文獻의 自動分類를 위한 판별분류 시스템 설계

金 賢 姬*
李 龍 禮**

<목 차>

Ⅰ. 序 論

1. 硏究의 目的

연구대상의 객체들을 어떤 관점에서 分類하는 것은 과학적 연구에서 가장 본원적 목표들 중의 하나이다. 특히 生物을 특성에 따라 분류하는 생물분류학, 석기(Stone tools)나 화석 등에 근거하여 문화발달과정을 기술하는 인류학, 그리고 문헌들을 主題別로 분류하는 도서관학·정보학에서는 분류학은

*명지대학교 도서관학과 부교수
**명지대학교 사회교육원 강사

1) M. S. Aldenderfer & R. B. Blashfield, "Cluster Analysis," Sage Univ. Paper Series on Quantitative Applications in the Social Sciences, Beverly Hills and London : Sage Pubns, 1985, pp. 7~16.

핵심 연구분야라고 할 수 있다.[1]

일반적으로 문헌의 분류를 위해서는 문헌의 표제, 초록, 또는 본문에 출현한 단어나 인용문헌이 屬性으로 선택되며, 또한 저자명과 문헌이 실린 학술잡지명 등도 분류에 도움이 되는 知識을 제공하기도 한다.[2]

문헌 분류는 수작업으로 수행해 오다가 1960년대에 들어서 자동 분류의 개념이 발전하기 시작하였다.[3] 문헌이나 용어 등을 주어진 분류체계에 따라 자동분류한 선행연구를 살펴보면, 먼저 Maron은 실험을 위해서 컴퓨터분야 문헌의 초록들을 실험문헌집단으로 정하고 분류체계는 임의로 32개의 주제카테고리로 설정하였다.[4] 그는 실험문헌집단을 각 주제카테고리로 분류한 후, 단어들을 분석하여 端緖語를 선택하고 단서어와 주제카테고리와의 관련도는 특정한 주제카테고리에 속하는 문헌들속에 단서어가 출현한 횟수로 정하였다. 소속 주제카테고리가 알려져 있지 않는 문헌을 분류할 때는 이 문헌이 갖는 단서어를 근거로 각 주제카테고리에 속할 확률을 계산한 다음 이 확률값이 가장 큰 주제카테고리에 문헌을 분류하였다.[5]

Borko와 Bernick은 Maron의 연구와 동일한 실험문헌집단을 이용하여 문헌의 자동 분류를 시도해 보았는데 Maron의 연구와 다른 점은 要因分析을 이용하여 분류체계를 구성한 점이다.[6] Borko등은 Maron의 실험에서 선택한 90개의 단서어를 색인어로 선택하고 이 90개의 색인어가 실험문헌집단에서 출현한 빈도수를 기술한 문헌—색인어 행렬을 이용하여 용어와 용어간의 90次 상관계수행렬을 구하였다. 이 상관계수행렬을 입력물로 해 요인분석한 결과 90개의 색인어가 21개의 요인으로 변환되었고, 각 요인에 적절한 主題名을 부여하여 주제카테고리를 구성하였다. 색인어와 주제카테고리와의 관련도는 표준화된 요인적재량으로 구했으며, 새로운 문헌을 분류할

2) K. A. Hamil & A. Zamora, "The Use of Titles for Automatic Document Classification," JASIS 31(6) : 396~402 ; 1980.
3) 정영미, 정보검색론, 서울 : 정음사, 1987, p. 181.
4) M. E. Maron, "Automatic Indexing : An Experimental Enquiry," JACM 8(3) : 404~417 ; 1961.
5) Ibid.
6) H. Borko & M. Bernick, "Automatic Document Classification," JACM 10(1) : 151~162 ; 1962.

때는 문헌에 출현한 색인어와 표준화된 요인적재량의 적을 모두 더해 준 값을 구한 다음 이 값이 가장 큰 주제카테고리(요인)로 문헌을 분류하였다.[7]

Hamil과 Zamora는 화학분야 문헌을 실험문헌집단으로 하고 분류체계는 CA(Chemical Abstracts)의 주제분류인 80개의 주제카테고리를 사용하여 문헌의 자동 분류를 시도하였다. 이들은 먼저 CA에 이미 색인되어 있는 문헌들의 표제, 초록, 키워드구로부터 단어를 추출하여 각 단어의 80개 주제카테고리내의 分布패턴을 구한 다음, 이 분포패턴을 이용하여 단어가 각 주제카테고리에 출현할 확률값을 구하고 이 값에 100을 곱하여 단어와 주제카테고리와의 관련도 가중치로 삼았다. 문헌을 분류할 때에는 문헌의 표제에 출현한 단어를 근거로 문헌과 각 주제카테고리와의 관련도를 산출하고 관련도의 값이 가장 큰 주제카테고리에 문헌을 분류하였다.[8][9]

Dillon과 Federhart은 앞의 연구들과는 달리 분류 대상으로 문헌대신 문헌에서 분석한 용어를 주어진 3개의 카테고리에 자동으로 분류하는 작업을 시도하였다. 구체적으로 이들은 判別(分類)分析을 이용하여 Harris설문지집단에서 추출한 어근들을 주제어, 類似주제어, 일반어의 3개 카테고리로 분류하여 색인어로 사용할 수 있는 의미있는 어근들을 가려내었다. 이때 판별변수로 사용된 것은 어근의 출현빈도외에 의미있는 어근이 출현하는 질문들은 무의미한 어근이 출현하는 질문들보다 더 밀접하게 관련되어 있을 것이라는 가정하에 어근이 출현한 모든 질문간의 평균 연관도(average of measure of association), 연관도 분포의 표준 편차(standard deviation) 및 첨도(skewness) 등을 이용하였으며 질문간의 연관도는 질문짝에 공통으로 출현한 어근에 기초하여 산출하였다.[10][11]

本稿에서는 판별분류분석기법을 이용해서 유기화학에 대한 주제 배경이 없

7) Ibid.
8) K. A. Hamil & A. Zamora, op. cit.
9) 정영미, op. cit., pp.185~186.
10) M. Dillon & P. Federhart, "Statistical Recognition of Content Terms in General Text," JASIS 35(1) : 3~10 ; 1984.
11) ———, "The Use of Discriminant Analysis to Select Content Bearing Words," JASIS 33(4) : 245~253 ; 1982.

는 분류자는 물론 주제 지식이 있는 정보전문가에게 분류 지식을 제공하면서 컨설팅해 줄 자동분류시스템을 구축하고자 한다.

2. 硏究의 方法

본 연구에서는 문헌의 자동분류시스템을 구축하기 위해서 판별분석을 이용했으며, 또한 집단의 數와 構造에 대한 事前 정보가 있을 경우 판별분석이 군집분석보다 더 정확하게 문헌을 분류할 수 있다는 가정을 검증하기 위해서 동일한 유기화학 문헌집단을 이 두 기법을 이용하여 9개의 下位主題別로 분류해 보았다.

실험 데이타로는 유기화학에 관한 269개의 문헌을 선택하여 218개는 실험문헌집단을 구성하고 51개는 검증문헌집단을 구성하였다.[12] 분류카테고리는 CA에서 사용한 14개의 하위주제 카테고리中 핵심 單語群이 다양한 유기화학일반과 유기물리화학을 제외하고, 脂肪族·脂環式化合物, 非縮合·縮合芳香族化合物, 複素環式化合物(異原子 1個 또는 2個 이상)을 한데 묶어 총 9개의 하위주제 카테고리를 설정하였다. CA에 이미 색인되어 있는 실험문헌집단의 표제, 키워드구로부터 불용어 및 기능어를 제외한 단어를 추출하여 핵심 단어 또는 단어 어근을 共有하는 단어들을 동일한 단어그룹으로 분류하여 47종의 단어그룹을 선정하여 각 단어그룹이 공유하는 단어 또는 단어 어근을 판별변수로 간주하였다. 이 47종의 판별변수中 판별에 貢獻度가 낮은 9종의 판별변수를 제외시켜 최종적으로 38종의 판별변수를 이용하여 분류시스템을 구축하였다.

이 시스템을 가지고 새로운 문헌의 표제를 이용하여 분류한 결과, 분류의 正確度는 판별변수 선택에 사용된 실험문헌집단(218개 문헌)의 경우 91%, 검증문헌집단(51개 문헌)의 경우 84%로 나타났다. 또한 검증문헌집단을 군집분석기법으로 9개의 군집으로 분류한 결과 분류의 정확도는 57%로 판별분석을 사용한 결과인 84% 보다 27%가 낮았다. 데이타 처리는 SAS 통계 패

12) 원래의 검증문헌집단의 문헌수는 90개이었으나 38개의 판별변수가 주제를 커버하는 문헌은 그 중 51개로 주제 커버율이 57%(51/90)이다.

키지를 이용하여 처리하였다.

3. 假 說

본 연구에서 검증하고자 하는 가설은 다음과 같다.

① 판별분석을 이용하여 유기화학문헌을 9개의 하위주제 카테고리별로 분류할 수 있다.

② 집단의 수와 구조에 대한 사전 정보가 있을 경우 판별분석기법이 군집분석기법보다 더 정확하게 문헌을 분류할 수 있다.

Ⅱ. 多變量統計모델

1. 다변량 통계분석

1) 다변량통계분석이란?

다수의 客體에 대하여 둘 이상의 反應變數를 동시에 관측한 자료를 분석하는 학문이다.[13]

2) 다변량자료의 구조

(1) n 객체(object) 각각에 대해 p 개의 반응변수, $X_1, X_2, \cdots, X_p$ 를 측정한다.

(2) 자료행렬(data matrix) : X

$$X = \begin{array}{c} \\ (1) \\ (2) \\ \cdot \\ (i) \\ \cdot \\ (n) \end{array} \begin{array}{c} \text{변 수} \\ (1)\ (2)\cdots\cdots(j)\cdots\cdots(p) \\ \left(\begin{array}{l} X_{11}\ X_{12}\cdots\cdots X_{1j}\cdots\cdots X_{1p} \\ X_{21}\ X_{22}\cdots\cdots X_{2j}\cdots\cdots X_{2p} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ X_{i1}\ X_{i2}\cdots\cdots X_{ij}\cdots\cdots X_{ip} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ X_{n1}\ X_{n2}\cdots\cdots X_{nj}\cdots\cdots X_{np} \end{array}\right) \end{array} \quad (n \times p)$$

객체 / 체

13) 고려대학교 통계연구소, SAS를 이용한 통계 WORKSHOP, 서울 : 고려대학교 통계연구소, 1990, pp.121~152.

자료행렬 X의 각 객체(행)는 서로 통계적으로 독립이며 관찰된 반응변수(열)들은 서로 상관되어(correlated) 있다.

3) 다변량통계분석의 目的과 연관된 기법들

다변량통계분석의 중요한 두가지 목적은 자료/차원의 縮略을 통한 構造的 單純化와 동일 집단내에서는 높은 유사성을 갖고 서로 다른 집단간에는 높은 異質性을 가지도록 객체/변수를 分類(classification) 내지 集團化하는 것이다. 다음은 각 목적과 연관된 중요한 기법들을 기술한 것이다.

(1) 構造的 單純化

(A) 다차원 축적 : n 객체의 모든 가능한 짝들 간의 상사성(혹은 거리)이 주어져 있을 때 다차원 공간상(주로 2차원 혹은 3차원)에 조사된 객체들의 상대적인 위치를 좌표상에 나타내므로 形象化(configuration)하는 기법이다.[14]

(B) 요인분석 : 여러 개의 반응변수들을 보다 적은 수의 가설적 변수, 즉 요인으로 바꾸기 위한 통계적 기법을 의미한다. 환언하면 여러 개의 서로 연관되어 있는 원래 반응변수들 간의 복잡한 상관 구조를 "소수 몇개의", "공통의 뿌리를 가지는", "개념상 의미있는", 그리고 "상대적으로 독립적인" 공통요인을 통해 해석, 단순화시키고자하는 공분산(상관)중심의 기법이다.[15]

(2) 分類·集團化

분류 집단화와 연관된 기법에는 요인분석과 같이 변수간의 상관관계를 이용하여 유사한 변수들끼리 묶어주는 변수중심 분류기법과 판별분석 또는 군집분석과 같이 객체간의 상사성(proximity) 또는 거리에 근거하여 유사한 객체들끼리 모아 주는 객체중심 분류기법이 있다. 여기서는 앞에서 설명한 요인분석은 생략하고 판별분석과 군집분석에 대해서 설명한다.[16]

(A) 군집분석 : 군집들의 갯수나 구조에 관한 아무런 가정없이 객체들 사이의 상사성 또는 거리에 근거하여 '자연스러운' 군집을 찾고 나아가 자료

14) 이재창 & 박정섭, "다차원축척 (Multidimensional Scaling) 기법," 응용통계 1 : 61~80 ; 1986.

15) 김기영 & 전명식, SAS 입자분석, 서울 : 자유아카데미, 1990, pp. 1~14.

16) 고려대학교 통계연구소, op. cit.

의 요약을 꾀하는 원시적이고 탐색적인 통계방법이다.[17]

(B) 판별분석 : 이미 알려진 집단정보가 각 객체에 대해 일단 주어져 있을 경우 이들 집단간의 차이를 분석하고, 집단정보를 가지지 않은 새로운 객체를 이미 주어진 부분집단 중 하나에 분류하는 方法이다.[18]

2. 판별분석

1) 판별분석이란?

(1) 정의 : 직관이나 사전 정보에 의해서 집단이 구분되어 있을 때 그 집단을 분류(판별)하는 기준이 되는 판별함수를 도출하고 그 판별함수를 이용하여 임의의 표본에 대해 어느 집단에 속하는지를 糾明하고자 하는 통계분석기법이다. 판별분류분석은 두 부분(판별과정과 분류과정)으로 나눠서 생각하면, 우선 판별과정은 이미 주어진 다변량관찰값들로부터 전체 집단을 어떤 특성에 따라 상호 배반적인 부분집단 $G_1, G_2, \cdots, G_g$로 분류하기 위해 필요한 판별변수의 설정 및 이에 따른 해석과 연관된 과정을 뜻하며, 이에 대해 분류과정은 소속집단이 알려져 있지 않은 새로운 객체가 주어졌을 경우 앞에서 유도한 판별함수를 사용하여 이를 주어진 부분집단 $G_1, G_2, \cdots, G_g$ 중 어느 하나로 분류시키는 작업이라 할 수 있다. 그러나 위의 두가지 과정은 서로 분리되어 단독으로 처리되기 보다는 동시에 서로 복합되어 처리되는 경우가 많다.[19]

(2) 수학적 판별모델을 만들기 위한 기본과정

수학적 판별모델을 만들기 위하여 필요한 기본과정 몇가지를 살펴보면,

(A) 부분집단의 전체 갯수 $(g) \geq 2$,

(B) $0 <$ 판별변수의 수 $(p) <$ (전체표본의 크기 $(N) - 2)$,

(C) i번째 부분집단 G_i에 속하는 객체의 크기 $(N) \geq 2$,

(D) 판별변수$(X_1, X_2, \cdots, X_p)$는 區間尺度로 측정되며,

17) M. S. Aldenderfer & R. B. Blashfield, op. cit.

18) W. R. Klecka, "Discriminant Analysis," Sage Univ. Paper Series on Quantitative Applications in the Social Sciences, Beverly Hills and London : Sage Pubns, 1980, pp. 7~15.

19) Ibid.

(E) 각 판별변수는 다른 변수들의 線形결합이 되어서는 안되며,

(F) 판별변수 $X'=(X_1, X_2, \cdots, X_p)$가 다변량정규분포를 따라야 하며,

(G) 각 집단에 있어서 변수들간의 공분산행렬이 동일해야 하는 가정이 있다.

다른 변수의 선형결합으로 표현될 수 있는 변수는 그 변수만의 독립적인 판별정보를 전혀 제공하지 못 할 뿐더러 분석에서 필요한 행렬 조작상 문제를 야기시키기 때문에 가정(E)가 필요하다.[20] 위의 가정들중에서 가정(F)와 가정(G)를 충족시키기가 가장 어려운데 이 두 가정이 극단적으로 위배되지 않는 한 판별분석을 적용하는데 큰 지장은 없다. 특히 표본의 크기가 매우 클 때는 더욱 문제가 되지 않는다. 물론 이 두 가정이 다소 위배되면 판별분석의 결과가 그만큼 부정확하게 나타나게 된다.[21][22]

3. 판별분석절차

판별분석에 관련된 절차는 변수선정, 판별함수 계산, 판별분류 순이다.

1) 판별변수의 선정

(1) 目 的

때때로 연구자는 단순히 판별력을 가지고 있다고 생각되는 여러 개의 잠재적 변수에 관한 자료를 보유하고는 있으나 그 변수들이 실제로 지니고 있는 가치나 필요성에 대해서는 불확실해 하는 경우가 많다. 이때 그 중 어떤 변수는 집단간의 차이를 '충분히' 구별 못하는 변수로 판명될 경우도 있을 것이고, 혹은 개별적으로는 훌륭하나 다른 변수들과의 상관관계에 의해 그 변수만의 독립적인 판별 정보를 전혀 제공하지 못하는 경우도 있을 것이다. 따라서, 이와 같이 판별에 공헌도가 낮거나 중복되는 변수들은 어떤 논리적인 타당성이 인정되지 않는 한 분석과정에서 제외하는 것이 바람직하다.[23]

(2) 方 法

20) Ibid.
21) Ibid, pp. 60~63.
22) 김기영 & 전명식, SAS 판별 및 분류분석, 서울 : 자유아카데미, 1990, pp. 1~3.
23) Ibid.

판별변수의 선택방법으로 단계적 판별분석법이 있는데 이 방법은 다시 변수증가법, 변수감소법, 변수증감법으로 구분된다.[24]

(A) 變數增加法

가장 큰 판별력을 가지는 변수선택으로부터 시작해서 나머지 변수들 중에서 처음 선택된 변수와 짝을 이룰 때 가장 좋은 판별력을 가지는 변수를 택하는 方法이다.

(B) 變數減少法

일단 모든 변수들이 모델에 도입되었다고 생각하고 각 단계에서 가장 낮은 판별력을 가지는 변수가 제거되는 형식이다.

(C) 變數增減法

우선 변수증가법으로 시작하지만 각 단계마다 이미 앞서 선택된 변수들을 再考하게 된다. 만약 이들 중 판별에 충분한 공헌도가 인정되지 않는 변수들이 있다면 이들은 일단 제거되지만 완전히 고려 대상을 벗어나지는 않고 나중 단계에서 다시 선택의 대상이 될 수 있다.

2) 판별함수의 도출

판별분석은 두 개 이상의 집단 구분을 하는 데 있어서 구분오류를 최소화할 수 있는 판별함수를 이끌어 내는 데 촛점을 두게 된다. 판별함수는 판별변수들의 선형결합으로 이루어진다. 선형결합이란 아래식 (1)과 같이 각 독립변수(판별변수)에 일정한 가중치를 부여하고 이를 더한 형태를 띠고 있다. 판별함수의 수는 종속변수가 되는 집단수보다 하나가 적은 수가 된다.

정준판별함수 $D=B_0+B_1X_1+B_2X_2\cdots+B_iX_i$ …(1)[25]

여기서 선형함수의 계수 B_i는 판별점수(D)의 집단간분산(Between Group Sum of Square)／집단내분산(Within Group Sum of Square)이 최대화 되도록 함으로써 추정하게 된다.[26]

---

24) Ibid, pp.47～59.

25) 정준판별함수는 종속변수를 가변수의 종속변수들로 나타냈을 때, 독립변수들과의 정준 상관분석을 통하여 얻어진 정준함수와 같은 것이 된다는 의미에서 정준의 수식어를 사용한다. 정준상관(canonical correlation)은 판별변수들과 그룹들간의 관련정도의 측도이다.

26) M. J. Norusis, SPSS/PC+advanced statistics, Chicago : SPSS inc., 1986, B1～B40.

3) 分類過程

각 객체를 어떤 판정기준에 따라 사전에 규정되어 있는 部分集團에 적절히 분류하는 데는 여러가지 방법이 있으나 어느 한 객체를 그에 가장 '가까운' 집단에 분류시킨다는 기본 개념에는 큰 차이가 없다고 하겠다. 이때 가장 '가깝다'는 것은 일반적으로 그 객체와 집단중심(또는 평균)과의 '거리'의 최소를 의미한다.[27]

여기서는 위에서 언급한 정준판별함수(1)를 이용하여 분류하는 方法과 일반화된 거리함수와 사후 확률(posterior probability)에 의한 분류방법에 대해서 설명하고자 한다.[28]

(1) 정준판별함수에 의한 方法

정준판별함수를 이용하여 소속 집단이 알려져 있지 않은 객체 X를 분류할 때는 먼저 객체 X의 판별점수(D)를 구해야 한다. 객체 X의 판별점수(D)는 비표준화된 정준판별함수에 객체 X의 판별변수의 값을 대입하여 구한다.[29]

앞에서 구한 객체 X의 판별점수 (D)를 이용해서 객체 X가 각 집단에 속할 사후 확률을 구하고 확률값이 가장 큰 집단에 객체 X를 분류한다. 사후 확률을 구하는 공식은 다음과 같다.[30]

$$P(G_i|D)=\frac{P(D|G_i)\cdot P(G_i)}{\sum_{i=1}^{g} P(D|G_i)\cdot P(G_i)} \quad \cdots\cdots(2)$$

여기서 $P(G_i|D)$=판별점수가 D인 문헌이 집단 i에 속할 사후 확률

$P(G_i)$=문헌이 집단 i에 속할 사전 확률

$P(D|G_i)$=문헌이 집단 i에 속한다는 조건하에서 판별점수 D가 얻어질 조건확률. 이 조건확률은 각 집단에서 판별점수

27) W. R. Klecka, op. cit.
28) Ibid.
29) M. J. Norusis, op. cit.
30) Ibid.

의 분포가 정규분포에 따른다는 가정하에 정규분포의 母數値를 알면 구할 수 있다.

(2) 일반화된 거리함수와 사후 확률에 의한 방법.

일반화된 거리함수와 사후 확률에 의해 새로운 객체 X를 분류하기 위해서는 먼저 객체 X와 각 집단 중심까지의 거리를 아래의 공식 (3), (4)의 일반화된 거리함수로 구한다음, 이러한 일반화된 거리 $D_i^2(X)$를 아래의 공식 (5)에 대입하여 객체 X가 집단 G_i에 속할 사후 확률을 구하고, 확률값이 가장 큰 집단에 객체 X를 할당하게 된다.[31] 이는 결국 객체 X를 일반화된 거리 $D_i^2(X)$를 최소화하는 집단 G_i에 분류하게 된다. 여기서 흥미로운 점은 집단들의 분산구조가 서로 다를 경우, 그 사실을 거리 계산에 고려하여 집단들의 분산행렬이 다를 경우와 모두 같을 경우에 각기 다른 함수를 사용하여 일반화된 거리를 구한다는 사실이다.[32]

〈분산행렬이 다를 경우〉

$$D_i^2(X) = \log|S_i| + (X-\bar{X}_i)'S_i^{-1}(X-\bar{X}_i) - 2\log(\pi_i) \quad \cdots\cdots\cdots (3)$$

〈분산행렬이 모두 같을 경우〉

$$D_i^2(X) = (X-\bar{X}_i)'S_p^{-1}(X-\bar{X}) - 2\log(\pi_i) \quad \cdots\cdots\cdots\cdots\cdots (4)$$

여기서, S_i=집단 G_i의 공분산행렬

$$P_i(X) = \mathrm{Exp}(-.5D_i^2(X)) \Big/ \sum_{i=1}^{n} \mathrm{Exp}(-.5D_i^2(X)) \quad \cdots\cdots\cdots\cdots\cdots (5)$$

여기서, $P_i(X)$=객체 X가 집단 G_i에 속할 사후 확률($P_i(X)$가 $P_i, \cdots, P_k$ 중에서 최대값일 때 X를 집단 G_i에 분류)

Ⅲ. 판별분류시스템

본 장에서는 유기화학문헌을 실험 데이타로 하여 판별분류시스템을 구축

31) 김기영 & 선명식, SAS 판별 및 분류분석, pp.34~35.
32) Ibid.

한 후 이 분류시스템을 이용하여 실험·검증문헌집단을 각각 분류해 보고, 끝으로 시스템의 분류 정확도를 평가해 보았다.

1. 시스템設計

1) 判別過程

(1) 기초데이타와 분류체계

실험 데이타는 유기화학에 관한 269개의 문헌을 선택하여 218개는 실험문헌집단으로 하고 51개는 검증문헌집단으로 구성하였다. 분류카테고리는 CA에서 사용한 14개의 하위주제 카테고리中 비슷한 주제는 통합하고 일부 하위주제는 제외시켜 총 9개의 하위주제 카테고리를 설정하였다. 〈表 1〉에 하위주제코드표를 표시하였다.

〈表 1〉 하위주제코드표

코 드	하 위 주 제 명
1	지방족. 지환식 화합물
2	비축합·축합방향 화합물
3	복소환식 화합물(이원자 1개, 2개 이상)
4	유기금속 및 유기 metalloid 화합물
5	Terpenes and Terpenoids
6	Alkaloids
7	Steroids
8	탄수화물
9	아미노산, 펩타이드, 단백질의 합성

(2) 판별변수의 선정

CA에 이미 색인되어 있는 218개로 구성된 실험문헌집단의 표제, 키워드 구로부터 불용어 및 기능어를 제외한 단어를 추출하여 유기화학분야의 핵심단어 또는 단어 어근(word roots)을 공유하는 단어들을 동일한 단어그룹으

로 분류하였다.[33] 분석된 단어그룹中에서 문헌빈도가 3미만인 단어그룹들을 먼저 제외시키고 또다시 판별에 공헌도가 낮은 4종 이상의 하위주제들에 분포되어 있는 단어그룹들도 역시 제거해 최종적으로 47종의 단어그룹을 선정하여, 각 단어그룹이 공유하는 47개의 단어 또는 단어 어근을 판별변수로 간주하였다.

47개의 판별변수중 의미있는 판별변수들을 선정하기 위해서, 실험문헌집단(218개 문헌)의 각 문헌의 표제 및 키워드구에서 추출한 단어들에서 각 판별변수를 포함하는 단어의 수를 표시한 218×47 행렬을 입력물로 하였으며 단계적 판별분석법(변수증감법)으로 32개의 판별변수를 선정하였다. 각 단계에서 추가된 판별변수의 기여부분은 부분 F-통계량(partial-F)으로 평가하였는데 F-통계량(집단간 분산/집단내 분산)은 변수의 공헌도가 높을수록, 즉 변수가 집단내 분산을 작게 하면서 집단간 분산을 크게 할수록 상대적으로 더 커진다.[34] 〈表 2〉는 유의수준을 0.15로 하여 SAS로 처리한 단계적 판별분석(변수증감법)의 전과정을 요약한 것이다.[35]

〈表 2〉 전과정의 요약

단계	판별변수		F-통계량	$PROB>F$
	도입(판별변수명)	제거		
1	v40(steroid)	·	255.737	0.0001
2	v5(alkaloid)	·	189.127	0.0001
3	v31(lithi)	·	52.695	0.0001
4	v26(heterocycl)	·	34.556	0.0001
5	v33(peptide)	·	23.387	0.0001
6	v15(carbohydrate)	·	19.522	0.0001
7	v41(terpen)	·	13.937	0.0001
8	v18(cycloprop)	·	11.986	0.0001
9	v29(isoprenoid)	·	6.004	0.0001
10	v17(cyclocondens)	·	4.694	0.0001

33) lithi→lithiation, lithium, organolithium...
34) 김기영 & 전명식, SAS 판별 및 분류분석, pp.47~59.
35) 확률값이 주어진 유의수준 0.15 보다 작은 판별 변수들이 선정되었는데 그 이유는 확률값이 주어진 유의수준보다 작은 영역에서 귀무가설(판별변수가 집단 판별에 유의하지 않다)이 기각되고 대립가설이 채택되기 때문이다.

단 계	판 별 변 수		F-통계량	PROB>F
	도 입(판별변수명)	제거		
11	v20(diaryl)	·	4.481	0.0001
12	v9(amino acid)	·	4.359	0.0001
13	v10(analog)	·	5.121	0.0001
14	v36(prenyl)	·	4.469	0.0001
15	v7(alkene)	·	4.064	0.0002
16	v43(zeolite)	·	4.778	0.0001
17	v4(aliphatic)	·	3.678	0.0005
18	v32(natural product)	·	3.717	0.0005
19	v13(benzene)	·	3.763	0.0004
20	v3(aldehyde)	·	4.001	0.0002
21	v30(lactone)	·	4.390	0.0001
22	v35(potassium)	·	3.558	0.0007
23	v21(epox)	·	3.409	0.0011
24	v44(imidaz)	·	3.235	0.0018
25	v24(glyco)	·	3.663	0.0005
26	v25(halogenation)	·	3.020	0.0033
27	v12(aromat)	·	3.070	0.0029
28	v38(pyrrol)	·	3.317	0.0015
29	v8(alkoxy)	·	3.038	0.0031
30	v11(antibiotic)	·	2.694	0.0080
31	v46(nitro)	·	1.994	0.0496
32	v34(photochem)	·	1.972	0.0524

판별변수의 선정과정에서 제외된 15종의 판별변수를 분석해 보니 ab-initio, indol 등 4개의 판별변수는 판별에 공헌도가 상대적으로 더 높은 다른 판별변수와 언제나 동시에 같은 문헌에 출현하여 독립적인 판별 정보를 전혀 제공하지 못하는 변수였으며, acetoxy, benzene 등 11개의 변수는 3개의 서로 다른 하위주제코드를 갖는 문헌들에 고르게 분포되어 판별에 공헌도가 낮은 변수들이었다. 앞에서 선정한 32개의 변수에다 판별력이 낮아서 제외된 11개의 변수중 6개의 변수를 첨가하여 최종적으로 38개의 판별변수를 선정하였다.

2) 分類過程

〈表 3〉 50×39 행렬데이타

문헌번호 \ 단어번호	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	하위주제코드
1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
5	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
6	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
7	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
8	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1
9	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
11	0	0	0	0	1	0	0	0	0	0	0	3	0	1	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
14	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
15	0	0	0	0	0	0	0	0	0	0	0	0	1	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
16	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
27	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
28	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
29	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
30	0	0	0	0	0	0	0	0	0	0	2	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	2
32	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
33	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
34	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
35	3	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2
36	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
37	2	0	0	0	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
38	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	2
40	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2	0	0	0	0	0	0	2
41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2
42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2
43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2
44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	2
45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	2
46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2
47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2
48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	2
49	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3

검증·실험문헌집단의 문헌들을 각 집단별 공분산행렬의 동일성 여부에 따라 선형 또는 이차 판별함수를 이용할 수 있는 일반화된 거리함수와 사후확률에 의한 方法에 의해 분류해 보고자 SAS 통계 처리를 했다.[36]

(1) 입력데이타와 프로그램

SAS 프로그램의 입력 데이타로 두 개의 데이타화일을 사용하였다. 첫 데이타화일은 분류함수인 일반화된 거리함수를 유도하기 위한 실험문헌집단의 각 문헌에서 최종적으로 선정한 38개의 판별변수가 출현한 빈도와 하위주제코드를 표시한 218×39 행렬데이타이며, 두번째 데이타화일은 유도된 분류함수를 기초로 검증문헌집단의 각 문헌을 9개의 하위주제별로 분류하기

〈表 4〉 판별변수코드표

코 드	판 별 변 수 명	코 드	판 별 변 수 명
1	aldehyde	20	fungicide
2	aliphatic	21	glyco
3	alkaloid	22	halogenation
4	alkane	23	heterocycl
5	alkene	24	hormone
6	alkoxy	25	isoprenoid
7	amino acid	26	lactone
8	analog	27	lithi
9	antibiotic	28	natural product
10	aromat	29	peptide
11	benzene	30	photochem
12	carbohydrate	31	potassium
13	chromium	32	prenyl
14	cyclocondens	33	pyrrol
15	cycloprop	34	steroid
16	deoxy	35	terpen
17	diaryl	36	zeolite
18	epox	37	imidaz
19	ethylene	38	nitro

36) SPSS에서는 정준판별함수를 기초로 하여 객체를 분류하는데 반해 SAS는 일반화된 거리함수와 사후확률에 의한 방법에 의해서 객체를 분류한다.

〈表 5〉 51×38 행렬데이타

문서번호 / 단어번호	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
25	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
27	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0
33	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
45	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
46	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
47	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
49	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
51	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

위한 51×38 행렬데이타인데 이 행렬데이타는 검증문헌집단의 각 문헌의 표제에서 38개의 판별변수가 출현한 빈도를 표시한 것이다. 〈表 3〉은 218×39 행렬중 일부인 50개의 문헌에 대한 50×39 행렬이며 〈表 4〉는 〈表 3〉의 열에 표시된 판별변수코드에 대한 판별변수명이다. 그리고 〈表 5〉는 두번째 데이타화일인 51×38 행렬데이타를 나타내고 있다.

SAS 프로그램은 먼저 집단별 공분산행렬 동일성의 유의성 여부를 검정하여 그 결과에 따라 합동공분산행렬 또는 집단공분산행렬 중 하나를 선택하도록 작성하였고, 또한 유도된 분류함수의 정확성을 검토하기 위해서 분류함수의 도출에 사용된 실험문헌집단과 검증문헌집단의 문헌들을 각각 분류하도록 작성하였다.

(2) 분류결과

SAS를 통한 판별분석결과로 검증문헌집단의 각 문헌에 대한 원래 소속 하위주제와 각 하위주제별 사후확률이 출력되는데 그 중 10개의 문헌에 대한 분류 결과가 〈表 6〉에 표시되었고, 이때 각 문헌은 가장 큰 사후확률을 갖는 하위주제에 할당되며, 원래 하위주제와 할당된 하위주제가 다를 경우 그 문헌에 별표(*)를 추가하여 誤分類임을 지적해 주었다.

〈表 6〉 사 후 확 률

문헌	원래소속	분류된	사	후	확	률					
번호	집 단	집 단	1	2	3	4	5	6	7	8	9
1	1	1	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	2	2	0.00	0.78	0.22	0.00	0.00	0.00	0.00	0.00	0.00
3	2	4*	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
4	3	3	0.00	0.29	0.69	0.00	0.01	0.00	0.00	0.00	0.00
5	4	4	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
6	5	5	0.00	0.00	0.00	0.00	0.64	0.00	0.00	0.00	0.36
7	5	5	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
8	8	8	0.29	0.05	0.00	0.00	0.09	0.00	0.00	0.30	0.27
9	8	8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
10	9	9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

〈表 7-1〉 검증문헌집단의 분류행렬

from \ to	1	2	3	4	5	6	7	8	9	total
1	5 (71.43)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	2 (28.57)	7 (100.000)
2	0 (0.00)	2 (40.00)	0 (0.00)	1 (20.00)	0 (0.00)	0 (0.00)	0 (0.00)	1 (20.00)	1 (20.00)	5 (100.00)
3	0 (0.00)	0 (0.00)	5 (100.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	5 (100.00)
4	0 (0.00)	0 (0.00)	0 (0.00)	6 (100.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	6 (100.00)
5	1 (25.00)	0 (0.00)	0 (0.00)	0 (0.00)	3 (75.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	4 (100.00)
6	0 (0.00)	0 (0.00)	1 (12.50)	0 (0.00)	0 (0.00)	7 (87.50)	0 (0.00)	0 (0.00)	0 (0.00)	8 (100.00)
7	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	8 (100.00)	0 (0.00)	0 (0.00)	8 (100.00)
8	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	2 (100.00)	0 (0.00)	2 (100.00)
9	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	1 (16.67)	5 (83.33)	6 (100.00)
total	6 (11.76)	2 (3.92)	6 (11.76)	7 (13.73)	3 (5.88)	7 (13.73)	8 (15.69)	4 (7.84)	8 (15.69)	51 (100.00)

〈表 7-2〉 검증문헌집단의 하위주제별 분류 잡음률

하위주제 코드	1	2	3	4	5	6	7	8	9	total
잡 음 률	0.29	0.60	0.00	0.00	0.25	0.13	0.00	0.00	0.17	0.16

〈表 8-1〉 실험문헌집단의 분류행렬

from \ to	1	2	3	4	5	6	7	8	9	total
1	24 (96.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	1 (4.00)	0 (0.00)	25 (100.00)
2	0 (0.00)	21 (91.30)	0 (0.00)	1 (4.35)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	1 (4.35)	23 (100.00)
3	2 (8.70)	1 (4.35)	20 (86.96)	0 (.000)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	23 (100.00)
4	0 (0.00)	1 (3.85)	0 (0.00)	25 (96.15)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	26 (100.00)
5	0 (0.00)	2 (8.70)	0 (0.00)	1 (4.35)	18 (78.26)	1 (4.35)	0 (0.00)	1 (4.35)	0 (0.00)	23 (100.00)
6	0 (0.00)	1 (3.57)	0 (0.00)	0 (0.00)	0 (0.00)	26 (92.86)	0 (0.00)	0 (0.00)	1 (3.57)	28 (100.00)
7	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	1 (4.17)	0 (0.00)	22 (91.67)	0 (0.00)	1 (4.17)	24 (100.00)
8	0 (0.00)	2 (10.53)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	17 (89.47)	0 (0.00)	19 (100.00)
9	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	1 (3.70)	26 (96.30)	27 (100.00)
total	26 (11.93)	28 (12.84)	20 (9.17)	27 (12.39)	19 (8.72)	27 (12.39)	22 (10.09)	20 (9.17)	29 (13.30)	218 (100.00)

〈表 8-2〉 실험문헌집단의 하위주제별 분류 잡음률

하위주제 코드	1	2	3	4	5	6	7	8	9	total
잡 음 률	0.04	0.09	0.13	0.04	0.22	0.07	0.08	0.11	0.04	0.09

〈表 7-1〉과 〈表 8-1〉은 유도된 분류함수의 정확도를 검토하기 위해서 검증·실험문헌집단의 문헌을 각각 분류한 결과이고, 〈表 7-2〉와 〈表 8-2〉는 검증·실험문헌집단의 하위주제별 분류 잡음률을 나타내고 있다.

〈表 7-2〉에 의하면 하위주제 3,4,7,8(복소환식 화합물, 유기금속 및 유기 metalloid 화합물, steroids, 탄수화물)의 문헌들은 모두 정확하게 분류되었고, 하위주제 1(지방족·지환식 화합물)은 문헌 2개가 하위주제 9에, 하위주제 2(비축합·축합방향 화합물)는 문헌 3개가 각각 하위주제 4,8,9에, 하위주제 5(Terpenes and Terpenoids)는 문헌 1개가 하위주제 1에 잘못 분류되었고, 또한 하위주제 6(Alkaloids)은 문헌 1개가 하위주제 3에, 하위주제 9(아미노산·펩타이드·단백질의 합성)는 문헌 1개가 하위주제 8에 잘못 할당되어 전체 51개 문헌 중 8개, 즉 9개의 하위주제에서 평균 16%의 오분류를 나타내고 있다. 9개의 하위주제 중 하위주제 2는 가장 높은 60%의 오분류를 갖고 있는 것으로 분석되었는데 실험문헌집단의 크기를 확대하는 등의 방법을 사용하여 판별변수들을 다시 검토해 보는 것이 바람직할 것 같다. 〈表 7-2〉와 〈表 8-2〉의 분류 결과를 비교해 보면 실험문헌집단의 분류 정확률이 91%로 검증문헌집단의 분류 정확률 84% 보다 약 7%가 더 높았다.

2. 시스템評價

판별분류시스템의 분류 정확도를 평가해 보기 위해서, 앞에서 판별 분석한 결과와 군집분석을 이용해서 분류한 것을 비교해 보았다.

1) 群集分析

군집분석은 객체들사이의 相似性이나 또는 거리에 근거하여 군집을 형성해 가는 방법으로 판별분석에 비해 사전 정보를 이용할 수 없을 뿐더러, 의미없는 변수를 제거할 수 있는 메카니즘이 없기 때문에 선정된 변수는 모두가 동일한 비중으로 유사성 평가에 투입된다. 따라서 변수의 선정이 잘못되면 엉뚱한 결과가 나타날 수 있다. 군집의 방법은 다음과 같이 네가지로 나

눠 생각할 수 있다.[37]

(1) 系譜的 군집방법 : 한 군집이 다른 군집의 내부에 포함되나 군집간의 중복이 허용되지 않고 가계보 혹은 나무모양의 형식을 취하며 '가까운' 객체들끼리 묶어감으로써 군집을 만들어 가는 併合的方法과 전체대상을 하나의 군집으로 출발하여 세분해 나가는 分割的方法으로 나눌 수 있다.

(2) 最適分離 군집방법 : 사전에 결정된 군집의 수로 객체들을 분류하는 방법으로, 크게 다음 세가지 절차를 따르고 있다. 그 첫째는 군집의 초기값을 설정하는데 K개의 군집이 있으면 K개의 초기치가 있어야 하며, 둘째는 가장 가까운 초기값을 갖는 군집들에 관찰치들을 할당하고 그 임시적 군집이 형성된 후 초기값을 임시 군집의 평균으로 대체한다. 세째는 바뀐 군집 초기치에 의해 재할당을 한 후 초기치를 바뀐 평균치로 대체하고, 이와 같은 과정을 군집 초기치의 변화가 ϕ에 가깝게 되었을 때까지 반복한다.

(3) 重複 군집방법 : 군집의 형태중에는 군집들이 상호 배반적이지 않고 서로 겹치는 부분이 있어 한 객체가 두개 이상의 군집에 속함을 허용하는 것이 더욱 타당한 경우들이 있다. 예를 들어, 언어학에서 어떤 단어들을 여러 가지 의미들을 가지고 있고 몇 개의 어원 집단에 동시에 속할 수 있다.

(4) Fuzzy 군집방법 : 주어진 객체들을 몇 개의 집단으로 분리하는 군집방법에서는 각 객체가 오직 하나의 집단에만 소속된다. 그러나 이러한 方法은 지나치게 문제를 단순화 하는 수가 있다. 즉, 객체들 중에는 어떤 특정 집단에 소속된 것이 분명한 것들도 있고 불분명한 것들도 있을 수 있다. 따라서 이는 객체들에 대하여 所屬函數를 대응시켜 특정 집단에 소속하는 수준을 나타내는 형식으로 군집하는 방법이다.

군집분석으로 자료를 분석할 경우 군집의 갯수와 분석방법의 선택은 매우 어려운 문제가 된다. 군집의 갯수를 결정하는 문제는 일반적으로 要因分析에서 공통인자의 갯수를 결정하는 것보다 더 어려울 뿐 아니라, 어떤 군집분석도 만족할만한 해결책을 제공하지 못하고 있는 실정이다. 계보적 군집방법을 적용할 때에 군집수의 결정은 일반적으로 dendrogram의 검토를 통

37) 김기영 & 전명식, 군집분석, 서울 : 자유아카데미, 1990, pp.13~61.

해 이루어진다. 즉, 병합되는 과정에서 대응되는 값(혹은 거리의 척도) 이 상대적으로 큰 변화를 보일 경우 이를 세밀히 검토할 필요가 있다. 한편 어떤 판정기준의 최적화를 따르는 군집방법들에 있어서는, 군집의 갯수에 대응하는 판정기준의 값을 플롯(plot)하여 판정기준의 값에서 급격한 증가(혹은 최소화 기준을 사용할 때에는 급격한 감소)가 발생하는 곳에서 대응되는 군집의 갯수를 정하는 방법이 통상적으로 많이 사용되고 있다. 이 밖에도 일단 선택된 군집의 수를 가지고 시작하여 군집의 수와 구조를 동시에 결정하고자 하는 방법 등 여러가지 가설검정 방법들이 있다.[38]

연구에 적합한 군집방법을 선택하는 데 있어서 고려되어야 할 사항으로써 수학적 혹은 계산상의 문제는 물론, 사용하고자 하는 분석방법에 내재하는 기본적 가정들과 그 가정들이 주어진 자료에 대해 가지는 의미 및 부합성의 여부, 그리고 고려되는 변수들의 특성 등의 여러가지 측면에서 세밀히 검토해야 될 것이다.[39]

SAS 패키지 프로그램에서 사용 가능한 방법은 계보적 方法과 최적분리 方法이 있다. 일반적으로 계보적 군집방법은 자료자체가 어떤 계보를 지니고 있을 때 유용하며 최적분리 군집방법은 사전에 군집의 수가 정해져 있거나 대규모의 표본을 분류할 때 유리하다.[40][41]

2) 군집분석과 판별분석의 比較

군집분석과 판별분석의 분류 정확도를 비교하기 위해서 검증문헌집단을 최적분리 군집방법인 K-평균(여기서는 군집의 갯수가 9개이므로 9-평균) 군집방법을 이용해 분류해 보았다. 군집분석한 결과와 판별분석한 결과를 〈表 9〉에 표시하였다.[42]

38) Ibid. SAS 패키지 프로그램에서는 군집의 갯수에 관한 판정기준으로서 CCC(Cubic Clustrering Criterion)를 시험적용시키고 있다. 이는 우선 초사각형(hyper-rectangle) 상에서 균일분포를 따르고 있다고 여겨지는 점들이 만약 어떤 군집들을 이루고 있다면 이들은 대체로 초입방체 형태로 구분되어 있을 것으로 가정하고 있다. 이와 같은 가정하에서 유도된 CCC 판정기준은 2내지 N/10 정도의 군집수를 CCC의 값에 플롯했을 때, 국부적 최고점(local peak)이 있으면 이 점에 대응되는 군집의 수가 적절하다고 본다.

39) Ibid.

40) Ibid.

41) M.S. Aldenderfer & R.B. Blashfield, op. cit.

42) 김기영 & 전명식, 군집분석, pp.13—61.

〈表 9〉 분석결과 비교

분 석 방 법	정 확 률	잡 음 률
군 집 분 석	29 (57)	22 (43)
판 별 분 석	43 (84)	8 (16)

( )은 %임.

판별분석을 이용한 분류의 정확률이 84%인데 반해 군집분석한 결과의 정확률이 57%로써 가정한 대로 판별분석의 분류 정확도가 더 높았다.

Ⅳ. 結 論

본 논문은 2가지 연구가설을 검증하기 위해서 먼저, 판별분류시스템을 이용하여 유기화학문헌을 9개의 하위주제별로 분류해 본 결과 분류의 정확률이 대체로 높은 91%(실험문헌집단), 84%(검증문헌집단)로 나타났다. 따라서 가설 1은 검증되고, 두번째는 판별분석한 결과의 정확률은 84%로 군집분석의 결과(57%) 보다 27%가 더 많기 때문에 가설 2도 검증된 셈이다.

그러나, 판별분류시스템으로 문헌을 분류하는 과정에서 여러가지 문제점 및 제한점이 나타났는데 이를 열거하면,

1) 단어 또는 단어어근으로 구성된 38개의 판별변수가 원래 선정된 검증문헌집단(90개 문헌)의 주제를 커버하는 율이 약 57%(51/90)에 불과했다.

2) 검증문헌집단(51개 문헌)에서 하위주제 비축합·축합 방향 화합물에 관한 문헌들의 오분류율이 60%로 정확률보다 무려 20%가 더 높았다.

3) 동일한 단어 또는 단어어근을 포함하는 단어들이 모두 동일한 의미를 갖고 있다고 가정한 점이다.

4) 한국 유기화학 문헌을 분류할 때 영문 서명이나 초록이 없는 경우는 이 분류시스템으로 분류할 수 없다는 점 등이다.

따라서, 이러한 문제점 및 제한점을 최소화할 수 있는 方法으로는 판별분류시스템을 설계할 때 다음과 같을 사항들을 보완하는 것이 바람직 할 것

이다.

1) 판별변수를 선정할 때 실험문헌집단의 표제 및 키워드구는 물론 초록까지 분석하는 方法

2) 새로운 문헌을 분류할 때 문헌의 표제에서 판별변수가 하나도 출현하지 않는 경우는 분석 범위를 초록으로 확대하는 方法

3) 판별변수의 선정을 위해서 좀 더 포괄적인 유기화학분야의 공통 단어 어근리스트를 작성하는 方法

4) 판별변수를 선정하는 실험문헌집단의 크기를 가급을 최대로 하는 方法

위의 사항들을 고려하여 문제점들을 최소화한 새로운 판별분류시스템을 설계한다 하드래도 이 시스템이 모든 유기화학문헌을 분류해 주리라는 기대는 하지 않으며, 다만 판별이 확실한 문헌들을 분류해 주고, 판별변수가 출현하지 않은 문헌들 및 각 집단별 사후확률이 거의 비슷한 문헌들은 分類專門家의 지적 판단이 최종적으로 필요하겠다.

<참 고 문 헌>

1. 고려대학교 통계연구소. SAS를 이용한 통계 WORKSHOP. 서울 : 고려대학교 통계연구소, 1990.
2. 김기영 & 전명식. SAS 인자분석 서울 : 자유아카데미, 1990.
3. _____. SAS 군집분석. 서울 : 자유아카데미, 1990.
4. _____. SAS 판별 및 군집분석. 서울 : 자유아카데미, 1990.
5. 이재창 & 박정섭. "다차원축척(Multidimensional Scaling) 기법." 응용통계 1 : 61—80 ; 1986.
6. 정영미. 정보검색론. 서울 : 정음사, 1987.
7. Aldenderfer, M. S. & Blashfield, R. B. "Cluster Analysis," Sage Univ. Paper Series on Quantitative Applications in the Social Sciences. Beverly Hills and London: Sage Pubns, 1985.
8. Borko, H. & Bernick, M. "Automatic Document Classification." JACM 10 (1) : 151—162 ; 1962.
9. Dillon, M. & Federhart, P. "Statistical Recognition of Content Terms in General Text." JASIS 35(1) : 3—10 ; 1984.
10. _____. "The Use of Discriminant Analysis to Select Content Bearing

Words." JASIS 33(4) : 245—253 ; 1982.

11. Hamil, K. A. & Zamora, A. "The Use of Titles for Automatic Document Classification." JASIS 31(6) : 396—402 ; 1980.
12. Klecka, W. R. "Discriminant Analysis," Sage Univ. Paper Series on Quantitative Applications in the Social Sciences. Beverly Hills and London: Sage Pubns, 1980.
13. Maron, M. E. "Automatic Indexing: An Experimental Enquiry." JACM 8 (3) : 404—417 ; 1961.
14. Norusis, M. J. SPSS/PC+advanced statistics. Chicago: SPSS inc., 1986.
15. SAS Introductory Guide for Personal Computers. SAS Institute Inc.: Cary, NC, 1985.

(접수일자 '90. 5. 17)

A Study on Design of Discriminant Classification System for the Automatic Classification of Documents

Hyun-Hee Kim*
Yong-Rye Lee**

Abstracts

This study suggests two hypotheses and verifies them. First hypothesis is that discriminant analysis which is a statistical technique can be used to classify documents on the subject of organic chemistry by nine sub-areas. Second hypothesis is that discriminant analysis is superior to cluster analysis in classifing objects by fixed categories.

* Associate Professor, Myong Ji University
** Instructor, Myong Ji University