머신러닝 알고리즘 Ch9

데이터 사이언스 & AI 전공

정 화민 교수

Clustering Analysis (계층적군집)

1. 군집분석 개요

- 특성에 따라 고객을 여러 개의 배타적인 집단으로 나누는 것
  - 결과는 구체적인 군집분석 방법에 따라 다름
  - 군집 개수, 구조에 대한 가정없이 데이터로부터 거리 기준에 의해 자발적인 군집화 유도
- 군집분석의 목적
  - 적절한 군집으로 나누는 것
  - 각 군집의 특성, 군집간의 차이 등에 대해 분석

2. 전통적 군집분석

기존 세분화 방법의 유형

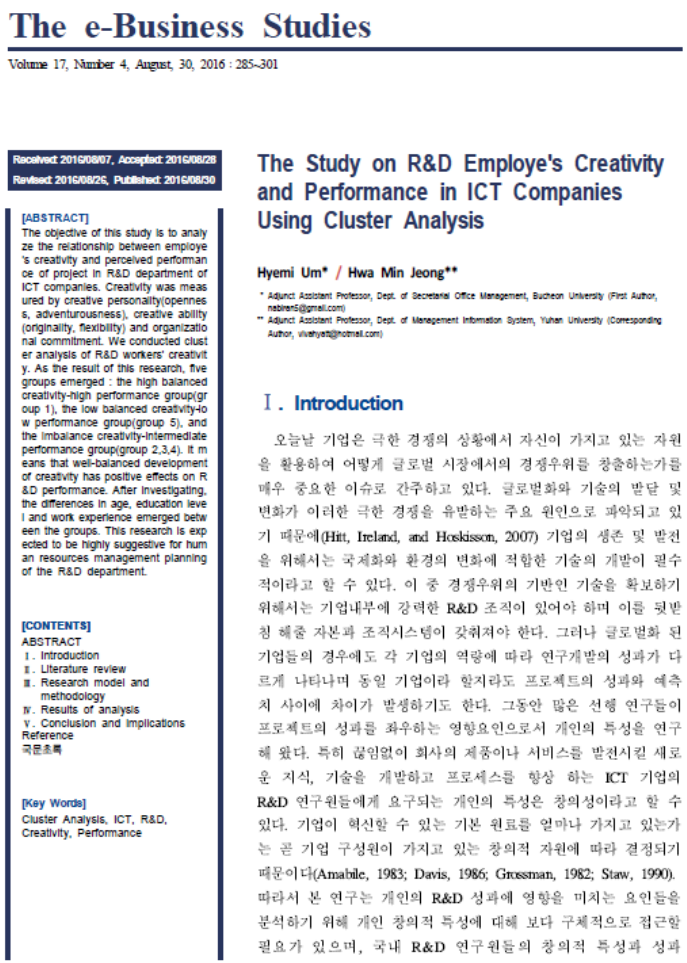
- 임의로 나누는 방법 : 고객등급/고객구분(신규/기존), 4분면, 9개 집단 등 다양
- 통계적 기법 : clustering, k-means 등.

3. 계층적 군집방법

- n개의 군집으로 시작해 점차 군집의 개수를 줄여나가는 방법
  - 관측 벡터 간의 거리뿐 만 아니라 군집간 거리에 대한 정의 필요

건강보험 공단 데이터를 이용한 건강 예측 플랫폼 사례

- AI 건강관리 서비스 플랫폼 데이터 분석에 사용된 분석기법 예



The e-Business Studies

Volume 17, Number 4, August, 30, 2016 : 285~301

Received: 2016/08/07, Accepted: 2016/08/28
Revised: 2016/08/26, Published: 2016/08/30

[ABSTRACT]

The objective of this study is to analyze the relationship between employe 's creativity and perceived performance of project in R&D department of ICT companies. Creativity was measured by creative personality(openness, adventurousness), creative ability (originality, flexibility) and organizational commitment. We conducted cluster analysis of R&D workers' creativity. As the result of this research, five groups emerged : the high balanced creativity-high performance group(group 1), the low balanced creativity-low performance group(group 5), and the imbalance creativity-intermediate performance group(group 2,3,4). It means that well-balanced development of creativity has positive effects on R&D performance. After investigating, the differences in age, education level and work experience emerged between the groups. This research is expected to be highly suggestive for human resources management planning of the R&D department.

[CONTENTS]

ABSTRACT
Ⅰ. Introduction
Ⅱ. Literature review
Ⅲ. Research model and methodology
Ⅳ. Results of analysis
Ⅴ. Conclusion and Implications
Reference
국문초록

[Key Words]

Cluster Analysis, ICT, R&D, Creativity, Performance

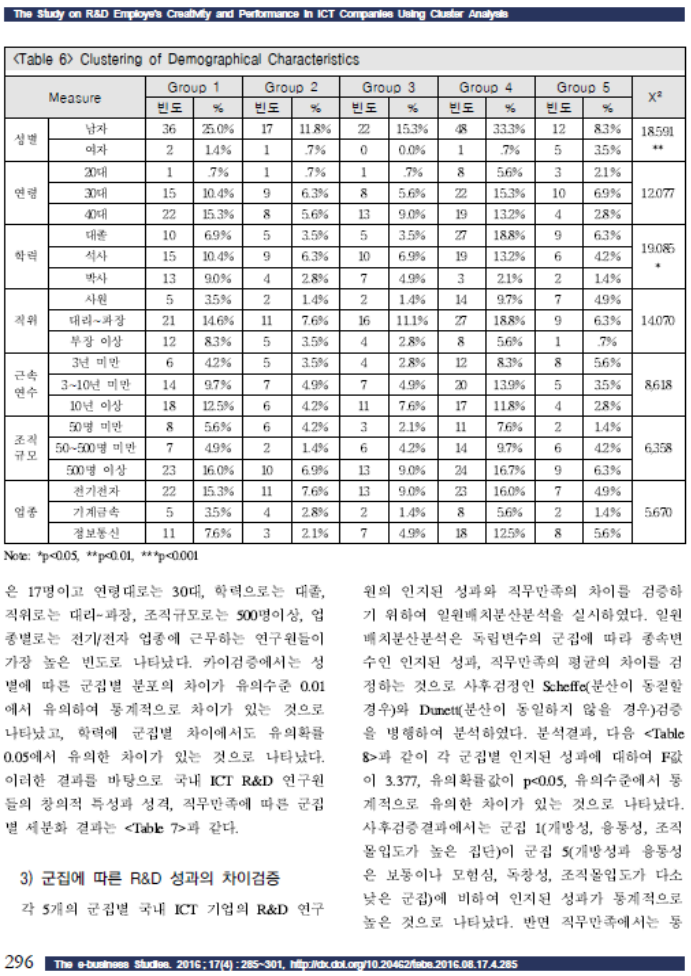
The Study on R&D Employe's Creativity and Performance in ICT Companies Using Cluster Analysis

Hyemi Um* / Hwa Min Jeong**

* Adjunct Assistant Professor, Dept. of Secretarial Office Management, Bucheon University (First Author, nabiren5@gmail.com)
** Adjunct Assistant Professor, Dept. of Management Information System, Yuhan University (Corresponding Author, vivahyat@hotmail.com)

Ⅰ. Introduction

오늘날 기업은 극한 경쟁의 상황에서 자신이 가지고 있는 자원을 활용하여 어떻게 글로벌 시장에서의 경쟁우위를 창출하는가를 매우 중요한 이슈로 간주하고 있다. 글로벌화와 기술의 발달 및 변화가 이러한 극한 경쟁을 유발하는 주요 원인으로 파악되고 있기 때문에(Hitt, Ireland, and Hoskisson, 2007) 기업의 생존 및 발전을 위해서는 국제화와 환경의 변화에 적합한 기술의 개발이 필수적이라고 할 수 있다. 이 중 경쟁우위의 기반인 기술을 확보하기 위해서는 기업내부에 강력한 R&D 조직이 있어야 하며 이를 뒷받침 해줄 자본과 조직시스템이 갖춰져야 한다. 그러나 글로벌화 된 기업들의 경우에도 각 기업의 역량에 따라 연구개발의 성과가 다르게 나타나며 동일 기업이라 할지라도 프로젝트의 성과와 예측치 사이에 차이가 발생하기도 한다. 그동안 많은 선행 연구들이 프로젝트의 성과를 좌우하는 영향요인으로서 개인의 특성을 연구해 왔다. 특히 끊임없이 회사의 제품이나 서비스를 발전시킬 새로운 지식, 기술을 개발하고 프로세스를 향상 하는 ICT 기업의 R&D 연구원들에게 요구되는 개인의 특성은 창의성이라고 할 수 있다. 기업이 혁신할 수 있는 기본 원료를 얼마나 가지고 있는가는 곧 기업 구성원이 가지고 있는 창의적 자원에 따라 결정되기 때문이다(Amabile, 1983; Davis, 1986; Grossman, 1982; Staw, 1990). 따라서 본 연구는 개인의 R&D 성과에 영향을 미치는 요인들을 분석하기 위해 개인 창의적 특성에 대해 보다 구체적으로 접근할 필요가 있으며, 국내 R&D 연구원들의 창의적 특성과 성과



〈Table 6〉 Clustering of Demographical Characteristics

Measure		Group 1		Group 2		Group 3		Group 4		Group 5		X^2
		빈도	%	빈도	%	빈도	%	빈도	%	빈도	%	
성별	남자	36	25.0%	17	11.8%	22	15.3%	48	33.3%	12	8.3%	18.591 **
	여자	2	1.4%	1	.7%	0	0.0%	1	.7%	5	3.5%	
연령	20대	1	.7%	1	.7%	1	.7%	8	5.6%	3	2.1%	12.077
	30대	15	10.4%	9	6.3%	8	5.6%	22	15.3%	10	6.9%	
	40대	22	15.3%	8	5.6%	13	9.0%	19	13.2%	4	2.8%	
학력	대졸	10	6.9%	5	3.5%	5	3.5%	27	18.8%	9	6.3%	19.085 *
	석사	15	10.4%	9	6.3%	10	6.9%	19	13.2%	6	4.2%	
	박사	13	9.0%	4	2.8%	7	4.9%	3	2.1%	2	1.4%	
직위	사원	5	3.5%	2	1.4%	2	1.4%	14	9.7%	7	4.9%	14.070
	대리~과장	21	14.6%	11	7.6%	16	11.1%	27	18.8%	9	6.3%	
	부장 이상	12	8.3%	5	3.5%	4	2.8%	8	5.6%	1	.7%	
근속연수	3년 미만	6	4.2%	5	3.5%	4	2.8%	12	8.3%	8	5.6%	8.618
	3~10년 미만	14	9.7%	7	4.9%	7	4.9%	20	13.9%	5	3.5%	
	10년 이상	18	12.5%	6	4.2%	11	7.6%	17	11.8%	4	2.8%	
조직규모	50명 미만	8	5.6%	6	4.2%	3	2.1%	11	7.6%	2	1.4%	6.358
	50~500명 미만	7	4.9%	2	1.4%	6	4.2%	14	9.7%	6	4.2%	
	500명 이상	23	16.0%	10	6.9%	13	9.0%	24	16.7%	9	6.3%	
업종	전기전자	22	15.3%	11	7.6%	13	9.0%	23	16.0%	7	4.9%	5.670
	기계금속	5	3.5%	4	2.8%	2	1.4%	8	5.6%	2	1.4%	
	정보통신	11	7.6%	3	2.1%	7	4.9%	18	12.5%	8	5.6%	

Note: *p<0.05, **p<0.01, ***p<0.001

은 17명이고 연령대로는 30대, 학력으로는 대졸, 직위로는 대리~과장, 조직규모로는 500명이상, 업종별로는 전기/전자 업종에 근무하는 연구원들이 가장 높은 빈도로 나타났다. 카이검증에서는 성별에 따른 군집별 분포의 차이가 유의수준 0.01에서 유의하여 통계적으로 차이가 있는 것으로 나타났고, 학력에 군집별 차이에서도 유의확률 0.05에서 유의한 차이가 있는 것으로 나타났다. 이러한 결과를 바탕으로 국내 ICT R&D 연구원들의 창의적 특성과 성격, 직무만족에 따른 군집별 세분화 결과는 <Table 7>과 같다.

3) 군집에 따른 R&D 성과의 차이검증

각 5개의 군집별 국내 ICT 기업의 R&D 연구원의 인지된 성과와 직무만족의 차이를 검증하기 위하여 일원배치분산분석을 실시하였다. 일원배치분산분석은 독립변수의 군집에 따라 종속변수인 인지된 성과, 직무만족의 평균의 차이를 검정하는 것으로 사후검정인 Scheffe(분산이 동질할 경우)와 Dunett(분산이 동일하지 않을 경우)검증을 병행하여 분석하였다. 분석결과, 다음 <Table 8>과 같이 각 군집별 인지된 성과에 대하여 F값이 3.377, 유의확률값이 p<0.05, 유의수준에서 통계적으로 유의한 차이가 있는 것으로 나타났다. 사후검증결과에서는 군집 1(개방성, 융통성, 조직몰입도가 높은 집단)이 군집 5(개방성과 융통성은 보통이나 모험심, 독창성, 조직몰입도가 다소 낮은 군집)에 비하여 인지된 성과가 통계적으로 높은 것으로 나타났다. 반면 직무만족에서는 통

주성분분석(PCA)

```
USArrests_PC.components_   # PC들을 column별 배치하려면 Transpose 해 주어야...
```

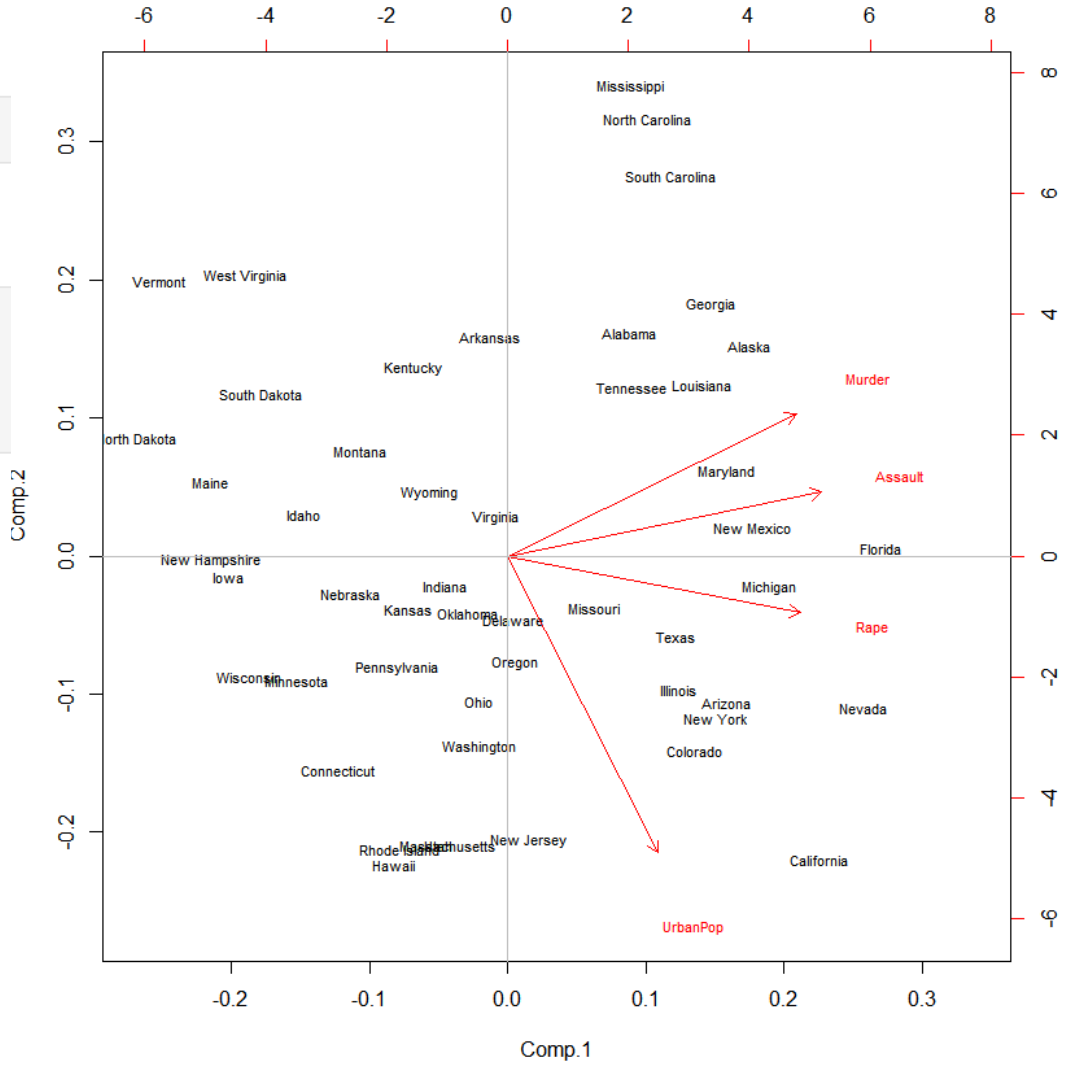
```
array([[ 0.53589947,  0.58318363,  0.27819087,  0.54343209],
       [ 0.41818087,  0.1879856 , -0.87280619, -0.16731864],
       [-0.34123273, -0.26814843, -0.37801579,  0.81777791],
       [ 0.6492278 , -0.74340748,  0.13387773,  0.08902432]])
```

```
PCs = pd.DataFrame(USArrests_PC.components_.T, index=USArrests.columns,
                   columns=['PC1', 'PC2','PC3','PC4'])

print(eigen_vecs)   # 'simple_PCA'로 직접 구한 PC loading 들
PCs                 # scikit-learn으로 구한 PC loading 들
```

```
[[ 0.53589947  0.41818087 -0.34123273  0.6492278 ]
 [ 0.58318363  0.1879856  -0.26814843 -0.74340748]
 [ 0.27819087 -0.87280619 -0.37801579  0.13387773]
 [ 0.54343209 -0.16731864  0.81777791  0.08902432]]
```

	PC1	PC2	PC3	PC4
Murder	0.535899	0.418181	-0.341233	0.649228
Assault	0.583184	0.187986	-0.268148	-0.743407
UrbanPop	0.278191	-0.872806	-0.378016	0.133878
Rape	0.543432	-0.167319	0.817778	0.089024



Source : https://github.com/hyunblee/ISLR-with-Python

Clustering Analysis (계층적군집)

Hierarchical Clustering

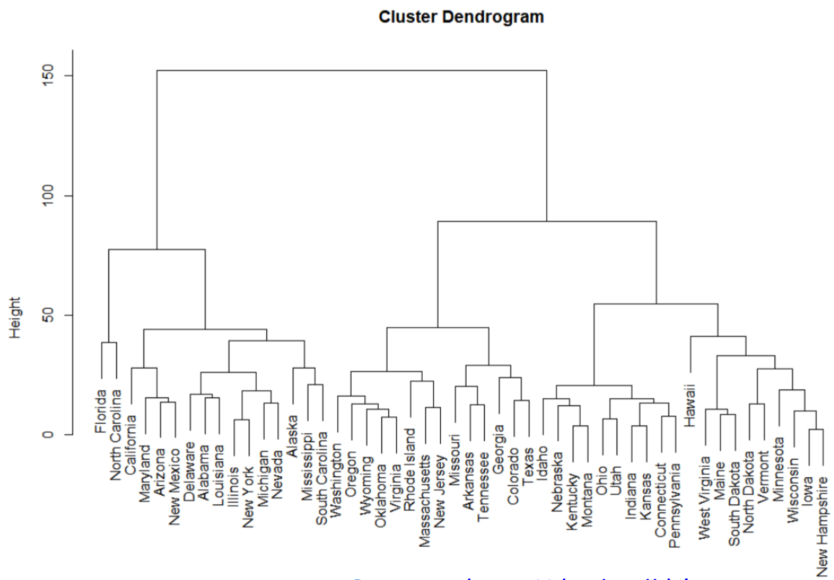
```
agg_cluster_complete=AgglomerativeClustering(distance_threshold=0, n_clusters=None,affinity='euclidean',linkage='complete').fit(X)
agg_cluster_single=AgglomerativeClustering(distance_threshold=0, n_clusters=None,affinity='euclidean',linkage='single').fit(X)
agg_cluster_ward=AgglomerativeClustering(distance_threshold=0, n_clusters=None,affinity='euclidean',linkage='ward').fit(X)
```

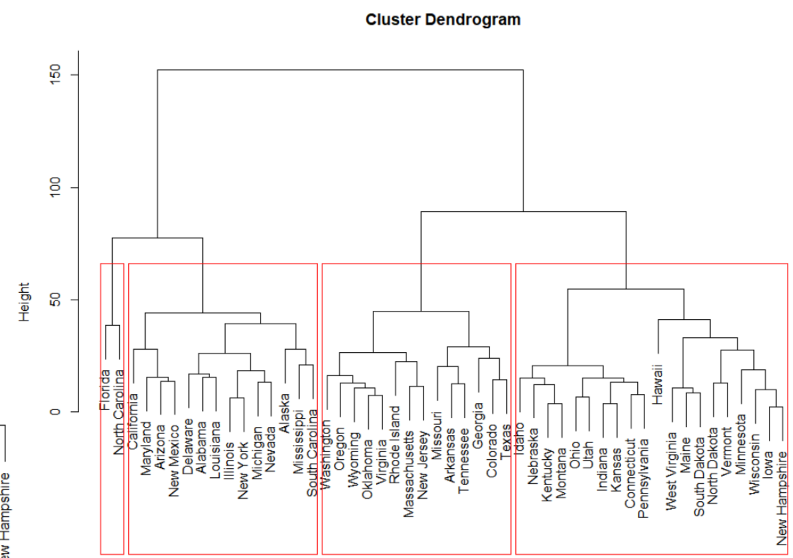
```
print(f'#clusters:{agg_cluster_complete.n_clusters_}')
print(f'#labels:{agg_cluster_complete.labels_}')
print(f'#leaves:{agg_cluster_complete.n_leaves_}')
```

```
#clusters:2
#labels:[0 0 0 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 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1]
#leaves:50
```

```
print(f'#clusters:{agg_cluster_single.n_clusters_}')
print(f'#labels:{agg_cluster_single.labels_}')
print(f'#leaves:{agg_cluster_single.n_leaves_}')
```

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





Source : http://denizstij.blogspot.com/2021/02/chapter-10-unsupervised-learning.html

Clustering Analysis (비계층적 군집 K-means clustering)

- K-평균법(K-means method)가 대표적

  ▪ 원하는 군집개수, 초기값 정해 seed 중심으로 군집 형성→각 데이터를 거리가 가장 가까운 seed가 있는 군집으로 분류→각 군집의 seed값 다시 계산→모든 개체가 군집으로 할당될 때까지 반복

  ▪ K-평균법은 한 개체가 속한 군집에서 다른 군집으로 이동해 재배치가 가능. 초기값에 의존. 군집의 초기값 선택이 최종 군집 선택에 영향 미침. 몇 가지 초기값 선택 후 결과 비교하는게 유용

가. 비계층적 군집화의 장점

- 주어진 데이터의 내부구조에 대한 사전정보없이 의미 있는 자료구조 찾을 수 있다.
- 다양한 형태의 데이터에 적용 가능
- 분석방법의 적용이 용이

나. 비계층적 군집화의 단점

- 가중치와 거리정의가 어려움
- 초기 군집 수를 결정하기 어려움
- 사전에 주어진 목적이 없으므로 결과 해석이 어려움