 2009. CINC'09. International Conference on* (Vol. 2, pp. 15-18). IEEE.

Itoh, H., Nishiwaki, M., & Funahashi, K. (2015, October). “A method of identifying students who require guidance using Bayesian network”. In *2015 IEEE 4th Global Conference on Consumer Electronics (GCCE)* (pp. 283-287). IEEE.

Kay, J., Reimann, P., Diebold, E., & Kummerfeld, B. (2013). “MOOCs: So many learners, so much potential”. Technology, 52(1), 49-67.

Margaryan, A., Bianco, M., & Littlejohn, A. (2015). “Instructional quality of massive open online courses (MOOCs)”. *Computers & Education*, *80*, 77-83.

Picciano, A. G. (2002). “Beyond student perceptions: Issues of interaction, presence, and performance in an online course”. *Journal of Asynchronous learning networks*, 6(1), 21-40.

Singh, I., Sabitha, A. S., & Bansal, A. (2016, January). Student performance analysis using clustering algorithm. In *Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference* (pp. 294-299). IEEE.

Soong, M. B., Chan, H. C., Chua, B. C., & Loh, K. F. (2001). “Critical success factors for on-line course resources”. *Computers & Education*, 36(2), 101-120.

West, P., Rutstein, D. W., Mislevy, R. J., Liu, J., Choi, Y., Levy, R., . . . Behrens, J. T. (2010). “A Bayesian Network Approach to Modeling Learning Progressions and Task Performance”. CRESST Report 776. National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Xiaorong, H. (2009, November). Research on network-course learning effect evaluation model based on three-layer Bayesian network. In *Intelligent Information Technology Application, 2009. IITA 2009. Third International Symposium on* (Vol. 1, pp. 167-170). IEEE.

Yu, X., Liu, J., Yang, Z. J., Liu, X., Yin, X., & Yi, S. (2016, October). Bayesian Network Based Program Dependence Graph for Fault Localization. In *Software Reliability Engineering Workshops (ISSREW), 2016 IEEE International Symposium on* (pp. 181-188). IEEE.

Zhao, J. W., Gu, S. M., & He, L. (2010, June). A novel approach to clustering access patterns in e-learning environment. In *2010 2nd International Conference on Education Technology and Computer* (Vol. 1, pp. V1-393). IEEE.