Journal of the Korean Data & Information Science Society
2010, **21**(3), 471–480

한국데이터정보과학회지

2009년 여자프로골프선수 프로파일을 이용한 군집방법비교

민대기[1]

[1]덕성여자대학교

접수 2010년 4월 9일, 수정 2010년 5월 17일, 게재확정 2010년 5월 23일

요 약

군집방법은 탐색적 통계기법에서 매우 유용한 방법이나 최종 의사결정을 지지할 검정 통계량이 없는 것이 단점이다. 자료구조에서 살펴보면 군의 성격을 파악하는 변수가 있느냐 없느냐가 군집분석과 판별분석의 차이이다. 군집분석이 가장 이상적으로 이루어졌다면 그 프로파일의 분석결과가 판별분석과 같을 것이다. 이 점에 근거하여 비계층 분석의 대표적인 K-평균법 방법과 자기 조직화지도 군집분석의 유효성을 2009년 여자프로골프 선수들의 프로파일 분석을 통하여 비교 연구하였다.

주요용어: 분류 비율, 엔트로피, 자기 조직화지도.

1. 서론

컴퓨터의 발달과 함께 최근에 군집분석은 EM (Expectation and Maximization) 알고리즘이나 인공신경망에 근거한 SOM (Self-Organization Maps) 등 다양한 연구가 진행되었다. 신경망을 이용한 예측모형에 관한 연구로는 Cho와 Park (2008)의 보험회사 이탈고객 분석에 관한 것이 있다. 자료구조상 판별분석과 비교하면 군의 소속을 판단할 변수가 없이 전체 자료가 몇 개의 군으로 어떻게 형성되었는지 연구하는 군집분석은 여러 분야에서 매우 유용한 방법론이나 최종적 의사 결정을 지지하는 통계량이 없는 것이 매우 큰 단점이다. 김재희와 고윤실 (2009)은 유사성이 큰 것들은 동일한 군집을 형성하며, 군집의 패턴이 인식되고 새로운 개체를 패턴인식을 통해 분류할 수 있다고 가정하였다. 이러한 가정은 군집분석이 이상적으로 이루어졌다면 그 결과를 판별 분석한 결과와 같을 수 있다 하겠다. 본 연구에서는 이러한 근거로 군집분석의 유효성을 의사결정나무 (decision tree)에서 활용되는 분류비율과 엔트로피를 근거로 군집분석의 유효성을 연구하였다.

2. 연구방법

2.1. 연구목적

군집분석은 데이터에 근거한 탐색적 방법으로, 자료가 몇 개의 군집으로 구성되어 있는가를 결정하는 매우 유용한 방법이다. Kim (2009)은 SPSS를 활용한 군집분석 방법론에 대하여 연구하였다. 본 연구에서는 군집의 수가 결정됐을 경우 비 계층적 방법으로 가장 많이 활용되는 K 평균법과 인공 신경망을 근거한 SOM을 2009년 LPGA 선수자료에 적용하여 각 방법에 따라 얼마나 군집이 이상적으로 이루어졌는가를 정보지수를 통하여 비교하였다. 즉 얼마나 군집의 개체가 의미 있게 분류되었는지를 실제 자료에 적용하여 결과를 비교하였다.

[1] (132-714) 서울특별시 도봉구 근화교길 19, 덕성여자대학교 통계학과, 부교수.
E-mail: dkmin@duksung.ac.kr

2.2. 자료설명 및 변수의 기술적 분석

자료는 2009년도 LPGA 상위권 선수들의 골프기술에 관한 자료 중 드라이브거리, 드라이브 정확도, 그린 적중률, 그린 적중 시 평균 퍼팅 수, 라운드당 버디 수, 이글 수, 샌드 세이브 등에 관한 것이다. 표 2.2에 의하면 121명의 선수에 대한 경기력 순위 평균이 이글을 제외한 분야에서 비슷한 값을 나타내고 있다. 이글은 파워와 퍼팅 실력이 우수해야 하므로 상위권 선수들에 순위 평균이 가장 작은 것으로 예측된다. PGA에서는 위에 열거한 항목의 순위를 합하여 All-Round라는 항목으로 순위를 매긴다. 정신력을 제외한 경기력에 관한 모든 것을 위 항목을 통하여 측정할 수 있다. 일반적으로 드라이브거리는 파워를 나타내는 지표가 되며, 그린 적중률은 아이언 샷의 정확도를, 그리고 샌드 세이브는 숏 게임 능력을 나타낸다. 표 2.3을 근거로 순위변수 간의 유의한 상관관계를 살펴보면 드리이브 거리 순위와 정확도 순위는 -0.35로 음의 상관관계가 있고, 즉 거리가 길면 정확도가 결여된다는 뜻이고 버디 순위와 이글 순위와는 각각 상관도가 0.61, 0.38로 아주 밀접한 관계를 보이고 있다. 드라이브거리 순위와 그린 적중률 순위는 0.52, 퍼팅 순위와 0.34로 높은 양의 상관계수를 보여주나 샌드 세이브는 관계가 없는 것으로 나타난다. LPGA에서 드라이브거리가 경기력에 미치는 영향력은 민대기와 현무성 (2009)의 PGA에 대한 연구에서도 알 수 있다. 골프기술 중 가장 관계가 깊은 것은 온 그린 시 퍼팅 순위와 버디 순위로 0.79를 나타내고 있다. 그만큼 버디는 정확한 퍼팅 기술과 정교한 아이언샷에 의해서 만들어지는 것을 말한다. 퍼팅 순위와 정확도 순위는 상관계수가 없는 것으로 나타났고 샌드 세이브와는 아주 미약한 관계를 보여주고 있어, 다른 기술과는 별도의 영역임을 알 수 있었다.

표 2.1 변수의 정의

변수	변수정의	변수역할
Distrank	드라이브거리 순위	예측변수
Girrank	드라이브 정확도 순위	예측변수
Accrank	페어웨이 적중률 순위	예측변수
Puttsrank	그린적중 시 평균 퍼팅수 순위	예측변수
Birdyrank	라운드당 버디 순위	예측변수
Eaglerank	이글순위	예측변수
Ssaverank	샌드 세이브 순위	예측변수
Scorerank	평균스코어 순위	예측변수

표 2.2 변수의 기술 통계량

변수	빈도수	평균	표준편차	최소값	최대값
scorerank	121	71.55	42.85	1.00	150
birdyrank	121	71.23	42.90	1.00	149
girrank	121	70.06	42.98	1.00	150
accrank	121	78.79	43.42	1.00	150
eaglerank	121	59.11	35.91	1.00	131
puttsrank	121	68.36	42.58	1.00	149
distrank	121	67.28	40.74	1.00	148

표 2.3 각 변수간의 상관관계

	birdyrank	girrank	accrank	eaglerank	puttsrank	distrank	ssaverank
birdyrank	1.00	0.67	0.06	0.27	0.79	0.61	0.29
girrank		1.00	0.31	0.27	0.32	0.52	0.13
accrank			1.00	-0.11	0.08	-0.35	0.07
eaglerank				1.00	0.23	0.38	0.07
puttsrank					1.00	0.34	0.30
distrank						1.00	0.08
ssaverank							1.00

3. 군집분석 결과

3.1. 군집 수의 탐색

군집분석에서 가장 중요한 것은 초기 군집 수를 몇 개로 결정할 것인가 하는 것이고 최종군집을 어떤 근거에 의하여 몇 개로 결정하는가이다. 골프 경기력은 앞서 언급한 바와 같이 숏 게임을 제외한 3가지 측정 요소로 분류하면 각 항목에 두 경우가 해당되기 때문에 최대 8개의 군집으로 분리할 수 있다. 가장 일반적인 방법으로 군집 수를 결정하기 위하여 주성분분석을 시도한바 6개의 고유치가 전체변이의 80%를 설명하여 초기 군집 수를 6으로 하고 군집분석을 시도하였다.

3.2. K-평균법 (K-means method)

Hartigang와 Wong (1979)은 K-평균법을 소개하였다. K-평균법은 비계층적 군집방법으로 가장 많이 활용되는 방법으로 하나 이상의 정량적 변수들로부터 계산되는 유클리드 거리에 기초하여 군집을 형성하며 군집 초기 값의 선택, 초기 군집의 형성, 개체들의 재 할당 단계를 거쳐 분석이 수행된다.

3.2.1. 최종 군집 수 결정 및 K-평균법 군집분석 결과

초기에 군집 수를 6으로 분석을 시도하고 그 결과에 대하여 나무 도표를 이용하여 프로파일 분석을 시도하였다. 그 결과 군집 수 6이나 4가 군집 수 5보다 이상적으로 분리되어 최종 군집 수를 5로 정하였다. K-평균법 분석결과 군집 1, 3, 5는 빈도수가 고르게 분포하였으나, 군집 2는 비교적 많게 군집 4는 작게 분포하였다. Martingnon (2005)는 작은 빈도수는 일반적으로 극한값을 포함하는 경우를 말한다고 하였다. 군집의 표준편차는 각 군집에서 중심으로부터 각 자료 값 간의 거리의 합을 의미하며 또한 군집의 상대적인 크기를 결정한다. 아래 표 3.1의 결과는 각 변수를 표준화하여 군집 분석을 시행한 결과이다. 표 3.1에 의하면 각 군집은 유사한 크기의 군집으로 형성되었으나, 군집 4의 빈도수가 다른 군집에 비하여 작은 것을 보면 군집을 형성하는 과정이 어려움을 알 수 있었다.

표 3.1 K-평균법 군집분석 결과 통계량

군 집	빈 도 수	군집 표준편차	씨앗부터 최대거리	근사군집	근사군집거리
1	20	0.71	2.89	2	2.24
2	40	0.74	2.71	1	2.24
3	25	0.76	2.97	5	1.91
4	15	0.70	2.93	2	2.29
5	21	0.76	2.89	3	1.91

그림 3.1의 평균프로파일 도표를 살펴보면 군집 1은 드라이브 거리에서 군집 2보다 약간 열세이지만 샌드 세이브를 제외한 나머지 부분에서 우수함을 보여주고 있다. 반면 군집 1은 정확도와 샌드 세이브

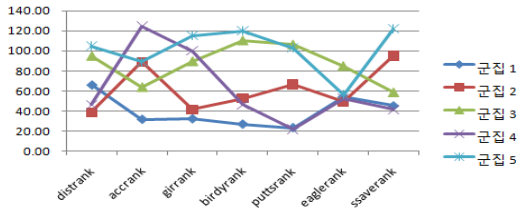

그림 3.1 평균프로파일도표

를 제외한 항목에서 우수함을 보여주고 있다. 군집 3과 5는 전반적으로 열세함을 나타내고 있고 군집 4는 버디. 퍼팅, 이글, 샌드 세이브 등에서 우수함을 보여주고 있다.

3.2.2. 나무 도표를 이용한 군집 프로파일 분석 및 유효성

본 연구에서는 정해진 군집 수에 대하여 어떤 방법이 좀 더 의미 있는 군집을 형성하는가를 비교해 보기 위함이므로 나무 도표를 이용하여 프로파일을 분석하였다. 예측된 군집을 목표변수로 하여 분류분석방법인 의사결정나무를 적용하였다. 각 군집에 대한 특성을 나무 도표를 이용하여 표시하면 아래 그림 3.2와 같다. 아래의 나무 도표는 어떤 변수가 군집을 형성하는 데 가장 영향력이 있는지를 보여 주고 있다. 또한, 나무 도표 모형을 평가하는 분리 기준인 순수도와 엔트로피를 이용하여 각 군집의 형성에 대한 적절성을 평가할 수 있다. 나무 도표를 표 3.2 테이블에 정리하였다. 테이블의 결과에 의하면 birdyrank가 전체 군집을 결정하는 가장 중요한 변수이다. 군집 1은 birdyrank가 91.5 미만이고 girrank가 89.5 미만이며 accrank가 56 미만인 그룹의 선수들로 구성되어 있다. 군집 2는 군집 1과 비슷하나 accrank 가 56 이상이며 ssaverank가 44 미만인 선수들로 구성되었다. 변수들의 성격상 birdyrank가 낮은 군집 1과 2가 우수한 선수들로 구성되어 있음을 예측할 수 있다. 군집 4는 birdyrank가 89.5 미만 그러나 girrank가 89.5 이상인 선수들로 구성되어 있다. 마지막으로 군집 3은 birdyrank가 91.5 이상인 선수 중 ssaverank가 84.5 미만인 선수로 구성되어 있고 군집 5는 ssaverank가 84.5 이상인 선수들로 구성되어 있다. 각 군집에 대한 scorerank 평균을 계산했을 때 군집 1에 속한 선수들의 scorerank 평균이 가장 낮았고 다음으로 군집 2, 4, 3, 5 순이다. 각 군집파일을 표 3.2에 정리하였다.

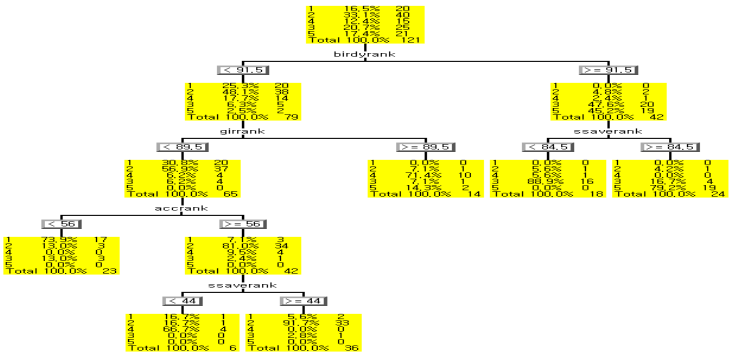

그림 3.2 나무 프로파일도표

각 군집에 속한 대표적 선수들을 아래의 표에 정리하였다. 몇 가지 아쉬운 것은 군집 1에 Ochoa와 신지애 (Jiyai Shin)와 같이 전혀 다른 경기 스타일인데 같은 군으로 분류됐다는 점이다. Ochoa 선수는 Cristie Kerr 선수와 마찬가지로 드라이버 거리가 길고, 페어웨이 정확도가 우수하지 않으나 그린적중률이 좋고 버디가 많은 선수이다. 반면 신지애는 드라이버 거리는 짧지만, 정확도가 우수하고 버디가 많은 선수이다. 또 다른 아쉬운 점은 Michelle Wie는 버디 능력이 뛰어나고 그린 적중률이 우수함에도 군집 4에 속했다는 점이다. 군집 4는 성적 순위가 중간그룹의 선수들로 구성되어 있고 그린 적중률이 우수하지 못하다. 그러나 Michelle는 표 3.7에 의하면 그린 적중률이 21위에 있다. 이러한 아주 다른 경기운영을 하는 선수가 같은 군집에 속해 있는 것과 전혀 다른 분류의 결과는 K-평균법에 대한 신뢰를 감소시킨다.

표 3.2 군집 1의 대표적 선수

군집 1								
name	scorerrank	birdyrank	girrank	accrank	eaglerank	puttsrank	distrank	ssaverank
Lorena Ochoa	1	1	7	54	31	1	9	88
Jiyai Shin	2	3	15	2	46	1	98	27
Cristie Kerr	3	10	2	61	20	17	12	82
Ai Miyazato	4	6	13	18	31	5	46	9

표 3.3 군집 2의 대표적 선수

군집 2								
name	scorerrank	birdyrank	girrank	accrank	eaglerank	puttsrank	distrank	ssaverank
Yani Tseng	5	7	11	119	4	51	3	63
Karrie Webb	18	30	9	80	12	76	23	26
Catriona Matthew	21	44	29	91	12	8	54	111

표 3.4 군집 3의 대표적 선수

군집 3								
name	scorerrank	birdyrank	girrank	accrank	eaglerank	puttsrank	distrank	ssaverank
Mi Hyun Kim	50	95	94	1	99	30	145	75
Candie Kung	70	118	58	66	46	111	95	14

표 3.5 군집 4의 대표적 선수

군집 4								
name	scorerrank	birdyrank	girrank	accrank	eaglerank	puttsrank	distrank	ssaverank
Michelle Wie	9	2	21	144	25	5	4	43

표 3.6 군집 5의 대표적 선수

군집 5								
name	scorerrank	birdyrank	girrank	accrank	eaglerank	puttsrank	distrank	ssaverank
Jane Park	84	92	107	96	31	66	105	85
Joo Mi Kim	99	113	105	63	79	66	100	112

3.3. 자기 조직화지도 (Self organization maps) 군집방법

SOM (Self Organization Maps)은 입력 벡터를 훈련 집합에서 대응되도록 가중치를 조정하는 인공 신경망 (neural network)에 기초한 자율학습의 한 방법이다. 신경망은 복잡하게 연결된 뉴런 (neuron)의 망으로 구성된 생체 학습 시스템을 수학적 모델을 빌려 만들어진 정보처리 시스템이다 (Na와 Kwon, 2010). SOM은 Helsinki 대학의 Teuvo Kohonen에 의해 최초로 소개하였으며 Kohonen map이라고도 불린다. 코호넨 네트워크는 입력층과 경쟁 층으로 구성되어 있으며 입력층을 통해서 들어온 data들은 입력층의 Unit에서부터 경쟁 층의 모든 구성 뉴런들과의 경쟁을 통하여 data 자신이 가장 유사한 뉴런 쪽으로 흡수되어 군집을 형성하는 방법이다. 코호넨 네트워크의 학습 철학은 승자 전취 (winner take all) 방식을 따른다. 그래서 승자만이 출력을 낼 수 있으며, 승자와 그의 이웃들이 그들의 연결강도를 조정한다. 승자 뉴런의 연결강도 벡터 (연결 가중치)는 입력 벡터와 가장 가까운 것이며 이 뉴런과 그의 이웃 반경 안의 뉴런들은 연결강도를 조정해 가면서 학습한다.

3.3.1. SOM 군집분석 결과

SOM 군집결과 군집 1, 2, 4는 빈도가 고르게 분포하였으나 군집 5는 비교적 많게 군집 3은 작게 분포하였다. 각 군집의 표준편차 비교결과 군집 1이 가장 작았고 군집 수가 작은 3이 가장 크게 나타났다. 각 군집의 중심으로부터 최대거리는 빈도수가 많은 군집 5가 가장 길었으며 군집 1이 가장 짧았다. 이러한 통계량을 근거로 군집 1이 가장 이상적으로 형성되었음을 알 수 있다.

표 3.7 SOM 결과 통계

군 집	빈 도 수	군집표준편차	중심으로 부터 최대거리	근사군집	근사군집거리
1	24	0.60	2.38	2	2.04
2	28	0.71	2.56	1	2.04
3	15	0.79	2.67	2	2.47
4	21	0.78	2.75	5	2.16
5	33	0.74	2.98	4	2.16

그림 3.3의 평균프로파일 도표를 살펴보면 군집 2는 드라이브 거리에서 군집 1보다 약간 열세이지만 나머지 부분에서 우수함을 보여주고 있다. 반면 군집 1은 정확도를 제외한 항목에서 우수함을 보여 주고 있다. 군집 4와 5는 전반적으로 열세함을 나타내고 있고 군집 3은 버디, 퍼팅, 이글, 샌드 세이브 등에서 우수함을 보여주고 있다.

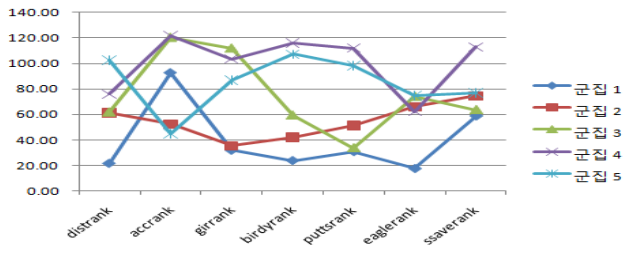

그림 3.3 평균프로파일 도표

그림 3.3의 평균프로파일 도표를 살펴보면 군집 2는 드라이브 거리에서 군집 1보다 약간 열세이지만 나머지 부분에서 우수함을 보여주고 있다. 반면 군집 1은 정확도를 제외한 항목에서 우수함을 보여 주

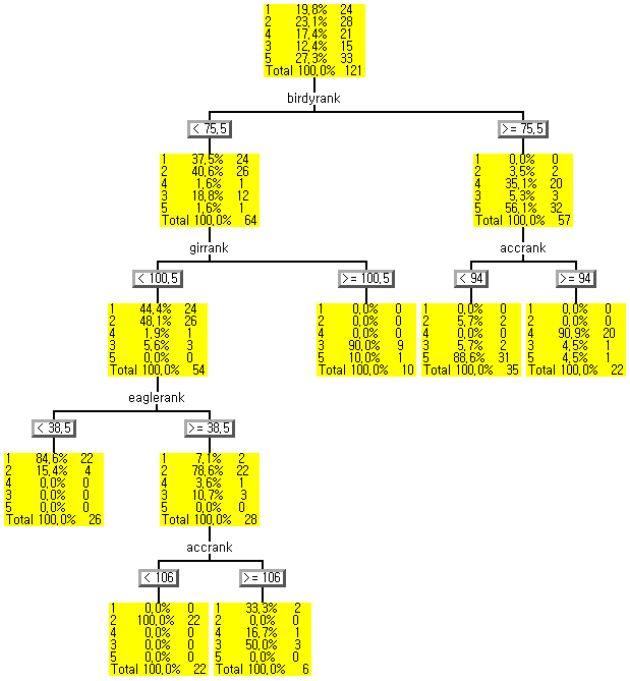

그림 3.4 SOM에 의한 군집 프로파일 나무 도표

고 있다. 군집 4와 5는 전반적으로 열세함을 나타내고 있고 군집 3은 버디, 퍼팅, 이글, 샌드 세이브 등에서 우수함을 보여주고 있다. birdyrank 가 75.5 이상인 선수 중 accrank 가 94 미만인 선수로 구성되어 있고 군집 4는 accrank 가 94 이상인 선수들로 구성되어 있다. 각 군집의 성격으로 성적을 관찰했을 때 군집 1과 2에 속한 선수들의 scorerank 평균이 가장 낮았고 다음으로 군집 3, 5, 4 순이다. 각 군집에 속한 대표적 선수들을 아래의 표에 정리하였다. 군집 1과 군집 2를 비교했을 때 K-평균법 군집방법과 다른 결과를 볼 수 있었다. 군집 1의 특성을 살펴보면 드라이브거리가 많이 나고 정확도는 떨어지나 그린 적중률이 좋고 버디를 많이 만들어내는 선수로 대표적인 선수는 Ochoca나 Cristie Kerrt 그리고 Michelle 등이 있다. 군집 2는 군집 1에 비하여 드라이브거리는 짧으나 정확도가 높고, 그린 적중률이 좋고, 숏 게임도 잘하며 버디를 많이 만들어 내는 선수다. 한마디로 요약하면 힘만 부족하고 나머지 분야에서 모두 우수한 기량을 가진 선수이다. 군집 3은 아주 이해하기 어려운 군집이라 할 수 있다. 일반적으로 버디 능력은 퍼팅과 그린 적중률이 높아야 한다. 그러나 군집 3은 버디를 잘하는 선수 중 그린 적중률이 낮은 선수로 분류가 형성되니 빈도수가 매우 낮았다.

3.4. K-평균법과 SOM의 비교

군집이 아주 이상적으로 이루어졌다면 그에 대한 분류도 잘 이루어졌을 것이라는 가정에 따라 군집분석의 유효성에 대한 판단 근거를 엔트로피와 분류비율로 하였다. 분류비율은 나무 도표의 초기 마디의 수를 최종 군집이 결정된 마디 군집의 빈도수로 나눈 비율로 정하였다. 최종 군집은 최종 마디에서 다수를 차지하는 군집의 빈도수에 의하여 결정하였고 다른 군집에 비하여 상대적으로 빈도수가 아주 작은

표 3.8 군집 1의 대표적 선수

군집 1								
name	scorerank	birdyrank	girrank	accrank	eaglerank	puttsrank	distrank	ssaverank
Lorena Ochoa	1	1	7	54	31	1	9	88
Cristie Kerr	3	10	2	61	20	17	12	82
Ai Miyazato	4	6	13	18	31	5	46	9
Michelle Wie	9	2	21	144	25	5	4	43

표 3.9 군집 2의 대표적 선수

군집 2								
name	scorerank	birdyrank	girrank	accrank	eaglerank	puttsrank	distrank	ssaverank
Jiyai Shin	2	3	15	2	46	1	98	27
Paula Creamer	10	23	1	3	46	21	86	49
Natalie Gulbis	19	15	47	26	79	8	80	27
Se Ri Pak	37	66	32	85	79	88	21	58

표 3.10 군집 3의 대표적 선수

군집 3								
name	scorerank	birdyrank	girrank	accrank	eaglerank	puttsrank	distrank	ssaverank
Beth Bader	46	24	92	114	79	17	71	108
M.J. Hur	55	47	113	139	31	8	53	24
Karine Icher	66	92	90	112	118	51	135	38
Inbee Park	67	50	138	142	131	1	28	17

1개의 마디는 계산에서 무시하였다. 유사한 논리로 초기 마디와 최종 군집 마디를 결정해서 군집의 유효성을 비교하기 위하여 초기 마디와 최종 군집결정 마디에서 엔트로피를 구한 후 그 차를 이용하여 정보 이득을 구했다. 분류비율이 높을수록, 정보 이득이 클수록 군집방법에 의한 이상적인 분류의 기준이 된다고 할 수 있다. 아래의 표에서 보듯이 SOM에 군집분석이 K-평균법에 결과보다 분류비율이나 정보 이득 면에서 높아 우수한 분석방법으로 나타났다.

표 3.11 군집방법 프로파일비교

	K- 평균		SOM	
	분류 비율	Information Gain	분류 비율	Information Gain
군 집 1	17/20=85%	2.032	22/24=92%	2.142
군 집 2	33/40=83%	2.093	22/28=79%	2.274
군 집 3	16/25=64%	2.147	9/15=60%	2.236
군 집 4	10/15=67%	2.114	20/21=96%	2.178
군 집 5	19/21=90%	2.062	31/33=94%	2.093
총	95/121=79%	10.448	104/121=86%	10.923

4. 결론

상황에 따라 주어진 자료에 최적의 군집개수와 이상적인 군집을 형성한다는 것은 쉬운 일이 아니다. 군집방법에 계층적, 비계층적 많은 연구가 수행되었으나 중요한 것은 최적 군집의 수와 형성된 군집이 의미 있어야 한다는 것이다. 본 연구에서는 가장 최적 군집은 최대한 이상적인 분류를 형성할 것이라는 가정에 따라 결정된 군집 수를 기반으로 LPGA 선수들의 프로파일을 K-방법과 SOM 방법으로 비교해

보았다. SOM을 이용한 군집 결과가 K-평균법을 이용한 결과보다 해석이 훨씬 의미가 있었고 분류비율이나 정보 이득에서도 좋은 결과를 보여 주었다.

참고문헌

김재희, 고윤실 (2009). 군집분석 비교 및 한우관능평가데이터 군집화. <응용통계연구>, **22**, 745-758.

Cho, M. and Park, E. (2008). Analyzing customer management data by data mining: Case study on churn prediction models for insurance company in Korea. *Journal of Korean Data & Information Science Society*, **19**, 1007-1018.

Hartigan, J. A. and Wong, M. A. (1979). K-means clustering algorithm. *Applied Statistics*, **28**, 100-108

Kim, D. (2009). A practical application of cluster analysis using SPSS. *Journal of Korean Data & Information Science Society*, **20**, 1207-1212.

Martignon, R. (2005). *Data mining using SAS enterprise miner*, Authorhouse, Cary.

Min, D. and Hyun, M. (2009). Prediction of a winner in PGA tournament using neural network. *Journal of Korean Data & Information Science Society*, **20**, 1119-1127.

Na, M. and Kwon, Y. (2010). Alternative optimization procedure for parameter design using neural network without SN. *Journal of Korean Data & Information Science Society*, **21**, 211-218.

Journal of the Korean Data & Information Science Society
2010, **21**(3), 471–480

한국데이터정보과학회지

A Comparison of cluster analysis based on profile of LPGA player profile in 2009

Dae Kee Min[1]

[1] Department of Information and Statistics, Duksung Women's University

Received 9 April 2010, revised 17 May 2010, accepted 23 May 2010

Abstract

Cluster analysis is one of the useful methods to find out number of groups and member's belongings. With the rapid development of computer application in statistics, variety of new methods in clustering analysis were studied such as EM algorism and Self organization maps. The goals of cluster analysis is finding the number of groupings that are meaningful to me. If data are analyzed perfectly with cluster analysis, we can get the same results from discernment analysis.

Keywords: Classification ratio, entropy, self-organization maps.

[1] Corresponding author: Associate professor, Department of Information and Statistics, Duksung Women's University, 419 Ssangmoon_dong Dobong_gu, Seoul 132-714, Korea.
E-mail: dkmin@duksung.ac.kr