Clustering Analysis of Academic Courses Based on LMS Usage Levels and Patterns:

Gaussian Mixture Model, K-Means clustering and Hierarchical clustering

Il-Hyun Jo
Department of Educational Technology
Ewha Womans University
Seoul, Korea
ijo@ewha.ac.kr

Jongwoo Song
Department of Statistics
Ewha Womans University
Seoul, Korea
josong@ewha.ac.kr

Yeonjeong Park
Department of Educational Technology
Ewha Womans University
Seoul, Korea
ypark78@ewha.ac.kr

Hyeyun Lee
Department of Educational Technology
Ewha Womans University
Seoul, Korea
hyeyun521@naver.com

Suyeon Kang
Department of Statistics
Ewha Womans University
Seoul, Korea
korea92721@naver.com

Abstract - **This study tried to find a group of academic courses based on the usage levels and patterns of Learning Management System (LMS) utilized in higher education using clustering techniques. LMS is an essential technology to support virtual learning environment where students have access to the learning materials that their instructors provide, submit the deliverables, and participate in various learning activities involving group projects, discussion forums, quizzes, and Wikis. However, the returns on large investment have not systematically performed in terms of what extent of students and instructors have utilized the system for their teaching and learning. In this study, 2,639 courses opened during 2013 fall semester in a large private university located in South Korea were analyzed with 13 observation variables that represent the characteristics of academic courses. Three clustering methods including Gaussian Mixture Model, K-Means clustering, and Hierarchical clustering contributed to (1) identifying large number of courses that show inactive and no usage of LMS, (2) disclosing the dramatically imbalanced clusters, and (3) identifying several clusters that present different usage patterns of LMS. The results of such *academic analytics* provide meaningful implications for academic leaders and university staff to make strategic decisions on the development of LMS.**

Keywords-Learning Management System; Clustering analysis; Gaussian Mixture Model; K-Means clustering; Hierarchical clustering.

I. INTRODUCTION

Currently, universities are increasing incorporating a Learning Management System (LMS) to support effective teaching and learning [1]. Whether focusing on campus-based learning in higher institute or distance learning, LMS is considered as an essential technology for virtual learning environment on e-learning systems where instructors provide various learning materials such as text, images, URL links and video clips to learners.

A common goal of LMS is to organize and manage different courses within an integrated system [1]. The integrated systems collect each learner's online behavior data in every class. Based on this data educational researchers and practitioners can analyze and interpret learners' status during the semester. University staff or decision makers can leverage such LMS usage trends analytics to derive proper treatment and policies to current learners.

Such a data-driven approach has been attempted in the field of higher education recently with the term of *academic analytics*. It has emerged after the widespread of data mining practices by the influence of business intelligence [2, 3]. This approach has been evaluated as a new tool to respond to increased concerns for accountability in higher education and to develop actionable intelligence to improve student success and learning environment [4]. For example, instructors and academic consultants are better able to understand the learner's learning behavior and performance, even their thoughts based on the rich data. Further, the academic analytics can help more strategic investment and development in a way to fulfill the needs of students and instructors based on the informed analytic results via the pattern-recognition, classification, and prediction algorithms of [5].

The data analytics in education has helped to develop prediction models for academic success of learners based on their behaviors and participation or identifying at-risk students for special guidance from their faculty and advisors [6, 7]. However, the previous applications of analytics have disclosed a further research to apply the elaborated analysis and develop more precise prediction models to prevent the drawbacks from the wrong feedbacks to students [8]. Therefore, as a preliminary research, this study highlights the need of the examinations of current usages and patterns of LMS. Instead of analyzing the individual student level data, the academic course data as a unit of analysis was utilized. We argue that without the thorough analysis on LMS usages and patterns and accurate clustering of the courses, it would not be able to build elaborated prediction models to estimate students' success and failure based on the online behavior records in LMS.

The data sets utilized for academic analytics can be diverse depending on the characteristics of institutions [5]. Not only the aforementioned LMS but also course management system (CMS), audience response systems, library systems are the examples. In this study, we utilized LMS dataset to analyze students' virtual learning behaviors and CMS data to collect the academic course's general information. By using both LMS and CMS data, the

 ISBN: 978-1-61208-423-7

clustering analysis of academic courses on the basis of virtual learning environment usage levels and patterns were synergistically performed. For the rigorousness and thoroughness on data analytics we employed multiple methods of clustering analysis: Gaussian Mixture Model, K-Means clustering and Hierarchical clustering. The specific research questions were as follows:

RQ1) To what extent have instructors and students utilized LMS for their teaching and learning?

RQ2) What clusters are formed as the patterns of LMS usages?

RQ3) How does clustering analysis detect academic courses that present inactive usage of LMS or unique LMS usage patterns?

II. METHODOLOGY

This study aimed to find a group of classes that are as homogenous as possible within group (cluster) and as inhomogeneous as possible between groups (clusters) based on their online activities and class sizes.

A. Research Context

The context of this study was a private university located in Seoul, Korea. With the supports of institution for teaching and learning in the university, we collected academic course data of the year of 2013 fall semester. All courses were opened using Moodle-based virtual learning environment regardless of the course type such as offline and online. Consequently, total 4,416 courses were analyzed at the initial data analysis step. However, since it was revealed that many courses did not use online campus, the exclusion of such non-active courses were performed. Finally, 2,639 courses were observed for this study with 13 variables.

A data set for analysis was prepared by combining two databases: CMS and LMS. CMS dataset contained course-related information indicating each student's hierarchical categorizations (i.e., graduate VS. undergraduate, mandatory VS. selective, affiliated colleges and department) and LMS dataset included online behavior tracks (i.e., total number of resources, notices, lecture notes, submissions, group works etc.). We integrated CMS and LMS dataset, and these data were divided in general indicator and activity-based indicator. Table I shows a total of 13 variables.

TABLE I. VARIABLE SUMMARY

	No.	Variable name	Variable explanation
General Indicator	1	**MEM**	Number of members
	2	**FRE**	Average log-in frequency per person
	3	**ACT**	Number of activity items
Activity-based Indicator	4	**RES**	Number of resources
	5	**NOT**	Number of notice
	6	**QNA**	Number of questions and answers
	7	**LEC**	Number of lecture notes
	8	**SUB**	Number of task submissions
	9	**GRO**	Number of group works
	10	**LIN**	Number of links
	11	**POS**	Number of discussion forum postings
	12	**QUI**	Number of Quiz
	13	**WIK**	Number of Wikis

B. Clustering Methods

1) Gaussian Mixture Model

GMM is a probabilistic model that assumes all data are from the mixture of normal distributions. The variables must be numeric since we assume that the data are from the multivariate normal distribution. The parameters (the proportion of each group, mean vectors, and variance matrix) are estimated by EM algorithm. In general, the number of clusters is very hard to estimate in the clustering analysis. However, we can estimate the optimal number of clusters in GMM using the Bayesian Information Criterion (BIC). We use the R-package "mclust" for GMM. The mclust package in R can estimate not only the number of clusters but also the optimal form of variance matrix. We will use the number of clusters from the GMM for the K-means and the hierarchical clustering, too.

2) K-means clustering

K-means clustering is one of the most popular clustering method because it is very fast to find clusters and very easy to understand. The objective function of K-means clustering is to minimize the sum of within scatters. Basically, it tries to find the k group that minimizes within-cluster sum of squares; therefore it maximizes the between-cluster sum of squares. Since it uses the squared Euclidean distances among the objects and the cluster centers are defined as the means of objects in each cluster, all variables must be numeric. We use K-means function in R for our analysis.

3) Hierarchical clustering

Hierarchical clustering method is used for building a hierarchy of clusters from data. Strategies for this clustering fall into two types: agglomerative for "bottom-up" approach and divisive for "top-down" approach. We use a bottom-up approach in this article. The algorithm finds the nearest two objects and merges them. It repeats this process until all objects are in one cluster. The final results are usually represented by the dendrogram. The hierarchical clustering methods can give different results depending on which distance metric we use between groups. There are several distance metrics between groups and we use the "complete-linkage" in our analysis. The "complete-linkage" is the maximum distance between two groups and it is known that the "complete-linkage" can find the compact clusters. We use "hclust" function in R for our analysis.

 ISBN: 978-1-61208-423-7

III. RESULTS

A. Descriptive Statistics

Before going to the clustering analysis, we examined descriptive statistics to find out the distribution of observations.

TABLE II. DESCRIPTIVE STATISTICS OF 2,639 COURSES

Name	Min	Max	Mean	SD	Skewness	Kurtosis
MEM	2	301	33.22	33.66	2.97	13.00
FRE	2	375	39.75	33.01	2.50	11.05
ACT	1	8	2.49	1.30	0.93	0.78
RES	0	**596**	11.87	21.49	12.22	263.56
NOT	0	132	6.64	9.26	3.21	20.15
QNA	0	**280**	2.95	14.25	12.09	183.95
LEC	0	176	3.74	9.69	5.16	51.87
SUB	0	36	0.95	2.82	4.97	32.73
GRO	0	**1612**	17.52	88.42	8.23	91.71
LIN	0	72	0.32	2.57	14.97	312.54
POS	0	**2810**	6.45	75.32	24.44	788.77
QUI	0	**215**	0.61	8.34	17.82	366.00
WIK	0	15	0.01	0.31	42.92	2005.64

As shown in Table 2, most variables have extremely high values. For example, the maximum values of variables indicate that 596 resources (RES), 176 lecture notes (LEC), 1612 board postings of group works (GRO), 2,810 discussion postings (POS), and 215 Q&A postings (QNA). These values present extremely high utilization level of few courses.

On closer inspection, one course which posted 2,810 forum discussion postings was big-sized basic requirement course and there were over a hundred students who signed up for class. There were 11 groups and they discussed enthusiastically with each other, so such very high postings were possible. Next, the other course which had 1,612 group works was the major course of educational technology and the instructor assigned team project during the semester. There were 10 groups and they used group board for team-based learning. Because they uploaded all the related materials for project, opinions and chatting messages in group board, so this high value was also possible. These cases looked as errors but it tells the 'real aspects' of unique courses.

Furthermore, the data were sparse by showing many observations with zero values. The variables from QNA to WIK have zero values for more than 50% of data. We can predict that there will be a single one big cluster with a lot of 'zero' observations. This one big cluster will have all the classes with minimal online activities. This cluster was not our interest but we were more interested in other clusters of small size and how different they are.

B. Gaussian Mixture Model

As Figure 1 indicates, Mclust finds the best model is three clusters with EEV (ellipsoidal, equal volume and shape covariance). In the point of three components (clusters), the increase of BIC starts decrease. However, four-cluster model is also close. Thus, we decided to investigate both three-cluster and four-cluster model.

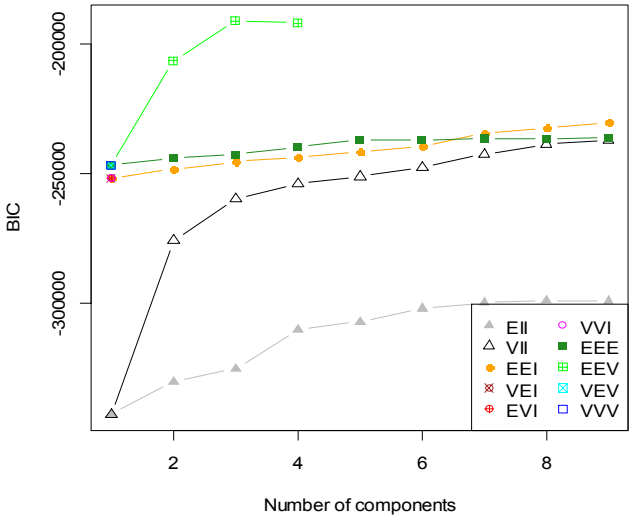

Figure 1.Results of EEV in Mclust

1) GMM with three clusters

The size of three clusters were 212, 2,360, and 67 respectively as seen in Table III.

TABLE III. CLUSTERING TABLE WITH THREE CLUSTERS

	Cluster 1	Cluster 2	Cluster 3
Number of Class	212	2360	67
Mixing Probability	0.08068	0.89393	0.02539

We checked the mean vectors (cluster centers) of three clusters. As Figure 2 indicates, cluster 3 (size 67, green line) has the higher mean values (more online activities) and cluster 1 (size 212, black line) is in the middle, and cluster 2 (size 2,360, red line) has the least online activities. On a closer view, cluster 3 has greatly high value of POS and cluster 1 has high GRO value.

 ISBN: 978-1-61208-423-7

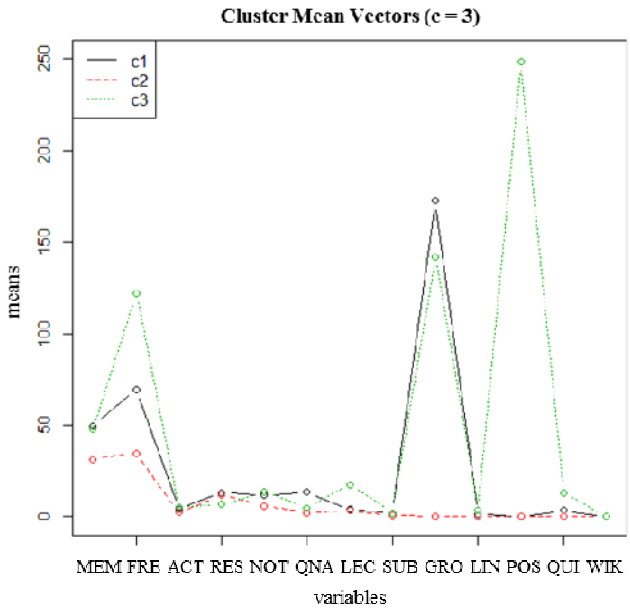

Figure 2. Mean vector plot of three clusters

Look inside the clustering table, 2,360 out of total 2,639 classes were included in cluster 2 where having at least online activities. They were inactive classes. Approximately 89% of total class did marginal performance at online campus. On the other hand, cluster 3 was the most active online classes. We can guess that these classes were actively discussed about their topic since both number of forum discussion postings and average log-in frequency per person are quite high. The rest courses in cluster 1 also participated in group work much but the average frequency mean is in-between cluster 2 and 3. This cluster is specialized in team project.

2) GMM with four clusters

We divided total classes into four clusters this time. The size of four clusters were 71, 2,322, 230, and 16 as sheen in Table IV.

TABLE IV. CLUSTERING TABLE WITH FOUR CLUSTERS

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Number of Class	71	2322	230	16
Mixing Probability	0.02705	0.87962	0.08727	0.00606

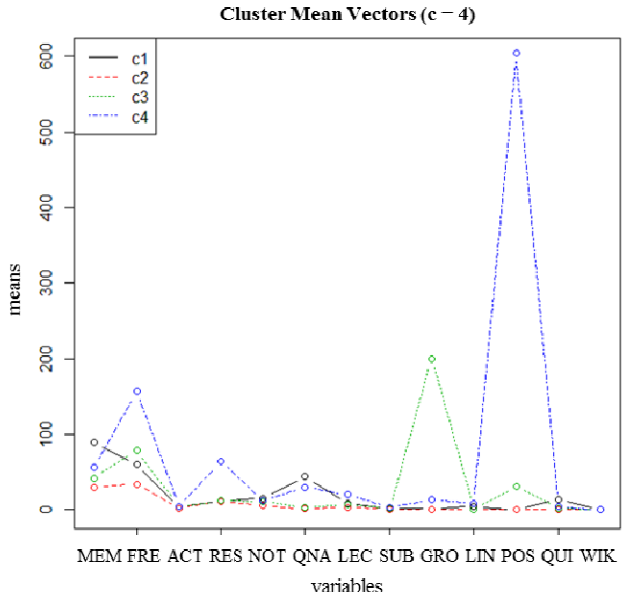

Figure 3. Mean vector plot of four clusters

When reviewing the cluster mean vector plot in Figure 3, cluster 4 (size 16, blue line) has extremely high mean values in POS while cluster 2 (size 2,322, red line) has low mean values in general. Cluster 4 shows the equal appearance with the cluster 3 in GMM with three clustering analysis. Moreover, in common with GMM with three clustering results, 9th variable (GRO) is shown the highest value in cluster 3 (size 230, green line), not the cluster 4 which has higher values in the gross. Courses which involved in cluster 3 were inactive in most of online activities except group works. Newly-drawn cluster 1 (size 71, black line) has the highest MEM value and it represents number of members including an instructor, teaching assistant and students. We are able to call its name, 'big-sized courses'.

The last thing we should observe carefully is that when we clustered total courses into four clusters using GMM, number of courses with highly active in online activities such as forum discussion postings and log-in frequency were decreased from 67 (see Table III) to 16(see Table IV).

C. K-means clustering

In addition to GMM, we also performed a clustering analysis using K-Means. As a first step, we analyzed with non-standardized dataset to see overall clusters and compare the results with GMM. However, due to the large scale differences among variables, we also conducted clustering with standardized dataset because we like to see the clustering results when all variables have the similar contributions in distances.

1) Using non-standardized data

The results of K-means clustering with non-standardized data showed similar results with GMM analysis. But this process was meaningful because the results identified fewer active online courses.

a) K-means clustering with three clusters

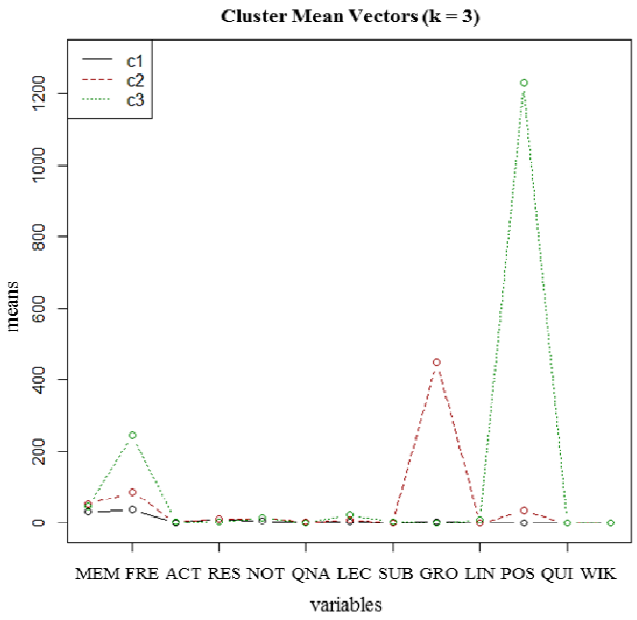

Figure 4. Mean vector plot of three clusters

Most of mean vector values about learners' online behavior were quite similar, similarly low but FRE, GRO, POS variables were distinguished among clusters. Learners who were included in cluster 3 (size 6, green line) classes logged LMS in the most frequently and wrote up the postings on the forum very much. Cluster 2 (size 71, red line) has high value of GRO which means group works. Cluster 1 (size 2,562, black line) which the most of classes were in has less online action.

TABLE V. CLUSTERING TABLE WITH THREE CLUSTERS

	Cluster 1	Cluster 2	Cluster 3
Number of Class	2562	71	6
Mixing Probability	0.97082	0.02690	0.00227

Six courses included in cluster 3 are listed on Table VI. They were super active classes in university. As shown in mean vector plot on Figure 4, these courses have high value of log-in frequency (FRE) and forum discussion postings (POS).

TABLE VI. DETAILED VARIABLE VALUES OF CLUSTER 3 COURSES

No.	MEM	FRE	ACT	RES	NOT	QNA	LEC	SUB	GRO	LIN	POS	QUI	WIK
255	103	**167**	7	8	71	7	37	3	0	0	**2810**	0	0
894	43	**264**	4	0	0	0	7	14	0	16	**991**	0	0
1299	30	**375**	3	0	0	0	27	0	0	14	**944**	0	0
1403	37	**204**	4	0	0	0	62	1	0	23	**715**	0	0
1630	46	**217**	8	2	22	1	13	12	0	1	**1297**	0	3
2049	18	**245**	4	13	5	1	0	0	0	0	**638**	0	0
M	46	**245**	5	4	16	2	24	5	0	9	**1233**	0	0

b) K-means clustering with four clusters

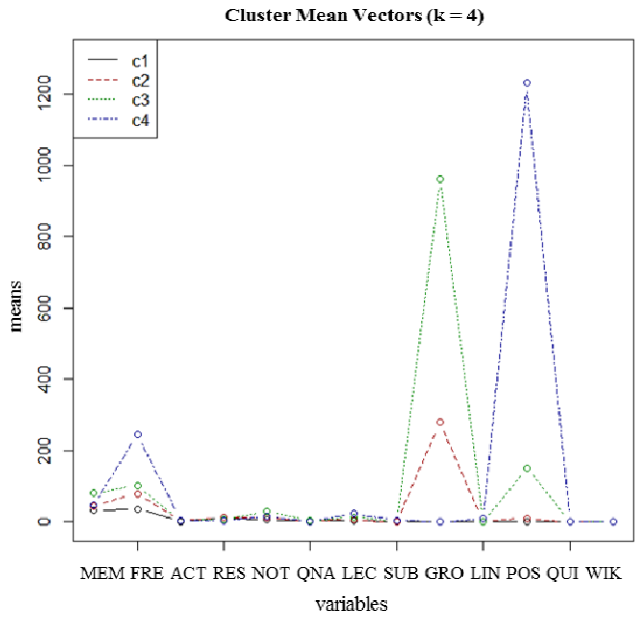

Figure 5. Mean vector plot of four clusters

When we were partitioning total courses into four clusters, cluster 3 and 4 were somewhat unique. Cluster 3 (size 11, green line) has high value of GRO variable and cluster 4 (size 6, blue line) is shown much online action in FRE and POS variables. As mentioned earlier, students in cluster 4 courses discussed with one another constantly and this fact can be proved by FRE and POS. Like the preceding, cluster 3 performed intensive group works. Newly created cluster 2 (size 109) compared to previous results was shown the middle activeness in LMS.

TABLE VII. CLUSTERING TABLE WITH FOUR CLUSTERS

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Number of Class	2513	109	11	6
Mixing Probability	0.95225	0.04130	0.00417	0.00227

Six classes included in cluster 4 are exactly the same courses with the cluster 3 in K-means clustering with three clusters analysis (see Table VI).

 ISBN: 978-1-61208-423-7

2) Using standardized data

Prior to clustering data, we rescaled variables for comparability. So standardized data was utilized in this step. It showed quite different figures in mean vector plots for the plot of non-standardized dataset. Since K-means uses the squared Euclidean distance, the outliers can affect the clustering results significantly. However, if we use the standardized dataset, then the effect of outliers will be reduced, therefore it is unlikely to see very small sized clusters.

a) K-means with three clusters

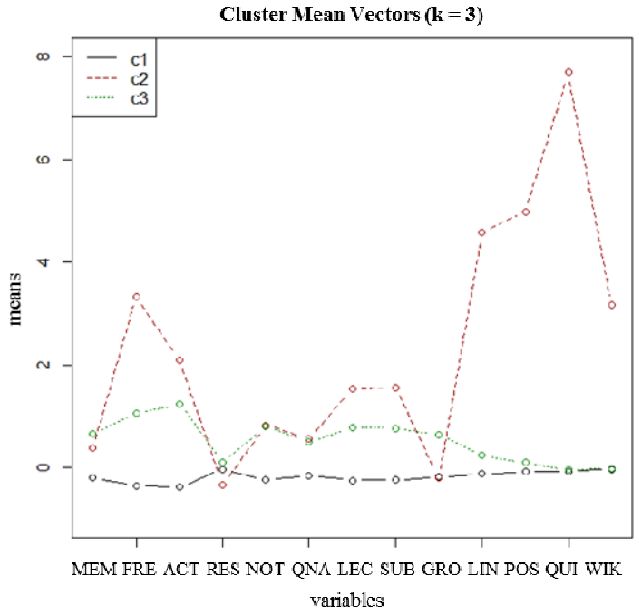

Figure 6. Mean vector plot of three clusters

As shown in Figure 6, cluster 2 (size 22, red line) has high mean vector value on the whole. FRE, ACT, LEC, SUB, LIN, POS, QUI and WIK values of cluster 2 were high. Among these, those courses used quiz function very frequently, so QUI was shown excessive activity log in comparison with other clusters.

TABLE VIII. CLUSTERING TABLE WITH THREE CLUSTERS

	Cluster 1	Cluster 2	Cluster 3
Number of Class	2030	22	587
Mixing Probability	0.76923	0.00834	0.22243

Cluster 1 (size 2,030, black line), about 77% of courses contained, was shown the low activeness in general without exception. However, cluster 2 was generally active. Cluster 3 (size 587, green line) was middle-active according to the LMS usage levels, but it had top-of-the-line value in MEM, RES and GRO. In contrast with non-standardized clustering results, these clusters were distinguished by the level of usage, not the unique extreme values.

b) K-means with four clusters

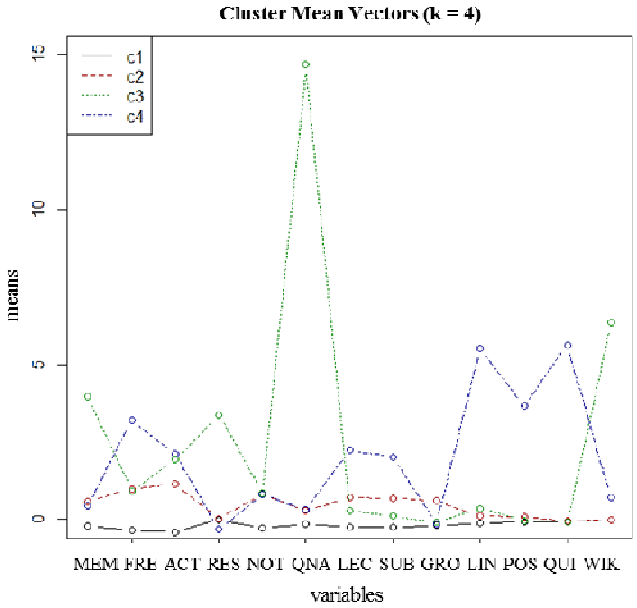

Figure 7. Mean vector plot of four clusters

Cluster 3 (size 8, green line) had unusually high mean vector values in QNA, compared to other clusters. Moreover, such MEM, RES and WIK values were also high. We can interpret this situation that there were many members in class, so lots of questions came out together. Another cluster 4 (size 30, blue line) has high action value in FRE, ACT, LEC, SUB, LIN, POS and QUI. In other words, students eagerly participated in LMS in average since the average log-in frequency per person value was the biggest among other cluster. Furthermore, we could assume that the courses provided both a great deal of course-related materials and the grade-related assignment. High values of SUB (number of task submission) and QUI (number of quiz) as well as LEC (number of lecture notes) and LIN (number of URL links) are the evidences. However, cluster 1's (size 1,979, black line) action was minor despite it took most of virtual learning environment courses. Likewise the previous analytic results of cluster 3 (size 587, green line), cluster 2 (size 622, red line) was shown the middle activeness.

TABLE IX. CLUSTERING TABLE WITH FOUR CLUSTERS

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Number of Class	1979	622	8	30
Mixing Probability	0.74991	0.23570	0.00303	0.01137

 ISBN: 978-1-61208-423-7

D. Hierarchical clustering

Lastly, we analyzed academic courses with hierarchical clustering method. Standardized dataset was used to clustering.

1) Hierarchical clustering with three clusters

TABLE X. CLUSTERING TABLE WITH THREE CLUSTERS

	Cluster 1	Cluster 2	Cluster 3
Number of Class	2637	1	1
Mixing Probability	0.99924	0.00038	0.00038

The result of hierarchical clustering displays an unprecedented appearance. The only 158th class in Table XI came under cluster 2 (size 1) and a 255th class in Table XII was included in cluster 3 (size 1). Except those two certain classes, the rest of courses were clustered together in cluster 1 (size 2,637).

TABLE XI. DETAILED VARIABLE VALUES OF CLUSTER 2 COURSE

Class No.	MEM	FRE	ACT	RES	NOT	QNA	LEC	SUB	GRO	LIN	POS	QUI	WIK
158	144	93	7	19	8	108	29	0	0	9	30	0	**15**

158th course utilized many activity items (ACT = 7) in moderate way and interestingly used Wiki function in its course. It was the course of economics department. Actually, 15 times was not that huge usage number but as almost the whole courses had not used Wiki (M = .01, SD = .31), this class was chosen for the sole course in cluster 2 because of WIK.

TABLE XII. DETAILED VARIABLE VALUES OF CLUSTER 3 COURSE

Class No.	MEM	FRE	ACT	RES	NOT	QNA	LEC	SUB	GRO	LIN	POS	QUI	WIK
255	103	167	7	8	71	7	37	3	0	0	**2810**	0	0

255th class represents extremely high value of forum discussion postings. This course also utilized many activity items (ACT = 7) and specifically in POS, it showed unparalleled usage. It was possible because there were lots of members in class. Every person uploaded 27.28 postings averagely and it would be an acceptable number.

2) Hierarchical clustering with four clusters

TABLE XIII. CLUSTERING TABLE WITH FOUR CLUSTERS

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Number of Class	2634	1	1	3
Mixing Probability	0.99811	0.00038	0.00038	0.00114

Four clusters analytic result was pretty similar with those *three* clusters hierarchical clustering. Cluster 2 and 3 courses (158th and 255th class) were the same with the previous result. However, newly created cluster 4 (size 3) differed from the previous one. Three classes out of 2,637 courses had high mean value in RES (number of resources).

TABLE XIV. DETAILED VARIABLE VALUES OF CLUSTER 4 COURSES

Class No.	MEM	FRE	ACT	RES	NOT	QNA	LEC	SUB	GRO	LIN	POS	QUI	WIK
514	46	49	2	**596**	0	227	0	0	0	0	0	0	0
1151	84	58	4	**276**	44	19	0	1	0	0	0	0	0
1557	52	83	4	**401**	5	1	29	0	0	0	0	0	0
Mean	60.67	63.33	3.33	**424.33**	16.33	82.33	9.67	0.33	0.00	0.00	0.00	0.00	0.00

These three courses did not utilized many activities so the variables from SUB to WIK got almost zero value. Specifically, courses had the highest RES values. We can interpret that instructors in these courses chose the resource application instructional method and provided many useful resources for the subject.

3) Hierarchical clustering with five clusters

TABLE XV. CLUSTERING TABLE WITH FIVE CLUSTERS

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Number of Class	2629	5	1	1	3
Mixing Probability	0.99621	0.00189	0.00038	0.00038	0.00114

Cluster 3 (size 1), 4 (size 1) and 5 (size 3) were same with cluster 2, 3 and 4 in hierarchical clustering with *four* clusters results. Cluster 2 (size 5) was broken loose from cluster 1 (size 2,629) and five classes in Table XVI were included.

TABLE XVI. DETAILED VARIABLE VALUES OF CLUSTER 2 COURSES

Class No.	MEM	FRE	ACT	RES	NOT	QNA	LEC	SUB	GRO	LIN	POS	QUI	WIK
31	183	49	4	0	10	6	22	0	0	**33**	0	0	0
571	103	59	5	12	6	12	0	2	0	**31**	0	0	0
594	101	52	5	21	24	18	0	1	0	**30**	0	0	0
1243	46	163	7	0	8	8	41	5	0	**48**	201	28	0
1694	55	52	4	0	12	0	33	1	0	**72**	0	0	0
Mean	97.6	75.0	5.0	6.6	12.0	8.8	19.2	1.8	0.0	**42.8**	40.2	5.6	0.0

These five classes had actively shared useful URL links during the semester as a course material. Mainly, instructors provided references from the web in big-sized courses.

4) Hierarchical clustering with six clusters

TABLE XVII. CLUSTERING TABLE WITH SIX CLUSTERS

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Number of Class	2620	5	9	1	1	3
Mixing Probability	0.99280	0.00189	0.00341	0.00038	0.00038	0.00114

 ISBN: 978-1-61208-423-7

This case, cluster 2 (size 5), 4 (size 1), 5 (size 1) and 6 (size 3) took on an exactly the same aspect with *five* clusters hierarchical clustering. All the courses which were included in each cluster were the same. Some courses in previous cluster 1 (size 2,620) were divided into two clusters in here as cluster 1 and 3 (size 9).

IV. DISCUSSION AND CONCLUSION

The purpose of study was to cluster academic courses in higher education in accordance with virtual learning environment usage levels and patterns. For this goal, we employed three methodologies: Gaussian mixture model, K-Means clustering and hierarchical clustering and could draw several implications.

The results of this study found that clusters were considerably imbalanced. Descriptive statistics revealed that outliers from dataset were so abnormally high in some variables. On the other hand, most of values were quite low and most of them were zero. This initial data condition led to disproportionate result and this would be the reason that some enthusiastic courses continuously came up. It may not be the cluster in which decision makers wanted to see from LMS usage patterns in higher education institute. However, this real combined data and the results emphasize the true status quo. We can infer that instructors do not know much how to use virtual learning environment well and they might have a hard time to facilitate the LMS use. It is time for academic leaders and university decision makers to form a practical plan which can improve utilization in a balanced way.

Nevertheless, this study revealed that certain enthusiastic courses were drawn out repeatedly when using these three clustering methods. Since such courses had unique characteristics distinguishing from other courses, this study suggests a further in-depth study to examine remarkable instructional methods and challenges in aspect of teaching and learning.

Three methodologies (Gaussian Mixture Model, K-means clustering and Hierarchical clustering) for clustering analysis of academics was meaningful respectively. GMM, as an initial step, was essential to check overall clusters of 2,639 academic courses opened during one semester. As classic and the most popular algorithm, K-means with both non-standardized and standardized dataset contributed to identify prototypical LMS usage patterns by revealing clusters of course utilized *forum-based online instruction*, *quiz-based online instruction*, and *wiki-based instruction*. Hierarchical clustering method was also valuable for the detection of extreme outlier courses that revealed *resource-based online instruction*. Because of hierarchical analytic approach, few outlier could not be included in other cluster naturally but it was left in isolation. This study confirmed that the different strengths of three methodologies leveraged to escalate the effectiveness and robustness of clustering analysis.

Finally, this study represented that online learning activity was fairly marginal despite the advance of information and communication technology (ICT) and its applications for promoting blended learning policy in higher education. We conclude that offline courses were central to most of higher education. We found too many courses did not incorporate a variety of activity items. LMS such as Moodle and Blackboard provides lots of meaningful activity opportunity like discussion, group works, quiz and Wiki. Although we consider that there might be some cultural characteristics of university in South Korea, we believe the analytics methods and approaches incorporated in this study contribute to the area of academic analytics in the field of higher education.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2013S1A5A2A03044140).

REFERENCES

[1] C. Dalsgaard, "Social software: E-learning beyond learning management systems," *European Journal of Open, Distance and E-Learning,* vol. 2006, 2006.

[2] P. Baepler and C. J. Murdoch, "Academic analytics and data mining in higher education," *International Journal for the Scholarship of Teaching and Learning,* vol. 4, p. 17, 2010.

[3] P. J. Goldstein and R. N. Katz, *Academic analytics: The uses of management information and technology in higher education*: Educause, 2005.

[4] J. P. Campbell, P. B. DeBlois, and D. G. Oblinger, "Academic analytics: A new tool for a new era," *Educause Review,* vol. 42, p. 40, 2007.

[5] K. E. Arnold, "Signals: Applying Academic Analytics," *Educause Quarterly,* vol. 33, p. n1, 2010.

[6] K. E. Arnold and M. D. Pistilli, "Course signals at Purdue: Using learning analytics to increase student success," in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 2012, pp. 267-270.

[7] A. Essa and H. Ayad, "Student success system: risk analytics and data visualization using ensembles of predictive models," in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 2012, pp. 158-161.

[8] A. Kruse and R. Pongsajapan, "Student-centered learning analytics," *CNDLS Thought Papers,* pp. 1-9, 2012.

 ISBN: 978-1-61208-423-7