8

Data Mining for Instructional Design, Learning and Assessment

Lluís Vicent and Xavier Gumara
Business Engineering School La Salle, Ramon Llull University
Spain

1. Introduction

Statistical analysis is widely used in many different areas: medicine, business, natural and social sciences, and of course, in education.
In this last topic, it is common that teachers make simple statistical analysis on the results of the students at the end of an exam or a course, and this is useful for the evaluation of that course. However a more powerful use of statistics can and must be done if the analyses are used to modify the methodology of learning personalizing contents and methods for groups of students with similar skills. To make a realistic personalization of learning, data mining techniques must be used. They are also useful to manage big amounts of information mainly composed by: contents, skills, tools, grades and students.
In this chapter, we present data mining techniques used in instructional design, in learning and in the assessment of the students. In order to reduce, interpret and classify the information, factor and cluster analysis have been used.
Factor analysis is a technique that extracts few unobserved new variables (factors) from a big number of data. These factors are linear combinations of the observed variables and the expert analyzer must define the information that underlies each factor. Cluster analysis classifies all the information in some sets (clusters) of items with common features.
Let's present here two examples of the use of Data Mining in e-learning:

- **Example 1.** An institution must decide its learning methodology, and it has planned to use a Learning Management System (LMS). Of course, an LMS contains many tools, and teachers and students must learn how to use these tools. But not all these tools add value to learning, and probably many of them are redundant, that is, students can acquire the same competences using different tools. In (Vicent, 2007) teachers were asked to value (from 0 to 3) the performance of each tool (24 were considered) to develop each skill. Using factor and cluster analysis, an LMS of only 5 tools was defined to run an engineering online degree in the European Higher Education Area.
- **Example 2.** If an LMS is used for learning, much information of the students is available: results of questionnaires, number of post in the forums, number of visits to the contents, etc. It is possible to classify the students in function of their behavior with a cluster analysis. This way, lazy, willing, active, brilliant, etc.

students can be detected. Results must be used to modify the behavior of some students if needed.

It is obvious that students have different capacities to learn one topic or skill. And each student is better in some skills than in others. When the teachers create assignments or questionnaires, each of these assignments is assessing one or more skills. Let's assume that at some point of a course, a teacher has collected 500 data of each student: questions answers, grades of assignments, forums posts, etc. Data mining techniques are definitively useful to interpret such amount of information. Factor analysis will simplify these 500 data in a few factors, each factor representing an unobserved variable with a real meaning that must be interpreted by the teacher. This factor will represent a skill or a set of skills. This technique suppose an automatic tool to grade skills, even in the case that the teacher did not define, in the assignment or question, which skills were going to be developed and assessed.

In addition, if a unique teacher manages a big group of students, they can be classified in function of their performance in the skill/s of each factor. Cluster analysis will do this classification. This analysis makes the teacher able to write a good report on the state of learning of their students, giving several grades (one per skill) to their students, and classifying the students in different groups taking into account their performance. If this analysis is done several times during a course, teachers can correct deficiencies in the achievement of some skills. In groups of students, teachers can prepare an adaptive learning plan for each group. This adaptive learning plan should be a must for teachers whose students have to achieve a predefined set of skills. This method can be also applied to a global degree, defining adaptive curricula for different groups of students.

In this chapter, the opportunities that the statistical analysis offer to teachers and managers of learning programs is presented.

2. Why Factor Analysis?

It is easy to understand the value of collecting data from students, but also to realize the need of leveraging this data to create knowledge. Data mining technologies offer a way to recognize and track patterns within data. Normally, there exist similarities between the variables analyzed so it is quite possible that we are dealing with redundant information and therefore it is possible to reduce the complexity of the results. In the world of technology we can find some analogies, for example in certain data compression algorithms applied to images or videos for its broadcast on the Internet.

The multivariate approaches for reducing the dimensions of the information can successfully combine some of the collected variables in a few fictitious variables in order to produce minimum information loss.

Factor Analysis is a common statistical method for extracting general information, as usually, many of data collected are related (correlated) to other data, and do not add significant information. Factor analysis detects these correlations and defines factors, which have meaningful information and are linear combination of the general data. Once the factors are found, the supervision of an analyst is needed to give a meaning to the factors.

Principal Components Analysis (PCA) is by far the most common form of factor analysis and its central concept is summarization; it tries to find the minimum variables: factors or principal components, linear combinations of the original variables that explain, with the minimum information loss, the global meaning of the original variables. The key parameter to estimate the information loss is the variance. A factor with high variance means that it carries a lot of information and vice versa.

So PCA is about sorting the factors taking into account the amount of variance that they explain. If with a few factors the most part of the variables can be explained it will mean that the original variables are correlated and the analysis has succeed, since we have been able to reduce the dimensionality of the problem. On the other side, if the original variables are completely uncorrelated all the factors would have approximately the same variance and we won't be able to reduce the dimension.

For the PCA to work properly, it is necessary to subtract the mean from each of the data dimensions. The mean subtracted is the average across each dimension. So, all the *x* values have $\overline{x}$ (the mean of the *x* values of all the data points) subtracted. And the same happens for *y*, *z*, and so on. This produces a data set whose mean is zero.

The next step is calculating the covariance matrix. The covariance matrix for an *N* dimensional dataset would be calculated as shown in (1).

$$C = \begin{pmatrix} \operatorname{cov}(x_1,x_1) & \operatorname{cov}(x_1,x_2) & \cdots & \operatorname{cov}(x_1,x_N) \\ \operatorname{cov}(x_2,x_1) & \operatorname{cov}(x_2,x_2) & & \operatorname{cov}(x_2,x_N) \\ \vdots & & \ddots & \vdots \\ \operatorname{cov}(x_N,x_1) & \operatorname{cov}(x_N,x_2) & \cdots & \operatorname{cov}(x_N,x_N) \end{pmatrix} \tag{1}$$

Since the covariance matrix is square, we can calculate the eigenvectors and eigenvalues for this matrix using a Single Value Decomposition (SVD), diagonalizating the matrix or resolving an eigenvalue equation. The equation for SVD of a matrix *X (m* × *n)* is (2).

$$X = USV^T \tag{2}$$

Where U is an m × n matrix, S is an n × n diagonal matrix, and V^T is also an n × n matrix. The columns of U are called the left singular vectors. The rows of V^T contain the elements of the right singular vectors. The elements of S are only nonzero on the diagonal, and are called the singular values. By convention, the ordering of the singular vectors is determined by high-to-low sorting of singular values, with the highest singular value in the upper left index of the S matrix. This gives the components in order of significance.

Note that for a square and symmetric matrix X (like the covariance matrix), singular value decomposition is equivalent to diagonalization, or solution of the eigenvalue problem.

One way to calculate the SVD is to first calculate V^T and S by diagonalizing X^TX. This process can be seen in (3).

$$\begin{aligned} X &= USV^T \\ X^T &= VSU^T \\ X^T X &= VSU^T USV^T = VS^2V^T \end{aligned} \tag{3}$$

And then, the only incognita left is U, that can be calculated as follows in (4).

$$U = XVS^{-1} \tag{4}$$

It is important to notice that the eigenvectors obtained are unit eigenvectors, that is, their lengths are 1. They are perpendicular to each other and give information about how the datasets are related in order of importance. So, by this process of taking the eigenvectors of the covariance matrix, we have been able to extract vectors that characterize the data. Each component's eigenvalue is called the "amount of variance" the component explains. It turns out that the eigenvector with the highest eigenvalue is the principal component of the data set.

When selecting the number of factors to be extracted it may happen that a minor number of principal components will explain all the variance, which will allow the perfect reconstruction of the original data (even though the number of components found is smaller than the number of original variables). However, in the absence of this event, there is no significance test on the number of principal components to choose.

In (Kaiser, 1960) it is suggested a rule for selecting a number of factors n less than the number needed for perfect reconstruction: set n equal to the number of eigenvalues greater than 1. Several lines of thought lead to Kaiser's rule, but the simplest is that since an eigenvalue is the amount of variance explained by one more factor, it does not make sense to add a factor that explains less variance than is contained in one variable. Since a component analysis is supposed to summarize a set of data, to use a component that explains less than a variance of 1 would be like writing a summary of a book where one section of the summary is longer than the book section it summarizes (Darlington, 1997).

Another criterion to select the number of principal components is to include just enough components that explain some arbitrary amount (typically 80%) of the variance. This can be calculated normalizing the eigenvalues and selecting, in order, the ones that explain the 80% of the variance.

So principal components are linear combinations of the original variables weighted by their contribution to explaining the variance in a particular orthogonal dimension and although the goal of PCA is dimension reduction, there is no guarantee that the dimensions are interpretable. In the next parts of this chapter we present two interpretations of Principal Components Analysis results applied to e-learning.

3. Factor Analysis to Decide which E-learning Tools are Needed in an LMS

3.1 Introduction

In 2005 at La Salle, a group of experts composed by faculty members and technician staff prepared the adaptation of the Engineering programs to the Euopean Higher Education Area (EHEA). As this programes were offered both in the campus and online at La Salle, there was the preocupation of knowing if it was possible to develop all the degrees, taking into account the generic competences the students were supposed to acquire (Tuning, 2001) in a purely on campus way (without the use of an LMS), in a purely online way (with no phisical attendance) or in a blended learning way.

For this reason, on the one hand, all the competences to learn were considered, and on the other hand, an important set of 24 tools for learning (face to face and technological) were stablished. These lists can be seen in Table 1.

Competences	Tools
Conceptual comprehension	Text
Capacity for analysis	Hypertext
Capacity for synthesis	Synthetic video
Planning and time management	Video lesson
Oral communication in the native language	Recording of an on campus class.
Written Communication in the native language	Non teaching purpose videos
Communication in a foreign language	Remote laboratory
Use of information technologies	Simulator
Information management	Virtual library
Ability for mathematical developments	Wiki
Problem solving	Blog
Decision making	Textual forum
Critical and self-critical abilities	Graphical forum
Communication with experts from other areas	Chat
Appreciation of diversity and multiculturalism	Virtual classroom
Teamwork	E-mail and mailing lists
Ethical commitment	News
Ability to work autonomously	Calendar
Adaptation to new situations	Personal folder
Creativity	Working group
Ability for design	Lectures
Leadership	Debate
Initiative and entrepreneur spirit	Interview
Openness to learning all along one's life	Laboratory
Identity, development and professional ethics	
Concern for quality	

Table 1. List of competences and tools considered.

The group of experts wanted to answer 3 questions:

1. Is the face to face class enough to develop the generic competences the Bologna Process indicates?
2. Are the LMS tools enough to develop the generic competences the Bologna Process indicates?
3. Which are the minimum set of tools (online or face to face) good enough to develop all the competences?

To answer these questions experimented face to face and online engineering faculties were polled. The tool was a table in which the resources are placed in the abscissas axis and the

competences in the ordinates axis. The faculty were asked to fill the table answering the following question:
"Qualify in an ascending order, from 0 (slightly suitable) to 3 (very suitable), the educational resources in each column, according to their adequacy for the development of the competences indicated in each line:
0: This skill cannot be developed with this resource
1: This skill might be developed with a non conventional use (different from the usual one) of this resource
2: It is possible to develop the skill with a normal use of the resource
3: This resource is very useful for the development of this skill"

The pool was answered by 38 faculty, and the number of data to analyze was of *26 competences x 24 tools x 38 faculty = 23.712 data*.
The first simplification was to caculate the average of the answers of each association competence – tool, obtaining Table 2 (Vicent, 2007).

	Educational Resources	Text	Hypertext	Synthetic video	Video lesson	Recording of an on campus	Non teaching purpose vide	Remote laboratory	Simulator	Virtual library	Wiki	Blog	Textual forum	Graphical forum	Chat	Virtual classroom	E-mail and mailing lists	News	Calendar	Personal folder	Working group	On Campus resources	Lectures	Debate	Inteview	Laboratory
	Skills	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20		21	22	23	24
A	Conceptual comprehension	2,3	2,3	2,6	2,5	2,1	1,5	2,2	2,5	1,3	0,8	1,0	1,4	1,7	1,3	2,2	0,9	0,4	0,2	0,5	0,6		2,6	2,2	1,7	2,5
B	Capacity for analysis	2,4	2,2	2,3	2,2	2,0	1,7	2,2	2,5	1,4	0,9	1,1	1,3	1,5	1,2	2,1	0,9	0,4	0,2	0,5	0,6		2,6	2,3	1,7	2,5
C	Capacity for synthesis	2,2	1,8	2,0	2,0	1,7	1,4	2,0	2,3	1,4	1,1	1,1	1,4	1,6	1,5	2,1	1,0	0,6	0,3	0,7	0,7		2,5	2,5	1,7	2,4
D	Planning and time management	1,5	1,5	0,7	1,2	1,1	0,8	1,0	0,9	0,9	0,8	1,0	1,4	1,3	0,8	1,3	1,7	2,4	3,0	1,7	2,0		1,4	1,4	1,4	1,2
E	Oral communication in the native language	0,9	0,9	0,9	1,7	1,8	1,7	0,5	0,3	0,3	0,4	0,5	0,5	0,4	0,5	2,1	0,4	0,4	0,3	0,3	0,3		2,2	3,0	2,9	1,2
F	Written Communication in the native language	2,3	2,0	0,7	1,2	1,1	0,9	0,5	0,4	1,6	1,2	1,8	2,5	1,4	2,1	1,6	2,0	1,1	0,6	0,7	1,0		1,4	0,8	0,8	0,9
G	Communication in a foreign language	2,3	2,0	1,2	1,7	1,8	1,7	0,8	0,6	2,2	1,1	1,5	2,0	1,2	1,7	1,8	1,8	0,8	0,6	0,8	1,0		1,9	2,0	1,8	1,1
H	Use of information technologies	1,2	2,0	1,9	2,0	1,7	1,6	2,3	2,1	2,5	1,8	2,0	2,4	2,4	2,3	2,5	2,1	1,4	1,4	1,6	1,8		0,9	0,6	0,7	1,4
I	Information management	1,9	2,1	1,2	1,6	1,4	1,2	1,0	1,0	2,4	1,8	1,8	1,7	1,5	1,1	1,5	1,5	1,2	1,2	1,7	2,1		1,2	1,1	0,9	1,0
J	Ability for mathematical developments	2,2	2,1	1,4	1,7	1,6	1,0	1,4	1,9	1,2	0,5	0,6	0,8	1,3	0,7	1,9	0,6	0,2	0,2	0,4	0,5		2,7	1,7	1,4	1,6
K	Problem solving	2,1	2,0	1,7	2,0	2,0	1,0	2,2	2,3	1,4	1,0	0,9	1,7	1,7	1,5	2,2	1,0	0,5	0,2	0,6	1,0		2,7	2,5	1,9	2,5
L	Decision making	1,3	1,4	0,8	1,4	1,2	1,0	2,0	2,2	1,0	0,8	0,9	1,3	1,2	1,1	1,3	1,0	0,5	0,6	0,5	0,8		1,3	2,0	1,6	2,0
M	Critical and self-critical abilities	1,6	1,3	0,8	1,1	1,2	1,0	1,6	1,8	1,0	1,1	1,3	1,9	1,8	1,5	1,7	1,0	0,4	0,2	0,5	0,9		1,4	2,5	2,0	1,9
N	Communication with experts from other areas	1,3	1,2	0,6	0,9	0,9	1,3	1,0	0,8	1,2	1,4	1,6	2,3	2,3	2,2	1,9	2,3	0,6	0,3	0,2	1,3		1,3	2,1	1,5	1,0
O	Appreciation of diversity and multiculturality	1,4	1,3	0,7	1,0	0,9	1,6	0,4	0,4	1,0	1,4	1,7	1,9	1,5	1,7	1,8	1,6	0,5	0,2	0,3	1,2		1,3	2,3	1,5	1,0
P	Teamwork	0,8	0,6	0,4	0,9	0,6	0,9	0,8	0,6	0,6	1,2	1,3	2,0	1,9	2,4	2,2	1,8	0,4	0,2	0,2	1,3		1,5	2,7	2,2	1,6
Q	Ethical commitment	1,2	1,0	0,7	1,0	1,0	1,5	0,8	0,7	0,8	1,1	1,4	1,7	1,6	1,6	1,6	1,4	0,6	0,4	0,6	1,2		1,6	2,3	1,9	1,5
R	Ability to work autonomously	2,4	2,3	2,0	1,9	1,9	2,0	2,6	2,8	2,5	1,1	1,2	1,8	1,8	1,3	2,0	1,2	0,8	0,9	1,3	1,3		1,5	1,4	1,2	2,3
S	Adaptation to new situations	0,9	1,0	1,0	1,0	0,9	1,1	2,0	2,0	1,4	1,2	1,1	1,2	1,2	1,1	1,6	1,2	0,7	0,6	0,6	0,6		1,0	1,7	1,4	2,0
T	Creativity	1,4	1,4	1,3	1,3	0,9	1,1	1,8	2,0	1,1	1,4	1,6	1,6	1,7	1,2	1,6	0,9	0,3	0,2	0,4	0,5		1,2	2,2	1,5	2,3
U	Ability for design	1,7	1,6	1,2	1,4	1,5	1,2	2,0	2,3	1,3	1,0	1,0	1,3	1,5	1,1	1,6	0,9	0,3	0,1	0,4	0,6		2,1	1,7	1,5	2,6
V	Leadership	0,7	0,6	0,4	0,4	0,6	0,4	1,1	0,7	0,4	1,2	1,3	1,9	1,9	2,0	1,7	1,4	0,6	0,3	0,2	1,2		0,8	2,7	1,7	1,6
W	Initiative and entrepreneur spirit	0,8	1,6	0,5	0,5	0,6	0,6	1,4	1,5	1,6	1,4	1,5	1,9	1,9	1,7	1,4	1,4	0,6	0,4	0,4	1,1		1,0	2,7	1,8	1,8
X	Openness to learning all along one's life	1,7	1,5	1,0	1,2	1,2	1,2	1,2	1,4	2,1	1,1	1,2	1,4	1,4	1,1	1,5	1,3	0,7	0,4	0,8	0,8		1,6	1,6	1,4	1,6
Y	Identity, development and professional ethics	1,2	1,0	0,8	0,9	0,9	1,2	0,7	0,7	0,8	1,0	1,3	1,5	1,3	1,3	1,3	1,3	0,6	0,4	0,7	0,8		1,6	2,0	1,8	1,0
Z	Concern for quality	1,4	1,2	0,9	1,0	1,0	0,7	1,2	1,3	1,0	1,0	0,8	1,0	1,0	0,9	1,2	0,9	0,6	0,8	1,0	1,1		1,5	1,5	1,5	1,5

Table 2. Average of the answers of each association.

Looking at this table it is difficult to answer any of the three questions. So, a factor analysis was used to simplify the data. Since the goal is to develop all the competences, these must not be simplified. There is no problem, on the other hand, to use only some of the tools, if all the competences can be learned. So, the factor analysis was applied to the tools.
In these analysis, it was discovered that only 5 factors could explain the 90% of the information of the table. This meant that many of the tools were superfluous and they were not strictly necessary, as they are as usefull as others. In Table 3 the weight of each tool in each factor can be seen.

	Factor				
	1	2	3	4	5
Text	0,539	0,597	-0,035	0,237	-0,335
Hypertext	0,433	0,801	-0,031	0,078	-0,218
Synthetic video	0,738	0,602	0,138	0,001	0,080
Video lesson	0,755	0,552	-0,013	0,267	0,104
Recording of an on campus class.	0,772	0,495	-0,072	0,312	0,097
Non teaching purpose videos	0,394	0,385	0,113	0,533	-0,021
Remote laboratory	0,549	0,499	0,224	-0,568	0,203
Simulator	0,650	0,426	0,192	-0,573	0,085
Virtual library	0,005	0,863	-0,018	-0,029	-0,281
Wiki	-0,654	0,497	0,341	-0,211	-0,093
Blog	-0,712	0,450	0,376	0,122	-0,184
Textual forum	-0,660	0,453	0,511	0,126	-0,014
Graphical forum	-0,369	0,441	0,684	-0,214	0,232
Chat	-0,592	0,219	0,689	0,177	0,121
Virtual classroom	0,347	0,356	0,530	0,399	0,417
E-mail and mailing lists	-0,801	0,397	0,195	0,239	0,063
News	-0,449	0,490	-0,600	0,062	0,350
Calendar	-0,345	0,442	-0,700	-0,049	0,408
Personal folder	-0,181	0,709	-0,601	-0,057	0,173
Working group	-0,674	0,503	-0,244	0,071	0,310
Lectures	0,848	0,008	0,018	0,333	-0,004
Debate	0,246	-0,702	0,411	0,130	0,302
Interview	0,316	-0,764	0,120	0,285	0,347
Laboratory	0,708	0,084	0,317	-0,529	0,177

Table 3. Weight of each tool in each factor.

Therefore, a table where the appropriateness of each factor for the development of any competence was studied. At Table 4 we remark (blue colour) which factor is most suitable for developing each competence.

Many things can be now understood from this table. The first one is that the instrumental competences can be easily developed as the scores of the factors are high. On the other hand, interpersonal and systemic competences are more difficult to develop from the faculty point of view.

An important point in the factor analysis is the sign. Each factor (Table 3.)has positive weights of some tools and negative for others. Then, in Table 4, we can see that maximum scores for each competence can be positive or negative. How must it be read? Let's see an example. *Planning and time management* is well developed by factor 3 in its negative side. It means that the high negative tools of factor 3 (in Table 3.) we see that they are news, calendar and personal folder) are the idoneous tools for developing that competence.

Therefore, in Table 4. we can detect that positive resources of factor 1 are very important in many skills. Positive resources of factor 2 are important in some competences, and negative factors are needed in some systemic skills. Negative resources in factor 3 are indispensable for organization, and positive factors resources are important for interpersonal skills. Positive resources of factor 4 are essential for oral communication.

	Competence	F1	F2	F3	F4	F5
Instrumentals	Conceptual comprehension	1,81	0,66	0,75	0,30	0,39
	Capacity for analysis	1,66	0,56	0,56	0,32	0,10
	Capacity for synthesis	1,17	0,43	0,61	0,14	0,59
	Planning and time management	-0,95	0,47	-3,13	-0,14	2,42
	Oral communication in the native language	1,31	-2,07	-1,20	2,14	0,98
	Written Communication in the native language	-1,17	0,74	-0,19	1,01	-2,24
	Communication in a foreign language	-0,09	0,57	-0,06	2,09	-0,87
	Use of information technologies	-1,00	2,65	0,65	-0,07	1,40
	Information management	-0,91	1,56	-1,33	0,22	-0,78
	Ability for mathematical developments	1,39	-0,36	-0,66	0,08	-1,12
	Problem solving	1,26	0,39	0,72	0,00	1,05
	Decision making	0,40	-0,57	-0,52	-1,26	-0,18
Interpersonals	Critical and self-critical abilities	-0,04	-0,44	0,68	-0,53	0,26
	Communication with experts from other areas	-1,33	-0,01	1,34	0,83	0,13
	Appreciation of diversity and multiculturality	-0,92	-0,42	0,56	1,28	-0,72
	Teamwork	-1,01	-1,11	1,41	0,58	1,54
	Ethical commitment	-0,55	-0,65	0,25	0,68	0,22
Systemic	Ability to work autonomously	0,84	1,80	-0,08	-0,39	0,13
	Adaptation to new situations	-0,03	-0,42	-0,36	-1,54	-0,15
	Creativity	0,10	-0,28	0,71	-1,42	-0,78
	Ability for design	0,85	-0,14	0,22	-1,09	-0,65
	Leadership	-1,22	-1,22	0,79	-0,86	0,80
	Initiative and entrepreneur spirit	-0,97	-0,52	0,70	-1,37	-0,04
	Openness to learning all along one's life	-0,02	0,01	-0,51	-0,21	-1,15
	Identity, development and professional ethics	-0,52	-0,93	-0,48	0,35	-0,63
	Concern for quality	-0,08	-0,69	-1,45	-1,10	-0,70

Table 3. List of suitable factors for developing each competence.

3.2 Cluster Analysis

In spite of the analysis, cluster analysis can be performed to quickly view the relationships between tools. These relationships will show if the tools have a similar behaviour in the development of competences. Applying the cluster analysis as can be seen in (Vicent et al., 2007) we can detect the next 10 clusters:

1. Personal folder, calendar, news
2. Text, hypertext
3. Graphical forum, Chat, Textual forum
4. Working group, e-mail, blog, wiki

5. Not teaching purposes video, recording of an on campus class, video-lesson, synthetic video
6. Interview, debate
7. Laboratory, simulator, remote lab
8. Virtual library
9. Virtual classroom
10. Lecture

If the tools of a cluster have similar behaviour when developing the same competences, we can assume that we can work with only one tool of each cluster if all the clusters are needed. To know if they are needed we must locate the clusters in the factors as can be represented in Table 4.

	F1	F2	F3	F4	F5
+	Cluster 5 / 10	Cluster 2 / 8	Cluster 3 / 9		Cluster 9
-	Cluster 4	Cluster 6	Cluster 1	Cluster 7	

Table 4. Positive and negative location of clusters into factors.

From the factor analysis we can say that negative tools of factor 1, negative tools of factor 4 and factor 5 can be discarded. So, cluster 4 is prescindible: it is e-mail, wiki, blog, textual forum and group folder.

Now, the best tool of each cluster can be selected, and then we can check if with these tools all the competences can be developed. In Table 4 (Vicent et al., 2007), it was shown that with nine tools almost all the competences could be developed" into "In Table 5 (Vicent et al., 2007), it is shown that with nine tools almost all the competences can be developed. Even on that paper, it was explained that the video-lesson and the lecture could be avoided as they are not indispensable for any competence. Looking at that table, answers can be given to the 3 questions:

1. Is the face to face class enough to develop the generic competences the Bologna Process indicates?
   No. The use of the LMS is strictly necessary for some competences as planning or obviously, the use of IT technologies.
2. Are the LMS tools enough to develop the generic competences the Bologna Process indicates?
   No. There are some competences where presence is very interesting as leadership, or ethical commitment.
3. Which are the minimum set of tools (online or face to face) good enough to develop all the competences?
   Only six: Hypertext, Simulator, Graphical forum, Virtual classroom, Calendar, Virtual library and Debate.

Authors of these paper want to point out that these three answers do not necessary represent their thoughts. These answers are given by the collective of faculty that answered the polls.

	Conceptual comprehension		Capacity for analysis		Capacity for synthesis		Planning and time management		Oral communication in the native language		Written Communication in the native language		Communication in a foreign language		Use of information technologies		Information management		Ability for mathematical developments		Problem solving		Decision making		Critical and self-critical abilities		Communication with experts from other areas		Appreciation of diversity and multiculturality		Teamwork		Ethical commitment		Ability to work autonomously		Adaptation to new situations		Creativity		Ability for design		Leadership		Initiative and entrepreneur spirit		Openness to learning all along one's life		Identity, development and professional ethics		Concern for quality	
Hypertext	2,3	2,2	1,8	1,5	0,9	2,0	2,0	2,0	2,1	2,1	2,0	1,4	1,3	1,2	1,3	0,6	1,0	2,3	1,0	1,4	1,6	0,6	1,6	1,5	1,0	1,2
Video-lesson	2,5	2,2	2,0	1,2	1,7	1,2	1,7	2,0	1,6	1,7	2,0	1,4	1,1	0,9	1,0	0,9	1,0	1,9	1,0	1,3	1,4	0,4	0,5	1,2	0,9	1,0
Simulator	2,5	2,5	2,3	0,9	0,3	0,4	0,6	2,1	1,0	1,9	2,3	2,2	1,8	0,8	0,4	0,6	0,7	2,8	2,0	2,0	2,3	0,7	1,5	1,4	0,7	1,3
Virtual library	1,3	1,4	1,4	0,9	0,3	1,6	2,2	2,5	2,4	1,2	1,4	1,0	1,0	1,2	1,0	0,6	0,8	2,5	1,4	1,1	1,3	0,4	1,6	2,1	0,8	1,0
Graphical forum	1,7	1,5	1,6	1,3	0,4	1,4	1,2	2,4	1,5	1,3	1,7	1,2	1,8	2,3	1,5	1,9	1,6	1,8	1,2	1,7	1,5	1,9	1,9	1,4	1,3	1,0
Virtual classroom	2,2	2,1	2,1	1,3	2,1	1,6	1,8	2,5	1,5	1,9	2,2	1,3	1,7	1,9	1,8	2,2	1,6	2,0	1,6	1,6	1,6	1,7	1,4	1,5	1,3	1,2
Calendar	0,2	0,2	0,3	3,0	0,3	0,6	0,6	1,4	1,2	0,2	0,2	0,6	0,2	0,3	0,2	0,2	0,4	0,9	0,6	0,2	0,1	0,3	0,4	0,4	0,4	0,8
Lecture	2,6	2,6	2,5	1,4	2,2	1,4	1,9	0,9	1,2	2,7	2,7	1,3	1,4	1,3	1,3	1,5	1,6	1,5	1,0	1,2	2,1	0,8	1,0	1,6	1,6	1,5
Debate	2,2	2,3	2,5	1,4	3,0	0,8	2,0	0,6	1,1	1,7	2,5	2,0	2,5	2,1	2,3	2,7	2,3	1,4	1,7	2,2	1,7	2,7	2,7	1,6	2,0	1,5
Maximum online resources	2,5	2,5	2,3	3,0	2,1	2,0	2,2	2,5	2,4	2,1	2,3	2,2	1,8	2,3	1,8	2,2	1,6	2,8	2,0	2,0	2,3	1,9	1,9	2,1	1,3	1,3
Maximum on campus resources	2,6	2,6	2,5	1,4	3,0	1,4	2,0	0,9	1,2	2,7	2,7	2,0	2,5	2,1	2,3	2,7	2,3	1,5	1,7	2,2	2,1	2,7	2,7	1,6	2,0	1,5
Difference	0,1	0,1	0,3	-1,5	0,9	-0,6	-0,1	-1,7	-1,2	0,5	0,4	-0,2	0,7	-0,2	0,5	0,6	0,6	-1,2	-0,3	0,2	-0,3	0,8	0,7	-0,5	0,7	0,2

Table 5. Appropriateness of the optimal tools for the development of all competences.

4. Factor Analysis for Grading Competences

4.1 Introduction

As stated before in this chapter, in the context of the European Higher Education Area (EHEA) teachers are encouraged to evaluate not only the students' knowledge in a subject (contents) but also the skills they may acquire during the learning process (competences).

Contents have always been evaluated through a course by using traditional methods like exams, test questionnaires, continuous evaluation, practical tasks and depending on the subject the list can continue with much more methods.

But since the implantation of the EHEA, teachers must grade competences too. That means they have to evaluate the performance of a student in different skills like capacity of analysis, capacity of synthesis, capacity to assume new concepts from a subject, capacity to make maths developments or even more complex to evaluate skills like the capacity of working on a team.

For example, in a Computer Science subject where the teacher wants to evaluate the basis of Object Oriented programming it is easy to grade concepts like hierarchy, abstraction or interfaces; but it is difficult for them to evaluate how the students are acquiring the previously mentioned and other skills.

Having this in mind, we wanted to focus on question-based tests to try to extract hidden information between the questions to allow the grading of competences; that is because questionnaires produce a lot of information that is not analyzed by the teachers, sometimes because there is such an amount of information that it is impossible to handle it with the naked eye.

For example in a subject where the teacher uses tests every month to evaluate contents, at the end of the year he/she can have up to 500 questions answered by every single student and the only mark that it is being extracted right now from all this information is the mean.

If the assessment of tests is done online with the help of questionnaires tools from the LMS or any other interoperable tool, it is even easier to use the results of an exam to discover new information from the data.

Since teachers are required to have more than one mark per student and subject in EHEA, one of the motivations of the application of data mining in e-learning is to take profit of the automatic recording of the grades in LMS to automatically grade competences. In this context, an interoperable tool named *Stats Engine* has been developed at La Salle and presented in (Gumara et al., 2008).

4.2 Extraction of Competences

As stated before, it is difficult for the teacher to read reports from question-based test from students when they have answered a large number of questions. In order to simplify the understanding of the learning of the students a statistical technique can be applied to the data.

Since the objective is to reduce the amount of data, Principal Components Analysis (PCA) could be applied. PCA describes the variation among many variables in terms of a few underlying but unobservable random variables so it is an appropriate data mining technique that is going to help teachers to better understand test results.

In this study the data used to perform PCA are test results and our base point would be an X matrix (s × q) where the rows represent students and the columns the questions of question-based tests. A cell from that matrix has a specific score from a student to a question of a test.

By performing PCA we would get the matrix named V (q × f) storing the eigenvectors from the decomposition where q is the number of questions of the test and f is the number of factors found in the analysis. A cell from this matrix tells us how much of a question is important in a factor.

Besides that we also get S (f × f), a diagonal matrix holding the eigenvalues of the PCA that give information about how important is a factor in explaining the original variables and how much variance explains respect the other factors. This conversion can be seen in Fig. 1.

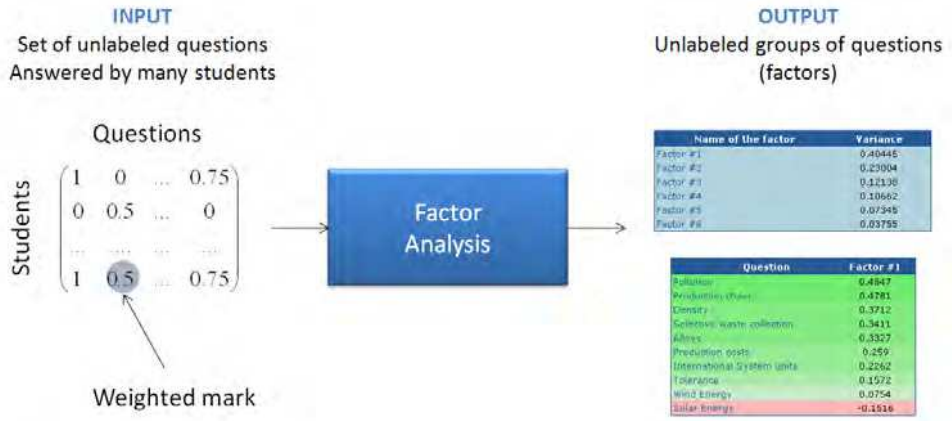

Fig. 1. Use of factor analysis to reduce the dimensionality of tests results.

The aim of our factor analysis is to find those factors which are inferred in the whole imported test dataset if it is reliable enough. For a test data set to be reliable remember that the teacher can reject questions which scores have a negative correlation with the overall

scores of the test or that the fact of dropping them from the exam makes Cronbach's alpha to increase.

So, by having a set of unlabeled and reliable questions answered by many students (the more results the better), PCA is able to group questions weightily under a set of unlabeled factors.

We can assume that, if a student has a real skill, he/she will correctly answer the questions related to that skill and vice versa. If the teacher looks carefully to the output from a factor analysis he/she might find some kind of relation between the questions with higher loads, both positive and negative, from the same factor. This relation can be motivated by two facts:

- The questions talk about the same concept, so the students who know about this concept answer correctly the same kind of questions and vice versa, or what is more interesting,
- the questions belong to the same competence or group of competences a student may acquire.

An example of that can be seen in Fig. 2 where an eigenvector for the second factor explaining the most variance is showed. The results belong to an exam of a Data Transmission subject performed at La Salle. The dark green and red questions are the ones that the factor places greater emphasis on; questions 6, 16, 18, 17, 12 and 11 (by this order) are the ones determining the underlying factor #2 as they have the highest positive loads. The dark red marked questions may also be considered by naming the factor if they belong to a clear opposite competence to former questions. Crossed questions belong to the original test but they were rejected from the analysis according to unreliability detection performed by *Stats Engine* and explained at (Gumara et al, 2008).

Question	Factor #2
6. Data communications model	0.5367
16. Analog modems	0.3858
18. Transmission impairments	0.2909
17. Analog modems	0.2267
12. Data communications concepts	0.211
11. Baseband data transmission	0.1034
5. A-Law and PCM (Pulse Code Modulation)	0.0767
10. Error detection and control	0.0586
7. Start bit	0.014
14. ADSL2+	-0.0761
1. ASK Modulation (Data Transmission)	-0.0826
8. Trellis graph	-0.1193
19. HDB (High Density Bipolar)	-0.1353
13. QAM Modulation	-0.1532
15. Modulations	-0.1564
4. Error-correcting codes	-0.2263
9. Data transmission system	-0.467
~~20. TDM (Time-Division Multiplexing)~~	-
~~3. Hamming code (error-correcting code)~~	-
~~2. A-Law and Uniform Quantification~~	-

Fig. 2. Vector of loads of a factor from PCA results.

In this example, the factor can be renamed to"Basic general knowledge in the field" since all the questions were theoretical concepts from Data Transmission as opposite from other

factors of the same test that grouped questions that implied the application of basic concepts formulas to be correctly answered. One of these factors was named "Modulation knowledge in practice".
Automatically label questions under skills or competences is a great feature that will surely help teachers when planning their future exams but right now we do not have any information about the performance of a student in each factor.
This performance can be measured with principal components scores. If a student gets a good positive score in "Basic knowledge in the field of study", it will mean that he/she is good in learning concepts by heart.
But scores are not directly generated by PCA, they have to be calculated with the help of the original qualification data and the PCA output.
Let V (q × f) be the matrix of eigenvectors and X (s × q) the original dataset matrix, (Jobson, 1994) uses the relationship Z = XV to find student scores to factors. Z (s × f) contains the principal components scores from each student to each factor. An example of this can be seen in Fig. 3 where principal components scores for a student are shown.

Factor #1	0.794
Factor #4	0.36
Factor #9	0.16
Factor #3	0.153
Factor #2	0.136
Factor #15	0.121
Factor #14	0.116
Factor #16	0.06
Factor #8	0.048
Factor #10	0.021
Factor #6	-0.0050
Factor #5	-0.022
Factor #7	-0.03
Factor #13	-0.052
Factor #11	-0.091
Factor #12	-0.093

Fig. 3. Principal component scores of a student.

Although the application described in this chapter is intended to be used with test results, the fact of extracting skills performance through PCA can be applied to other kind of sources. For example, a teacher would be able to score competences form other kinds of evaluation he/she may assign to the classroom like projects developed, personal interviews, continuous evaluation, etc.

4.3 Cluster Analysis

The objective of cluster analysis in this research is to automatically group students using the previously calculated factors as the data set to execute it. When the students clustering algorithm is fulfilled a set of unlabeled groups of students is given to the teacher as can be seen in Fig. 4.
In order to process the data, each cluster is assigned a vector of means and a vector of standard deviations, each pair of values belonging to a single factor. By looking at the results and having previously tagged factors, the teacher is now able to label groups.

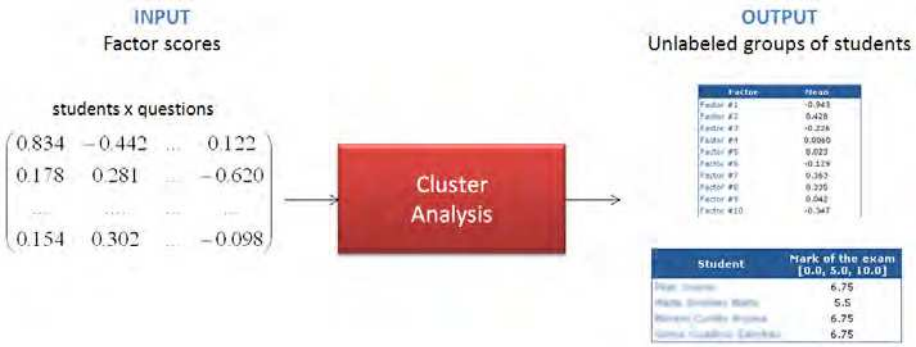

Fig. 4. Use of cluster analysis to group students according to performance on different skills.

5. Conclusions

In this chapter, two examples of data mining applications in e-learning are presented and proved to give knowledge to the teacher.

In the first example, factor analysis has been used in order to set the minimum group of tools for developing an online degree in the European Higher Education Area (EHEA). The research started by polling experienced faculty and then factor analysis was applied to the results to understand their thoughts. Finally, the data from factor analysis was interpreted stating that with a simple LMS of 5 tools and some on campus debates it was possible to cover all the competences. With this analysis it can be understood that a complementariety between face to face classes and online work was needed.

In the second example, a complementary application to questionnaire tools from LMS has been implemented and demonstrated to expand reporting options from open source LMS. The main power of the system is to provide an automatic classification of students according to their performance on different competences. In addition, the system automatically groups the whole group of questions in few factors. Each factor means one or more competences that the student acquires. So this tool becomes a good aid for the competences assessment in the context of an EHEA.

This feature is very useful for the teacher when the scoring periods come. With *Stats Engine* the teacher will not be able to score every single competence required to evaluate (capacity of working on a team may be hard to evaluate with question-based online test results) but will become a great support tool and sure make the scoring of competences and skills easier. The mathematics required have been tested and the algorithms contrasted in order to be suitable for this kind of data.

To summarize, factor analysis is a data mining technique that helped us to reduce a large number of variables in two different problems, both related to learning. By studying the patterns of relationship among many dependent variables, with the goal of discovering something about the nature of the independent variables that affect them, we wereable to give a better interpretation of the data, an interpretation that could not be performed with the naked eye.

6. References

Darlington, R. B. (1997), *Factor Analysis*. Retrieved from http://www.psych.cornell.edu/Darlington/factor.htm

Gumara X., Vicent L., Segarra M. (2008). QTI Result-Reporting Stats Engine for Question-Based Online Tests, *Proceedings of the IEEE ICALT*, pp. 717-721, 978-0-7695-3167-0, Santander, July 2008, CPS, Los Alamitos CA

Jobson J. D. (1994), *Applied Multivariate Data Analysis Volume 2*, Springer.

Kaiser H. F., *The application of electronic computers to factor analysis*, Educational and Psychological Measurement, 1960, 20, pp. 141-151.

Smythe C., Shepherd E., Brewer L. and Lay S. (2002), *IMS Question & Test Interoperability: An Overview*, Final Specification, Version 1.2, IMS.

Vicent, L., G. Bou, X. Avila, J. Riera, J. A. Montero (2007). Which are the best e-learning tools for an Engineering Degree in the European Higher Education area?, *Proceedings of the IEEE ICALT*, pp. 882-886, Niigata, July 2007, CPS, Los Alamitos CA

Tuning (2001). Tuning Education Structures in Europe. http://www.unideusto.org/tuning/

E-learning Experiences and Future

Edited by Safeeullah Soomro

ISBN 978-953-307-092-6
Hard cover, 452 pages
Publisher InTech
Published online 01, April, 2010
Published in print edition April, 2010

This book is consisting of 24 chapters which are focusing on the basic and applied research regarding e-learning systems. Authors made efforts to provide theoretical as well as practical approaches to solve open problems through their elite research work. This book increases knowledge in the following topics such as e-learning, e-Government, Data mining in e-learning based systems, LMS systems, security in e-learning based systems, surveys regarding teachers to use e-learning systems, analysis of intelligent agents using e-learning, assessment methods for e-learning and barriers to use of effective e-learning systems in education. Basically this book is an open platform for creative discussion for future e-learning based systems which are essential to understand for the students, researchers, academic personals and industry related people to enhance their capabilities to capture new ideas and provides valuable solution to an international community.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Lluis Vicent and Xavier Gumara (2010). Data Mining for Instructional Design, Learning and Assessment, E-learning Experiences and Future, Safeeullah Soomro (Ed.), ISBN: 978-953-307-092-6, InTech, Available from: http://www.intechopen.com/books/e-learning-experiences-and-future/data-mining-for-instructional-design-learning-and-assessment

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821