대학 강의평가에서 문항 추출에 관한 연구

황세명*, 김인택**

명지대학교 정보공학과 대학원*, 명지대학교 통신공학과 교수**

A Study on Effective Selection of University Lecture Evaluation

Se-Myung Hwang*, In-Taek Kim**

*Department of Information Engineering, Myongji University**

*Department of Communication Engineering, Myongji University***

국문요약

본 논문에서는, 강의 평가에 필요한 설문을 효과적이며 체계적으로 얻기 위한, 대표 문항 추출 방법을 비교하였다. 비교에 사용한 방법은 요인분석(Factor Analysis: FA), FCM(Fuzzy c-Means) 알고리즘과 군집분석(Cluster Analysis : CA) 등으로 이러한 방법들을 사용하여 고려할 수 있는 다양한 형태의 많은 문항들로부터 적은 수의 문항을 추출한다. 추출된 문항은 많은 수의 문항들이 형성하는 클러스터의 대표 문항을 이루고 있다. 이를 위해 여러 개의 설문지로부터 얻은 120 문항의 강의 평가서를 명지대학교 외 3 개 대학교 646명의 학생들에게 평가를 실시하여 데이터를 얻었는데 학생들은 주어진 문항에 대하여 "매우 그렇다", "그렇다","보통이다", "그렇지 않다", "매우 그렇지 않다" 그리고 "해당 없다"까지의 6등급으로 응답하였다. 각 문항에 대한 학생들의 응답 성향을 분석하여 약 25 문항을 추출하였다. 실험 결과 본 논문에서 비교 분석한 요인분석, FCM 알고리즘과 군집분석 등의 기법은 매우 유사한 설문을 추출할 수 있었다.

Abstract

In this paper, selecting survey items was performed using three clustering methods: factor analysis, fuzzy c-Means algorithm and cluster analysis. The methods were used to extract key items from various questionnaires. The key item represents several similar questionnaires that form a cluster. Test survey was made of 120 items obtained from several surveys and it was answered by 646 students from 4 universities. Each item contains 6 choices. Applying the clustering method chose 25 items which is reduced from the original 120 items. The results yielded by three methods are very similar.

주제어 : 요인 분석, FCM 알고리즘, 군집분석, 클러스터링
Keywords : Factor Analysis(FA), Fuzzy c-Means Algorithm, Cluster Analysis(CA), Clustering

I. 서 론

강의 평가 결과는 강사에게 강의의 질과 문제점을 인식시키고 강의 개선의 기회를 제공하며, 학생에게는 강의 선택의 길잡이가 될 수 있다. 뿐만 아니라 더 나아가 학교는 강사 평가의 근거로 사용할 수 있다. 따라서 강의 평가는 공정하고 객관적이며 종합적이어야 한다. 강의 평가 설문은 명확하고 간결해야 하며 동시에 그 설문의 수도 적절해야 설문을 작성하는 사람의 의견을 잘 반영할 수 있다. 예를 들어, 강의 평가 설문 중 "수업이 체계적이었느냐"라고 묻는 질문이 있다면 학생들이 이에 응답하기 그다지 쉽지 않을 것이다. 반면에 구체적으로 "수업을 강의 계획서대로 진행하였는지", "강의의 목표를 명확히 제시하였는지", "성적평가의 기준이 명확하였는지"등의 질문을 받는다면 응답하기 훨씬 수월할 것이다. 한편 설문의 수도 중요한 요소가 된다. 설문의 수가 너무 많으면 설문자로 하여금 성실한 답변을 기대하기 어려울 뿐만 아니라 중복되는 설문이 존재할 가능성도 높아진다. 강의 평가 문항이 어떤 특정 부분(예 : 교수법)에 대해서만 중복하여 질의한다면, 이는 강의에 대한 전반적인 평가라고 보기 어렵다(원태연, 2004).

강의 평가 설문지의 작성은 아직도 몇몇 관련자에 의하여 임의로 만들어지고 있다. 이를 위해 여러 종류의 설문지에서 문항을 추출하거나, 유사한 질문의 형태로 변형하여 강의 평가 설문지를 작성하고 있다. 이와 같이 강의 평가 설문지를 작성하기 위한 체계적인 방법론이 아직도 존재하지 않는다. 명지대학교 공학연구소에서는 이러한 문제점을 해결하기 위하여 체계적인 설문지 작성을 시도하였다. 이를 위해 기존에 사용되고 있는 설문지의 모든 문항을 취합하였다. 총 취합된 120 개의 문항의 설문지를 4개 대학교의 646명의 학생들에게 작성하여 설문 결과를 얻었다. 이렇게 얻은 설문지는 중복성을 제거하고 대표성을 가지는 문항을 얻기 위해, 추출과정에서 군집 분석 기법을 사용하였다(박광배, 2004).

본 논문에서는 설문 추출을 위한 효과적인 방법을 찾기 위해 군집분석이외에도 요인분석과 FCM(Fuzzy c-Means) 알고리즘을 적용하였는데 이들은 궁극적으로 클러스터링(Clustering)을 이용하는 방법이다. 다수 개의 질문에 대한 응답 결과가 어느 정도의 유사한 성향을 가질 때, 이를 하나의 클러스터로 취급함으로써 대표 문항으로 선택할 수 있다(이순묵, 1995 ; Frank Höppner et al, 1999).

본 논문의 구성은 다음과 같다. 2장에서는 대표 문항이란 무엇이며 이것을 추출하기 위한 방법으로 사용된 클러스터링 알고리즘에 대하여 전반적으로 살펴본다. 논문에서 사용된 요인분석, FCM 알고리즘과 군집분석의 주요 개념을 정리하고, 실험에 적용되는 내용에 대해 이론적으로 살펴본다. 3장에서는 대표 문항의 추출 예로, 제안된 방법의 적용을 설명하고, 각각의 실험 구성 및 결과를 비교 분석한다. 4장에서는 연구 결과를 요약하고 실험 결과에 대하여 논의한다.

II. 대표 문항 추출 방법

문항은 강의 평가서를 구성하는 요소이다. 강의 평가가 의도한 제대로의 기능을 다 하느냐 못하느냐는 그 속에 포함된 문항이 제대로 되어있는지에 따라 결정된다. Joseph Hoey(2000)는 좋은 문항은 아

래와 같은 특성을 가져야 한다고 소개했다.

- 한 문항에 오직 하나의 질문을 포함할 것
- 문항을 짧고 간결하게 또 명확하게 서술할 것
- 전문용어를 피하고 쉬운 단어로 작성할 것
- 유도질문을 피할 것
- 중복적인 질문을 피할 것
- 답하기 거북한 질문을 피할 것
- 부정의 부정적인(double negative) 질문을 피할 것
- 추상적인 질문을 피할 것 (p 11-28).

따라서 대표문항은 앞에서 서술한 좋은 문항의 특성을 가짐과 동시에, 클러스터를 형성하는 문항들을 대표하는 것으로 정의할 수 있다. 강의 평가 설문지는 여러 개의 설문 그룹들로 형성되어 있으며, 이 그룹들은 예를 들어, 과목의 내용, 교수법, 교수, 평가 및 과제, 학생, 시설 및 장비, 조교 및 스태프 등이 있을 수 있다. 대표문항은 각 그룹에서 그 그룹 목적의 성취도를 잘 반영하는 항목으로 선정되어야 한다. 이 과정에서 각 그룹에는 몇 개의 클러스터가 형성되며, 그룹 사이의 클러스터는 존재하지 않는다고 본다. 즉, 교수법과 교수를 동시에 대표할 수 있는 문항은 없다.

본 논문에서는 대표문항을 추출하기 위해 문항들을 클러스터링하여 문항 수를 줄이는 방법을 비교 분석한다. 이를 위해 주어진 문항에 대한 응답을 조사한다. 만일 어떤 두 개의 문항에 대한 응답이 주어진 특정 문턱값 이상의 유사성을 가지면 두 개의 문항은 하나로 축약될 수 있다. 클러스터는 여러 개의 문항들이 하나의 문항으로 대표될 수 있는 문항들의 집단이다. 클러스터링은 이렇게 비슷한 특성을 가진 항목들을 집단으로 구성하는 과정을 의미한다. 이 과정에서 항목들 간의 유사도 또는 거리를 측정하는 기준이 요구된다. 클러스터링은 데이터에 내재된 특성을 자동으로 추출하는 기법으로 무감독 학습(unsupervised learning)의 주된 접근법 중에 하나이다. 무감독 학습법의 목표는 데이터에 내재된 분포 특성에 대한 적절한 표현을 찾아내는 것이다. 구체적인 목표로는 중복성 절감, 정보량 최대화, 엔트로피 최소화, 재구성(reconstruction), 오차 최소화 들을 들 수 있다. 클러스터링은 고차원의 데이터를 시각화하거나 압축하는 방법 중 하나로서, 데이터 분석, 시각화, 압축 및 패턴 인식, 영상 처리 등 여러 공학 분야에서 전처리(processing) 과정으로 많은 분야에서 널리 응용되고 있다(Frank H?ppner et al, 1999 ; 김대원 · 이광형, 2003).

1. 요인 분석(Factor Analysis)

요인분석(factor analysis)은 다수 변수들 간의 상관관계(correlation)를 이용하여 그들 간의 체계적인 구조를 밝히고, 서로 유사한 변수끼리 묶어주는 분석기법 중의 하나이다. 즉, 여러 개의 변수들이 서로 어떻게 연결되어 있는가를 알아보고, 변수들의 저변에 존재하는 공통적인 특성을 설명하는 적은 수의 요인으로 축소시키는 분석방법이다(이순묵, 2000). 두 변수의 상관관계의 정도를 나타내는 상관계수(correlation coefficient)는 [-1, 1]의 값으로 상관계수는 공분산의 양에 비례한다. 공분산은 한 변수가 변할 때 다른 변수가 변하는 양으로, 두 변수가 동시에 변하는 정도를 나타내는 것으로 식 (1)과 같이 표현된(이광호, 2002 ; 박광호, 2004).

$$S_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n-1} \quad (1)$$

여기서 n은 변수의 개수이며, 공분산 S_{xy}는 한 변수 X_i가 그 평균값 $\bar{X}$로부터 또 다른 변수 Y_i가 그 평균값 $\bar{Y}$로부터 얼마나 떨어져 있는가를 의미한다. 한 변수가 평균값으로부터 멀리 떨어져 있을 때 다른 변수도 그 평균값으로부터 멀리 떨어져 있으면 공분산은 커지게 되며, 그런 경향이 일반화되어 있으면 상관은 높다. 상관계수 r는 공분산 S_{xy}을 각 변수의 표준편차 S_X, S_Y로 나누어 얻는다.

요인분석에서 상관관계가 높은 변수들끼리 동질적인 몇 개의 집단으로 묶어주는 목적은 다음과 같다. 첫째, 동일한 개념을 측정하기 위한 변수들 간에는 높은 상관관계가 존재하므로 동일한 개념을 측정한 변수들이 동일한 요인으로 묶이는지의 여부를 확인함으로써 측정도구의 타당성을 검증해 보는데 이용될 수 있다. 둘째, 많은 변수들을 공통요인으로 묶어 줌으로써 자료의 복잡성의 문제를 해결하고 요약 정보를 제공한다. 이는 불필요한 변수들의 제거로서 요인에 포함되지 않거나 포함되더라도 중요도가 낮은 변수들을 찾아낼 수 있게 해준다. 즉 , 변수 수보다 적은 수의 요인으로 변수를 대신하는 것이다. 이런 의미에서 요인분석은 자료를 축약시켜 단순하게 하는 효과가 있는데, 이러한 특성을 이용하여 다수의 문항으로부터 소수의 문항을 얻을 수 있다.

본 논문에서는 축소대상인 구체적인 질문의 문항들이 변수이며, 축소된 문항들이 대표하는 추상적인 개념이 요인이다. 요인은 지표(index)를 만드는데 활용된다. 이 경우 요인을 하나의 지표로 파악하고, 각각의 변수들은 지표를 구성하는 요소가 된다. 본 논문에서 언급하는 강의평가서의 예를 든다면, "과목의 내용"관한 세 질문의 문항을 종합하여 과목 내용지표를 만들 수 있다. 이렇게 만들어진 지표는 또한 세 질문 문항을 종합하여 만든 종합변수로서 연속변수의 형태를 갖는다. 이들 요인으로써 변수들을 표시하게 된다.

가. 요인분석의 적용

전술한 바와 같이, 요인분석의 기본원리는 상관관계에 의한 유사한 변수의 통합이다. 요인분석의 첫 번째 과정은 상관관계를 측정하는 일이다. 변수간의 높은 상관관계는 높은 동질성을 의미하는 것으로 이는 요인적재량(factor loading)이 가장 높은 곳에 변수들이 속하게 된다고 설명할 수 있다. 이렇게 요인적재량이 높은 변수들이 집단을 형성할 때, 그 집단들(변수의 수보다 작은 수의)을 요인(factor)이라 한다. 요인은 기본적으로 식 (2)과 같이 변수들 간의 선형조합(linear combination)으로 표현할 수 있다(이순묵, 1995; 이영준, 2002).

$$F_i = a_{1i}X_1 + a_{2i}X_2 + \cdots a_{ni}X_n \qquad (2)$$

F_i번째 요인, a : 계수
X : 변수 , n : 변수와 계수의 수

요인분석의 두 번째 과정은 요인도출 모형결정으로 주성분 분석(Principal Component Analysis: PCA) 방법을 사용한다. 주성분 분석은 여러 개의 변수들이 가지는 정보를 가능한 적은 개수의, 그리고 서로 중복되지 않는 주성분들로 축약한다. 이때 주성분들이 원래 변수들의 정보를 최대로 함축하게 된다. 주성분은 분산-공분산 행렬(variance-covariance matrix)을 이용하여 PCA로 얻어진다. 이 때, 첫 번째 주성분은 가장 분산이 많은 자료를 설명해 준다. 그 후의 두 번째, 세 번째,...., n 번째 주성분은 분산이 적어지는 순서로 자료의 나머지 정보들을 설명한다(박광배, 2004).

첫 번째 주성분 분석(First Principal Component)의 정의

$\underline{a}_1'\underline{a}_1=1$을 만족하는 벡터 $\underline{a}_1$ 중 $\underline{a}_1'(\underline{x}-\underline{\mu})$의 분산을 최대화하는 $\underline{a}_1$를 찾은 후 계산한 아래 벡터를 첫 번째 주성분이라(First Principal Component) 한다. $\underline{x}$는 변수 벡터를 나타낸다.

$$\underline{y}_1=\underline{a}_1'(\underline{x}-\underline{\mu}) \tag{3}$$

n번째 주성분 분석(n-th Principal Component)의 정의
$\underline{a}_n'\underline{a}_n=1,\ \underline{a}_1'\underline{a}_n=\underline{a}_2'\underline{a}_n=\cdots\underline{a}_{n-1}'\underline{a}_n=0$ 을 만족하는 벡터 $\underline{a}_n$ 중 $\underline{a}_n'(\underline{x}-\underline{\mu})$의 분산을 최대화하는)를 찾아 n번째 주성분을(n-th Principal Component)다음과 같이 구한다.

$$\underline{y}_n=\underline{a}_n'(\underline{x}-\underline{\mu}) \tag{4}$$

위의 방법을 계속하면 반복하여 변수의 개수만큼 주성분들을 $\underline{y1}, \underline{y2}, \cdots \underline{y_p}$구한다. 주성분은 변수의 개수만큼 존재하고 각 주성분은 서로 독립이다. (상관 계수 = 0)

$$\underline{y}=\begin{pmatrix}\underline{y}_1\\ \underline{y}_2\\ \vdots\\ \underline{y}_p\end{pmatrix}=\begin{pmatrix}\underline{a}_1'\\ \underline{a}_2'\\ \vdots\\ \underline{a}_p'\end{pmatrix}(\underline{x}-\underline{\mu}) \tag{5}$$

각 주성분 계수 벡터 $\underline{a}_j$를 구하는 방법은 변수 벡터의 분산-공분산 행렬 Σ의 고유값(eigenvalue)들에 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p$에 대응하는 $\underline{e1}, \underline{e2}\cdots\underline{e}_p$고유값들 이 주성분 조건을 만족한다.

$$\underline{a}_1=\underline{e}_1,\ \underline{a}_2=\underline{e}_2,\ \cdots,\ \underline{a}_p=\underline{e}_p \tag{6}$$

고유값에 의한 주성분과 분산은 아래의 관계를 가진다.

- 주성분 y_j의 분산은 고유값 λ_1이다.
- 분산-공분산 행렬의 대각선 요소의 합 trace는 변수, $x_1, x_2, \cdots, x_p$들의 변동(분산)의 합(=$tr(\Sigma)$)이다.
- 주성분의 변수 변동의 설명력은 $\lambda_j/\mathrm{tr}(\Sigma)$이다.이다.

각 주성분의 설명력은 $\lambda_j/\mathrm{tr}(\Sigma)$으로 나타난다. 첫 번째 주성분의 설명력이 가장 크고, 두 번째 주성분의 설명력은 그 다음으로 크다. 주성분 설명력의 합이 변수의 총 분산 중 약 80%가 되는 주성분까지만 사용한다.

$$0.8\leq \sum_{\min k}\frac{\lambda_k}{tr(S)} \tag{7}$$

분산-공분산 행렬 (Σ)대신 본 논문에서는 상관계수 행렬 (R)을 사용한다. 고유값이 1이상인 주성분

만 사용하면 총분산의 80%정도를 설명하므로 고유값이 1이상인 고유벡터로부터 구해진 고유값만으로 주성분을 구한다.

마지막 과정으로 요인값을 중심으로 어떤 방법으로 유사한 요인끼리 묶어야 하는가를 고려하는 요인의 분리 방법이다. 본 논문에서는 요인들 간의 축들이 직교하도록 유지하면서 요인이 추출되는 방식인 VARIMAX 방법을 사용하였다(Charles D. & Fyfe C., 2000).

2. FCM(Fuzzy c-Means) 알고리즘

FCM(Fuzzy c-Means) 클러스터링은 HCM(Hard c-Means) 클러스터링의 퍼지 모델로 알고리즘의 결과로 클러스터 중심과 FCM 행렬을 동시에 구할 수 있는 자기 조직화, 무감독 학습(unsupervised learning)의 대표적인 예이다. FCM 알고리즘은 각 데이터에 특정 클러스터 중심과의 거리에 따른 소속정도(degree of membership)를 부여한다. 그러므로 각 데이터는 모든 클러스터에 대한 소속정도를 가지며 이것이 분류의 근거로 사용된다. FCM은 n개의 벡터 x_i, $i=1$, $\cdots$, n의 집합을 c개의 퍼지 그룹들로 분할하고, 비유사성 측정의 비용함수가 최소가 되는 것과 같은 각각의 그룹 안에서 클러스터의 중심을 찾는다. 데이터 집합에 대한 소속감 정도의 합은 항상 1이다.

가. FCM의 적용

소속이 불확실한 변수 자료를 표현하려면 퍼지 분할 공간을 사용하는 퍼지 부집합을 사용하는 것이 바람직하다. 변수 자료 집합 X는 그 원소인 변수 자료 x_k가 p차원의 벡터로 표시되므로 식 (8)과 같이 나타내어진다(Frank HÖppner et al, 1999).

$$X = \{x_1, x_2, \cdots, x_n\} = \{X_k\} \subset R^p \quad = \text{불확실한 객체 자료} \tag{8}$$

c개 클래스의 분류 표시 자료로 변수 자료들을 분류한다고 가정했을 때, 변수 자료에 대한 퍼지 소속 등급 행렬은 식 (9)과 같다.

$$U = \{u_1, u_2, \cdots, u_c\} \in R^{c \times n} \tag{9}$$

여기서 $u_1(x_k)=u_{ik}$를 클래스 (i)에 속하는 변수 자료 x_k의 소속 등급이라고 하면 U는 식 (10)와 같이 표현된다.

$$U = \begin{pmatrix} & x_1 & \cdots & x_k & \cdots & x_n \\ 1) & & & \downarrow & & \\ \vdots & & & \downarrow & & \\ i) & \rightarrow & \rightarrow & u_{ik} & \rightarrow & \rightarrow \\ c) & & & \downarrow & & \end{pmatrix} \begin{matrix} \\ \\ \\ \text{객체} \\ \\ \text{클래스} \end{matrix} \tag{10}$$

퍼지 소속 등급 행렬 U의 가로는 클래스(i)에 해당하는 모든 객체의 소속 등급이며, 세로는 k번째 객

체에 대한 모든 클래스의 소속 등급이다.

$$U = \{u_{ik}\} \in M_{fuzy}(\text{FCM})=\text{미지의 퍼지 c-분할} \tag{11}$$

$$u_{ik} = \left\{ \sum_{j} \frac{|X_k - V_i|_A}{|X_k - V_j|_A} \right\}^{-2/(m-1)} \tag{12}$$

소속등급의 갱신 : [0,1]의 분포 (행렬의 세로방향)

$$V = \{V_i\} = R^p = \text{미지의 클러스터 중심} \tag{13}$$

$$V_i = \Sigma u_{ik}^m \cdot X_k / \Sigma u_{ik}^m = \Sigma \beta_k \cdot X_k \quad (\Sigma \beta_k = 1) \tag{14}$$

FCM 알고리즘의 목적함수J_m은 다음과 같고, 이를 최소화하는 (U, V)를 구하는 것이 그 목표이다. J_m은 식 (15)과 같이 구한다(Didier Dubois et al, 1997).

$$Min\left\{ J_m(U, V; X) = \sum_{k=1}^{N} \sum_{i=1}^{C} u_{ik}^m (|X_k - V_i|_A)^2 \right\} \tag{15}$$

식 (15)에서 C는 클러스터 개수, N은 입력 데이터 개수, m은 퍼지 정도를 나타내는 가중치(weighting exponent)이다. $\|X_k - V_i\|_A^2$은 입력 패턴 X_k와 군집의 중심인 V_j사이의 거리로서 유클리드 거리를 사용하고, 센터 값은 각 그룹 내의 데이터들의 평균값을 의미하며, U_{ik}^m는 군집 X_k와 입력 패턴 사이의 소속 등급(membership grade)이다.

3. 군집 분석(Cluster Analysis)

군집 분석은 많은 변수들을 일정한 속성에 따라 서로 유사한 것끼리 군집화(clustering) 하거나 상관관계가 큰 변수들끼리 집단으로 묶는 통계적 분석기법이다. 즉, 각 변수가 미리 정해진 기준에 맞추어 각 군집 내에 유사한 것들을 몇몇의 집단으로 그룹화하여, 각 집단의 성격을 파악함으로써 데이터 전체의 구조에 대한 이해를 돕고자 하는 방법이다(박광배, 2000 ; 원태연, 2004). 변수들 간의 유사성(similarity) 또는 이와 반대 개념인 거리(distance)에 근거하여 변수들을 집단으로 군집화 한다. 판별분석에 따른 분류(classification)와 다른 점은 분류는 이미 알려진 그룹의 구조와 수대로 나누어 각 그룹에 새로운 변수를 할당하는 것이 목적이지만, 군집분석은 그룹의 수나 구조가 가정되어 있지 않고 무엇을 기준으로 해서 데이터를 분류하느냐에 따라서 결과는 달라진다.

요인분석과 비교해 보면 군집분석은 설문에 응답한 응답자간의 상관관계를 통하여 유사한 응답을 한 집단끼리 묶어 집단을 구성하여 주는 방법이나, 요인분석은 변수간의 상관관계를 통하여 상호 상관관계가 높은 변수들끼리 통합하여 요인으로 묶는다는 점에서 구별된다.

군집화 과정은 우선 초기 군집을 정의하고 이들 군집간의 거리가 가까운 것끼리 다시 군집을 형성한다. 이 때 유사한 성격을 가진 변수들은 동일한 군집에 속하여 상이한 성격의 변수는 상이한 군집에 할당되어야 한다. 여기서 변수들 간의 거리는 유클리드 거리를 사용했으며, 군집들 간의 거리는 평균연결법을 사용한다(원태연, 2004). 본 논문에서는 요인분석을 통하여 사전에 변수들 간의 중복부분을 제거한 순수요인들을 도출하여 군집분석의 변수로 사용하였다.

가. 군집분석의 적용

복잡한 자료 집합에서 간단한 군집 구조를 만들어 내기 위하여 "근접도(proximity)" 또는 "유사도" 등의 측도가 필요하다. 유사도의 선택은 매우 주관적이기 때문에 변수의 성질(이산, 연속, 이항) 또는 측정척도(명목, 순서, 구간, 비) 및 주제 관련 지식 등을 고려해야 한다. 항목(변수 또는 사례)들의 군집화에 사용되는 근접도는 일종의 거리에 의해 나타내진다. 반면에 변수들은 상관계수 또는 연관성에 대한 유사도에 의해 군집화가 이루어진다(박광배, 2004). p차원 상의 두 관찰 값 $X=(x_1, x_2, \cdots, x_p)$와 $Y=(y_1, y_2, \cdots, y_p)$간의 거리 $d(X_p, Y_p)$는 그 값이 적을수록 거리가 가까움을 알려준다. 유클리드 거리는 식 (16)과 같이 정의 된다.

$$d(X, Y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots(x_p-y_p)^2} \\ =\sqrt{(X-Y)'(X-Y)} \tag{16}$$

본 논문에서 적용하는 클러스터링 방법으로 각 변수에서부터 시작하여 유사한 것들끼리 계속 묶어가는 응집 계층적 군집 방법을 사용할 것이다. 이의 절차는 다음과 같다.

1. 변수의 수가 N일 때 각 변수들 간의 거리들로 이루어진 $N\times N$ 대칭 행렬 $D=\{d_{ik}\}$를 구한다.
2. 거리 행렬 D에서 가장 가까운 거리를 갖는 쌍을 찾는다. 이 때 가장 근접한 쌍이 U, V라고 하자.
3. U, V를 합치고 난 후 거리 행렬을 조정한다. 우선 U, V에 해당하는 행과 열을 없애고, U, V와 나머지 클러스터에서 거리행렬을 만든다.
4. 1~3의 단계를 N-1개의 모든 변수가 한 군집이 될 때까지 반복한다.

위의 절차에서 두 변수로 한 군집을 이룬 뒤 그 군집과 다른 군집과의 거리는 여러 가지 방법으로 결정될 수 있다. 가장 많이 사용되는 방법으로 단일연결법(최소거리 기준, single linkage), 완전연결법(최장거리 기준, complete linkage), 평균연결법(average linkage) 등이 존재하는데, 본 논문에서는 군집 간 평균연결법을 적용하였다. 군집 간 평균연결법은 한 군집 내에 속해 있는 모든 변수와 다른 군집 내에 속해 있는 모든 변수의 쌍집합에 대한 거리를 평균적으로 계산한다. 또한 새로운 객체를 기존의 군집에 포함시킬 때 새로운 객체를 포함한 모든 객체 간의 평균거리가 최소가 되는 군집에 포함시키는 방법이다. 즉, 군집내의 객체들의 응집에 더 중점을 두는 군집화 방법이다[9]. 거리 행렬 $D=\{d_{ik}\}$에서 가장 가까운 거리의 변수가 U, V라 할 때 (U, V)를 한 군집으로 묶는다. 그리고 또 다른 군집 W와의 거리는 식 (17)과 같이 정의된다.

$$d_{(U, V)W}=\frac{\sum\sum d_{ik}}{N_{(U, V)}N_W} \tag{17}$$

식 (17)에서 d_{ik}는 (U, V)군집의 i변수 (i=U 또는 V)와 W군집의 k변수간의 거리를 뜻한다. 또한 $N_{(U, V)}$, N_W는 각각 (U, V)군집의 변수의 수, W군집의 변수의 수를 의미한다.

III. 대표 문항 추출 예

본 논문에서는 강의평가서 작성을 위해 대표 문항 추출 방법을 적용하였다. 문항 추출에 사용된 자료

는 명지대학교 외 3개 대학교 학생 646명을 대상으로 실시한 120 문항의 강의 평가서 (원강의 평가서)이며 이들로부터 약 25 문항 정도가 추출되었다. 원강의 평가서의 내용은 과목의 내용(A내용), 교수법/설명(B설명), 교수법/상호교류(B교류), 교수법/준비(B준비), 교수법/구성 및 진행(B진행), 교수(C교수), 평가 및 과제/평가(D평가), 평가 및 과제/ 과제물(D과제), 평가 및 과제/교재(D교재), 학생(E학생), 시설 및 장비(F시설), 조교 및 보조 스태프(G조교), 기타(H기타) 등의 임의의 13개 그룹으로 형성되어 있다. 〈표 1〉은 과목의 내용에 대한 문항들을 예시하고 있다. 학생들은 주어진 120개의 문항에 대하여 "매우 그렇다"에서 "해당 없음"까지 6등급으로 응답하였다. 실험을 위해 120개의 각 변수에는 해당하는 그룹명과 일련번호를 사용하였다. 현실적으로 사용 가능한 강의 평가서는 그 질의 항목이 20개 전후라는 점을 감안하여, 본 논문에서는 전술한 세 가지의 클러스터링 기법을 사용하여 대표 문항을 추출하였다.

〈표 1〉 원강의 평가서의 문항의 예

(1) 과목의 내용
① 이 과목의 내용은 전반적으로 많은 도움이 되었다. **(A내용1)**
② 과목의 내용이 시대에 뒤떨어지지 않았다. **(A내용2)**
③ 이 과목을 후배에게 추천하고 싶다. **(A내용3)**
④ 과목 내용은 실무 적용이 가능하다. **(A내용4)**
⑤ 과목에서 배우는 분량이 적절하다. **(A내용5)**
⑥ 과목 자체가 너무 어렵거나 너무 쉽지 않았다. **(A내용6)**
⑦ (프로젝트) 실무에 활용이 가능하다고 생각한다. **(A내용7)**
⑧ (프로젝트) 본인 및 팀원의 실력에 비추어 적절한 난이도를 가졌다. **(A내용8)**
⑨ (실험/실습) 관련 이론 강의와의 관계는 적절하였다. **(A내용9)**

1. 요인 분석의 적용 결과

요인분석의 첫 번째 과정은 여러 개의 변수가 동시에 서로 어떻게 관련되어 있는가를 알아보기 위하여 120개 문항의 상관관계행렬을 계산한다. 〈표 2〉는 평가서의 13개 그룹 중 A그룹에 대한 상관관계행렬이다. 이 행렬을 분석한 결과 기초요인행렬이 만들어지며, 이 행렬 내의 원소는 요인계수라 하고 변수와 요인간의 관계를 나타낸다. 상관관계는 다음과 같이 계산된다(이상호, 2002).

$$r = \frac{S_{xy}}{S_x \cdot S_y}$$

$$S_{xy} = X\text{와 } Y\text{의 공분산} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$S_x = X\text{의 표준편차} = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}}$$

$$S_y = Y\text{의 표준편차} = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n-1}}$$

〈표 2〉의 상관계수를 살펴보면, 변수 A내용1과 A내용3은 상관관계가 0.649로 상당한 상관이 있고, 변수 A내용1과 A내용7은 0.143으로 상관이 거의 없다고 해석될 수 있다. A내용1의 물음은 "이 과목의 내용은 전반적으로 많은 도움이 되었나?"이고, A내용3은 "이 과목을 후배에게 추천하고 싶은지?"에 대한 물음이며, 두 문항은 매우 비슷한 문항으로 분류할 수 있는 것이다. 또한 A내용7과 A내용8이 0.898의 상관관계를 가진다. 이를 정리하면, A내용1과 A내용2, A내용3을 한 클러스터로, A내용7과

A내용8을 또 다른 클러스터로 하는 것이 바람직하다.

〈표 2〉 상관관계행렬

	A내용1	A내용2	A내용3	A내용4	A내용5	A내용6	A내용7	A내용8	A내용9
A내용1	1.000	.547	.649	.466	.436	.319	.143	.166	.221
A내용2	.547	1.000	.524	.470	.387	.295	.134	.146	.173
A내용3	.649	.524	1.000	.482	.502	.368	.122	.156	.199
A내용4	.466	.470	.482	1.000	.389	.233	.194	.114	.175
A내용5	.436	.387	.502	.389	1.000	.387	.168	.211	.200
A내용6	.319	.295	.368	.233	.387	1.000	.212	.243	.217
A내용7	.143	.134	.122	.194	.168	.212	1.000	.898	.751
A내용8	.166	.146	.156	.114	.211	.243	.898	1.000	.777
A내용9	.221	.173	.199	.175	.200	.217	.751	.777	1.000

〈표 3〉 설명된 총분산

Compo-nent	Initial Eigenvalues			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumul-ative %	Total	% of Variance	Cumul-ative %
1	3.697	41.072	41.072	3.180	35.337	35.337
2	2.135	23.717	64.790	2.651	29.453	64.790
3	.822	9.139	73.928			

위의 〈표 3〉은 주성분 분석 요인 추출법에 의해 설명된 총분산이다. 초기 고유값(Initial Eigenvalues)은 요인변수의 크기를 나타내는 수치로서 이것이 크면 그룹 중에서 차지하고 있는 비중이 높음을 의미한다. %분산(% of Variance)은 각 고유치가 설명하는 분산의 양으로서 해당 변수의 공통분산에서 차지하는 비율이 된다. 따라서 요인으로 분류될 경우 이 값이 높을수록 독자적인 요인값으로서의 힘이 큰 것이다. 본 논문에서는 요인 선택 시 고유값이 1이상인 것을 요인수로 정하기로 한다. 따라서 여기에서는 2개의 요인이 추출될 수 있는 것이다. 이 경우 요인 중 성분1이 설명할 수 있는 양은 41.072%이고 성분2가 23.717%에 해당되어 전체적으로 요인에 의해 총 분산의 64.790%가 설명될 수 있음을 나타낸다.

아래의 〈표 4〉는 앞의 과정에서 추출한 2개의 요인을 나타내며, 이는 회전 후 적재값(Cumulative)을 나타낸다. 이는 변수들과 요인의 관계를 보다 선명하게 드러날 수 있도록 하여 요인의 회전을 용이하게 한다. 본 논문에서는 VARIMAX 기준의 직교 회전 방법을 택하였으며, 이 때 각 변수들은 요인의 축이 회전함에 따라 새로운 변수값을 가지게 된다. VARIMAX 회전은 요인의 분산을 최대화된다. 이 요인계수들의 분산을 구할 때는 우선 요인계수를 제곱한 것을 원자료로 해서 계산한다. 다음은 VARIMAX 기준을 수학적으로 표시한 것이다. 번째 요인에 대한 계수들의 분산을 S_p^2는 아래와 같이 얻는다(이순묵, 1995).

$$S^2 = \Sigma(ti - \bar{t})^2/n \tag{22}$$

ti는 요인의 세로줄에 있는 계수들을 제곱한 값이고 i는 1, 2, …, n(n은 변수의 수효)이며 t는 한 요

인의 세로줄 내에서 t값의 평균이다. 이때 m개의 요인에 대한 분산의 합은 식(22)와 같다.

$$S^2 = \Sigma S_p^2,\ p = 1, \cdots, m \tag{23}$$

여기서 m은 요인의 수효이다. 이 공식을 Kaiser(1958)은 "raw" VARIMAX기준이라고 한다. 즉 S^2이 최대로 되게 하는 회전기준이 원래의 VARIMAX기준이다. 요인구조를 돌릴 때, m개의 요인이 있으면 두 개씩 요인의 켤레를 만들어 돌리게 되므로 전체 요인의 켤레는 $[m(m-1)/2]$개가 된다. 이 전체 요인의 켤레들을 VARIMAX기준의 값이 최대한으로 커져서 더 이상 변화가 없을 때까지 거듭해서 돌리는 것이 VARIMAX기준에 의한 회전이다.

회전 한 결과 요인 1은 변수 A내용1, 2, 3에 대한 설명변량(variance account for : VAF)이 크다는 것을 알 수 있고, 변수 A내용7, 8, 9에 대해서는 설명변량이 작다고 말할 수 있다. 반면에, 요인 2는 변수 A내용7, 8, 9에 대해서는 많은 설명력을 가지고, 변수 A내용1, 2, 3에는 적은 설명력을 가짐을 볼 수 있다. 요인 분석 결과 2개의 요인이 추출되었고, 각각의 대표 문항을 선정할 수 있다

〈표 4〉 회전된 요인 행렬

	Component	
	1	2
과목 내용의 도움 정도(A내용1)	.804	7.436E-02
내용의 시대성(A내용2)	.753	5.112E-02
추천정도(A내용3)	.831	5.332E-02
실무 적용 가능성(A내용4)	.698	7.318E-02
분량의 적절성(A내용5)	.689	.135
과목의 어려움 정도(A내용6)	.518	.224
(프로젝트) 실무적용가능성(A내용7)	9.338E-02	.942
(프로젝트)실력에 따른 난이도(A내용8)	.107	.951
(실험/실습) 이론 강의와의 적절성(A내용9)	.162	.880

2. FCM 알고리즘의 적용 결과

일반적인 FCM 알고리즘은 주어진 변수 자료 집합 $X = \{x_1, x_2, \cdots, x_n\}$에 대해 다음과 같은 과정을 수행한다. 여기서 x_k는 k번째 문항에 대한 응답을 N차원 상에서 보여준다. 여기서 N은 응답자의 수를 의미한다.

[FCM/HCM 알고리즘](Joseph Hoey, 2000).

〈표 5〉에서는 A내용 그룹을 FCM한 후, 3개의 클러스터로 분류된 것을 볼 수 있다. 첫 번째 클러스터에는 변수 A내용5, 6이 비슷한 소속함수 값을 갖고, 두 번째 클러스터는 변수 A내용7, 8, 세 번째 클러스터에는 변수 A내용1, 2, 3이 비슷한 소속함수 값을 가짐으로 각각 클러스터링 되어 있는 것을 알 수 있다. 한편 변수 A내용4의 경우는 첫 번째와 세 번째 클러스터에 속하는 값이 거의 동등하여 어느 클러스터에 포함되어 있다고 표현할 수 없다. 따라서 표 5의 실행결과로부터 A내용에서는 3개의 대표 문항이 추출되는 것을 확인할 수 있다.

〈표 5〉 FCM 실행 결과

	A내용1	A내용2	A내용3	A내용4	A내용5	A내용6	A내용7	A내용8	A내용9
첫 번째 클러스터 소속함수값	0.2989	0.3406	0.3759	0.4611	**0.6048**	**0.5747**	0.0695	0.0594	0.1985
두 번재 클러스터 소속함수 값	0.0449	0.0487	0.0479	0.0748	0.0634	0.0873	**0.8664**	**0.8861**	0.6128
세 번째 클러스터 소속함수 값	**0.6562**	**0.6107**	**0.5762**	0.4641	0.3319	0.3381	0.0641	0.0544	0.1887

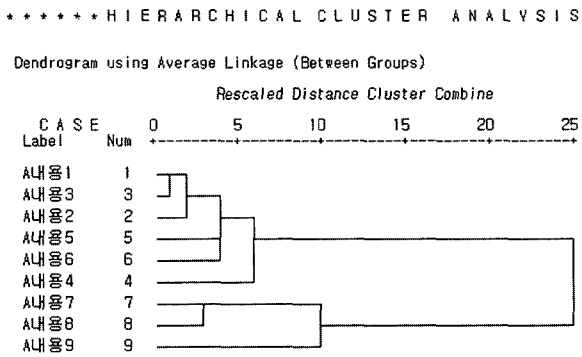
3. 군집 분석의 적용 결과

군집분석에 의하여 학생들의 응답이 군집화되어 가는 과정이 〈표 6〉에 표현되어 있다. 조합된 군집은 매 단계마다 남아 있는 가장 적은 계수(유클리드 거리)의 군집쌍으로 형성이 된다.

〈표 6〉 군집화 일정표

단계	조합된 군집		계수	최초출현 군집단계		다음단계
	군집 1	군집 2		군집1	군집2	
1	1	3	106.000	0	0	2
2	1	2	150.000	1	0	5
3	7	8	175.000	0	0	7
4	5	6	182.000	0	0	5
5	1	5	202.000	2	4	6
6	1	4	227.000	5	0	8
7	7	9	332.500	3	0	8
8	1	7	710.056	6	7	0

〈표 6〉에서와 같이 계수는 해당 변수들이(A내용1 ~ A내용9) 속해 있는 군집간의 거리정도를 나타내므로, 이 값이 클수록 군집화가 늦어지게 된다. 따라서 이 값이 가장 작은 변수 A내용1과 A내용3이 단계 1에서 군집화 된다. 단계 1에서 결합된 A내용1과 A내용2의 군집은 단계 2에서 결합되고 있음을 알 수 있다. 마지막 단계인 단계 8에서는 A내용1와 A내용7이 군집화되고 있다.



〔그림 1〕 덴드로그램

위 〈그림 1〉은 평균결합방식으로 결합된 덴드로그램(dendrogram)이다. 여기에서 세로축은 변수, 가로축은 상대적 거리를 나타내고 있다. 군집화 과정을 살펴보면 A내용1과 A내용3, A내용2는 하나의 클러스터로, A내용7과 A내용9는 또 다른 클러스터로 형성됨을 알 수 있다. 즉, 2개의 군집으로 분류될 수 있으며, 각각의 대표 문항을 추출할 수 있다.

4. 적용 결과 비교

본 절에서는 이상에서 설명한 세 가지의 클러스터링 방법, 즉 요인분석, FCM 알고리즘, 또 군집분석 등을 적용한 방법들의 결과를 비교해 보자. 〈표 7〉에 나타난 바와 같이 요인분석과 군집분석은 동일한 클러스터를 형성하였고, FCM 알고리즘을 사용한 경우는 A내용5와 A내용6이 하나의 클러스터를 추가로 더 생성했음을 확인할 수 있다.

〈표 7〉 적용 결과 비교

분석방법 / A 내용	요인분석	FCM 알고리즘	군집분석
1	1	1	1
2	1	1	1
3	1	1	1
4			
5		2	
6		2	
7	3	3	3
8	3	3	3
9			

〈표 8〉 대표 문항 추출

	대표 문항
과목의 내용	이 과목을 후배에게 추천하고 싶다.
	과목에서 배우는 분량이 적절하다.
	(프로젝트) 본인 및 팀원의 실력에 비추어 적절한 난이도를 가졌다.

〈표9〉 클러스터링 알고리즘을 적용했을 때의 결과

항 목	클러스터링 결과
A내용	FA=GA, FCM은 FA, GA보다 1개의 문항이 더 추출됨.
B설명	FA=FCM, GA는 FA, FCM보다 1개의 문항이 더 추출됨.
B교류	FA=FCM=GAB준비FA=FCM=GA
B진행	FA=FCM=GA
C교수	FCM=GA, FA는 1개의 문항이 더 추출됨. 클러스터의 구분이 다소 명확하지 않음.
D평가	FA=FCM=GA, 추출되지 않은 문항들도 대표 문항과 비슷한 효력을 나타냄.
D과제	FA=FCM, GA는 2개의 대표 문항중 1개의 문항이 추출되지 않음. 독자적인 대표 문항이 추출됨.
D교재	FA=FCM=GA
E학생	FA=FCM=GA
F시설	FA=FCM=GA
G조교	FA=FCM=GA

FA=요인분석, FCM=FCM 알고리즘, GA=군집분석

따라서 〈표 8〉은 세 가지 클러스터링 기법을 적용한 방법들의 결과를 조화시킨 최종 대표 문항들이다. 한편 표 9는 내용A뿐 만 아니라 나머지 12개의 집단에 대하여 세 가지의 클러스터링 알고리즘을 적용했을 때 발생한 결과를 요약하였다.

아래의 〈표 9〉에서 항목이 나타내는 것은 원강의 평가서 내용들이다. B교류, B준비, B진행, D교재, E학생, F시설, G조교의 항목에서 선정된 대표 문항은 세 가지 분석 기법의 결과가 모두 같다. 이는 추출된 대표 문항이 응답자에게 응답하기 수월한 문장으로, 효과적인 설문 추출을 하였다고 볼 수 있다. 다른 항목중 A내용의 경우는 요인분석(FA)과 군집분석(GA)은 같은 대표 문항을 추출하였으나. FCM 알고리즘에서는 1개의 문항이 더 추출되었음을 알 수 있다. 또한, D과제 항목에서는 군집분석에서만 다른 분석 기법에서보다 적은 수의 대표 문항이 추출되었음을 나타낸다.

IV. 결 론

본 논문에서는 설문 추출을 위한 효과적인 방법을 찾기 위해 군집분석이외에도 요인분석과 FCM(Fuzzy C-Means) 알고리즘을 적용하였는데 이들은 궁극적으로 클러스터링(Clustering)을 이용하는 방법이다. 요인분석은 상관계수와 주성분을 이용하여 문항수를 축약하고, FCM 알고리즘은 특정 클러스터 중심과의 소속 함수값에 따라 변수들을 분류한다. 군집분석은 학생들의 평가 점수간의 거리에 근거하여 클러스터를 생성한다. 이러한 분석 방법의 결과, 설문지의 120개 문항이 25개의 문항으로 축소가 가능함을 확인하였다.

향후 과제로는 요인 분석 결과에 의한 새로운 설문지를 작성하는 것이다. 이를 이용하여 여러 학생들에게 실제적인 설문을 실시하는 것은 흥미로운 일이다. 마지막으로 FCM 알고리즘과 군집분석 간의 클러스터링의 성능 분석을 통해 두 방법 간의 장단점을 파악하는 것도 실제 적용할 때 큰 도움이 되리라 본다.

[참 고 문 헌]

나성호(2002). 고객세분화를 위한 군집분석 기법 중 K-평균 군집분석과 코호넨 네크워크의 분류 성능에 관한 연구. 서울대학교 석사학위논문

김대원 · 이광형(2003). A Cluster Validity Index Using Overlap and Separation Measures Between Fuzzy Clusters. 한국 퍼지 및 지능시스템 학회 논문지, 13(8). pp. 455-460.

이순묵(1995). 요인분석 I. 학지사.

이순묵(2000). 요인분석의 기초 14. 교육과학사, Chap. 8.

이영준(2002). 요인분석의 이해. 석정, Chap. 14.

이상호(2002). 자료분석의 기초이론. 강원대학교 출판부.

노형진(2002). 한글 SPSS 10.0에 의한 조사방법 및 통계분석. 형설출판사.

박광배(2004). 다변량 분석. 학지사.

원태연(2004). 다변량 시각화 데이터 분석법. 교우사.

이순묵(1996). 설문조사법. 자유 아카데미.

Charles D. & Fyfe C.(2000). *Kernel Factor Analysis with varimax Rotation.* IEEE Int. Conference on Neural Networks IJCNN-2000.

Frank HÖppner, Frank Klawonn, Rudolf Kruse & Thomas unkler(1999). *Fuzzy Cluster Analysis*. WILEY COMPUTER PUBLISHING.

Joseph Hoey(2000). Introduction to Survey Research. *Rose-Hulman Institude of Technology*, pp. 11-28.

Didier Dubois, Henri Prade, and Ronald R.Yager(1997). *Fuzzy information Engineering*. WILEY COMPUTER PUBLISHING