Proceedings of KFIS 2002 Spring Conference, 2002. 5. 25

인자 점수를 이용한 이상치 데이터의 군집화

Outlier Data Clustering using Factor Score

전성해, 임민택, 오경환
서강대학교 컴퓨터과학과

Sung-Hae Jun, Min-Taik Lim and Kyung-Whan Oh
Department of Computer Science, Sogang University
E-mail : shjun@ailab.sogang.ac.kr

ABSTRACT

이상치를 포함한 학습 데이터의 군집화 전략은 일반적으로 이상치를 포함하여 학습하거나, 이상치를 제거하는 두 가지 선택이 가능하다. 이상치를 제거하지 않고 학습에 반영시켜야 할 경우 한 개 또는 소수의 이상치가 독자적인 군집을 형성하거나 객관적인 군집화를 방해하는 문제가 발생할 수 있다. 이 때 주어진 학습 데이터의 군집 결과가 이상치의 영향으로부터 벗어나기 위해 원래의 학습 데이터에 대한 변환 작업을 거친 후 군집화를 수행할 수 있다. 이러한 변환 방법으로서 본 논문에서는 차원 축소의 기법으로 알려진 인자 분석의 점수를 사용하였다. 인자 점수로 변환된 학습 데이터에 대해 계층적 군집화, K-means 그리고 자기조직화 지도 등과 같은 군집화 알고리즘을 적용하면 이상치가 자신만의 군집을 별도로 형성하지 않고 다른 학습 데이터의 군집에 소속되면서 이상치의 영향으로부터 벗어남을 실험을 통하여 확인하였다.

Key words : 인자점수, 군집화, 이상치, 인자분석

1. 서 론

일반적인 군집분석은 원래 데이터에 있는 각 변수의 관측 값을 이용하여 관측 개체의 거리 행렬, 혹은 유사성 행렬을 만들어 이를 이용하여 모든 개체들을 집단화 한다. 이때 거리(비유사성), 혹은 유사성의 측도로 무엇을 쓰느냐, 또는 어떤 근거로 집단화 하느냐에 따라 여러 방법을 생각할 수 있다. 이들 방법 중 특별히 어느 군집 분석 방법이 가장 좋다고는 말할 수 없다는 것은 잘 알려져 있다[3]. 그래서 여러 군집 방법 중에서 분석가가 원 데이터의 형태를 가지고 자신의 주관적인 결정에 의해 방법을 선택하게 된다. 하지만 이상치가 있을 때는 방법의 선택에 신중을 기해야 한다. 본 논문의 목적은 이러한 이상치를 제거하지 않고 분석에 포함시켜야 할 경우에 이에 대한 해결 방법에 대한 연구이다. 이러한 연구로 본 논문에서는 원 데이터의 군집분석 시 이상치의 영향에서 벗어나게 하기 위해 변수들의 인자분석(factor analysis)[1]을 통한 인자점수를 택하고 이것을 이용하여 최종 군집분석을 수행한다. 본 논문의 2절에서는 군집분석 시의 이상치를 어떠한 기준에 근거하여 판정하는가에 대해 알아보고, 3절에서는 이상치의 영향에서 벗어나기 위한 방법으로서 사용되는 인자점수에 대한 설명과 인자점수를 구하기 위한 분석기법을 알아본다. 군집분석시의 이상치의 구체적인 처리 과정에 대한 실험 및 결과는 4절에서 알아보고 마지막으로 5절에서는 결론 및 향후 연구과제에 대해서 언급한다.

2. 군집 분석시의 이상치

군집분석을 할 데이터에 이상치가 없으면 일반적으로 상사성에 의한 군집 방법을 사용하면 된다. 그런데 이상치가 있을 경우에는 이 이상치를 제거하고 일반적인 군집분석을 수행할 수 있다. 그러나 이상치를 제거할 수 없는 경우에

는 이를 포함하여 군집분석을 수행해야 하며 이 때 한 개 또는 소수의 이상치 개체가 한 개의 군집을 형성하는 문제가 발생할 수 있다. 이러한 문제를 해결하기 위하여 본 논문의 제안 방법을 적용시킬 수 있다. 군집분석이라는 방법 자체가 개체간의 상사성이나 거리에 의해 몇 개의 집단으로 나누는 것이므로 이상치의 판정 기준으로 사용되는 통계량 역시 이것에 기초를 두어야 한다. 본 논문에서의 이상치 판정은 Mahalanobis의 거리[4]를 이용한다. Mahalanobis 거리, d_i 는 다음과 같이 정의된다.

$$d_i = (x_i - \bar{x})' S^{-1} (x_i - \bar{x}) \quad (1)$$

다변량 Slutsky의 정리[2]에 의하여 비록 d_i 는 서로 독립이 아닐지라도 근사적으로 변수의 개수를 자유도로 하는 χ^2 -분포를 따르게 된다. 본 논문에서는 이 통계량을 이용하여 이상치를 판정한다.

3. 군집 분석시 이상치의 변환

3.1 인자 분석의 모형과 방법

이상치의 영향에서 벗어나기 위한 선형 변환이라고 생각할 수 있는 것에는 다변량 통계 분석 방법의 하나인 인자 분석이 있다. 이는 원 변수 사이의 상관 관계를 찾아내어 일차 결합으로 새롭게 생성되는 변수를 이용하여 서로 독립으로 만들고 이렇게 만든 작은 수의 새 변수로 원 변수의 총 변동을 충분히 설명할 수 있는 인자라고 부르는 새로운 변수의 계수를 이끌어 낸다. 인자 분석은 유사성(상관, 공분산)행렬의 구조에 관한 통계적 모형을 구축하고 생성시키는 몇 개의 인자를 유도하여 해석하는 기법이다. 따라서 변수들 중의 일부가 공통 인자에 대해 그 중요성을 나타내는 부하량(loading)이 클 때 이 인자는 새로운 의미를 갖는 변수로 명명된다. 이것은 변수들 간의 복잡한 구조를 단순화 시켜주는 것이라 할 수 있다. 본 논문의 목적인 인자점수에 의한 군집분석은 변수에 대해 상관이라는 유사성 척도를 통해 새롭게 명명되는 인자로 축소된 차원의 데이터를 가지고 개체들에 대해 군집분석을 실시한다고 할 수 있다. p개의 변수와 m개의 인자 사이의 인과관계는 다음과 같은 인자패턴(factor pattern)으로 표현된다.

$$X_{(p\times 1)} - \mu_{(p\times 1)} = \Lambda_{(p\times m)} F_{(m\times 1)} + \varepsilon_{(p\times 1)}, (m < p) \quad (2)$$

여기서 $\Lambda = \{\lambda_{ij}\}$ 는 인자부하(factor loading) 행렬 또는 인자패턴 행렬이고 λ_{ij} 는 개 개의 인자부하이다. 그리고 $F = \{f_1, f_2, ..., f_m\}'$ 는 공통인자(common factor) 벡터이고 f_i 는 공통인자이다. 또한 $\varepsilon = \{\varepsilon_1, \varepsilon_2, ..., \varepsilon_p\}'$ 는 특수인자(유일성 변수) 벡터이고 ε_j 는 특수 인자(special factor)이다. 위에서 인자 부하량 λ_{ij} 는 모형에서 고려된 i번째 변수 x_i 에 관한 j번째 공통인자 f_j 의 중요성을 의미하고 f_j 와 ε_j 등은 관찰될 수 없는 것이고, 특히 ε_j 는 j번째 변수 x_j 에만 연관되어 있는 특수 요인이다. 이 모형에 의해서 p개의 관찰 변수들이 m개의 인자에 의해 차원이 축약되었다고 한다. 여기서 m개의 인자들은 서로 상관되어 있지 않으면서 p개의 관찰 변수에 내재된 잠재변수로서의 역할을 한다. 인자부하 행렬과 공통인자 벡터를 구하기 위한 인자분석 방법에는 크게 주축 인자법(principal factor method)에 의한 방법과 최대가능도법(maximum likelihood method)에 의한 방법이 있다. 두 개의 방법은 각각의 장단점이 있으나 본 논문에서는 통계적 추론이 필요하지 않은 군집분석에 적용되기 때문에 인자적재나 인자회전(factor rotation) 등 제반 모든 분석을 주축 인자법으로 수행한다. 주축 인자법에 의한 인자부하 행렬을 이용하면 인자분석의 통계적 모형하에서 원 데이터의 공분산(covariance) 행렬 Σ 를 다음과 같이 인자 분해(factorial decomposition)할 수 있다.

$$\Sigma = \Lambda\Lambda' + \Psi \quad (3)$$

만일 모든 변수들의 분산이 1이 되도록 표준화 시킨다면 아래 식과 같이 표현 될 수 있다.

$$P = \Lambda\Lambda' + \Psi \quad (4)$$

단 $P = \{\rho_{ij}\}$ 는 상관(correlation) 행렬이다. 즉 공분산 행렬과 상관 행렬이 같게 되는 것이다. 일반적으로 군집분석을 할 때 데이터는 각 변수별로 그 척도가 틀리는 경우가 대부분일 것이다. 이 때는 데이터의 변수들을 표준화 시켜 주어야만 올바른 분석이 가능하다. 이 표준화 작업에 상관 행렬을 사용한다.

3.2 인자의 개수에 대한 결정

군집분석의 입력 변수로서 사용할 인자점수를 구하기 위해서는 인자의 개수를 사전에 주관적으로 결정해야 한다. 이 때 생각할 수 있는 방법과 기준은 다음과 같다.

- 고유값(eigen value)의 크기에 의한 방법 : 통상적인 상관행렬로부터 인자를 유도할 경우 상관 행렬의 고유값이 1보다 큰 것의 개수를 인자의 개수로 한다.
- 인자의 공헌도에 의한 방법 : 한 인자의 실제적으로 중요한 의미를 가지기 위해서는 그 인자가 전체 변이에 대해 가지는 공헌도가 최소한 얼마(예를 들어 70%) 이상이 되어야 한다는 것이다. 따라서 이 얼마 이상이 되는 설명력을 갖는 인자의 수로 최종 인자의 개수를 결정한다.

이외에도 가능도비(likelihood ratio)원리에 의한 방법, 카이 제곱(chi-square) 적합도 검정에 의한 방법 그리고 Scree 도형에 의한 방법이 있다.

3.3. 인자 점수(factor score)

인자점수는 본 논문에서 수행되는 군집 분석의 입력 변수로 사용된다. p개의 변수에 대해서 크기 N개의 개체를 관찰하여 상관 행렬을 구해서 이것을 기초로 인자분석을 수행하여 원 변수에 대한 m개의 각 인자의 인자 부하량을 구하고 이들 인자 부하량을 계수로 하는 일차 변환으로 다시 개체들의 인자점수를 구하게 된다. 일반적으로 개체가 N개이고 변수가 p개이며, 인자가 m개인 경우 변수와 인자와의 관계는 다음과 같다.

$$X_1 = \lambda_{11} f_1 + \cdots + \lambda_{1j} f_j + \cdots + \lambda_{1m} f_m = \sum_{j=1}^{m} \lambda_{1j} f_j$$

$$X_i = \lambda_{i1} f_1 + \cdots + \lambda_{ij} f_j + \cdots + \lambda_{im} f_m = \sum_{j=1}^{m} \lambda_{ij} f_j$$

$$X_p = \lambda_{p1} f_1 + \cdots + \lambda_{pj} f_j + \cdots + \lambda_{pm} f_m = \sum_{j=1}^{m} \lambda_{pj} f_j$$

(5)

인자 부하량 λ_{ij}는 변수 X_i가 공통인자 f_j에 대해서 가지는 가중치 값이다. 이 인자 부하량을 이용해서 각 개체 $O_1,\ldots,O_N$에 대한 인자점수를 구한다. o_{kl}이 k번째 개체의 l번째 변수에 대한 관찰값을 나타낸다고 하면 O_k는 $O_k = (o_{k1},\cdots,o_{kp})$와 같이 표현할 수 있다. g번째 개체가 h번째 인자에 대해서 가지는 인자점수 s_{gh}는 다음과 같이 표현된다.

$$s_{11} = \lambda_{11} o_{11} + \cdots + \lambda_{l1} o_{1l} + \cdots + \lambda_{p1} o_{1p} = \sum_{l=1}^{p} \lambda_{l1} o_{1l}$$

$$s_{1m} = \lambda_{1m} o_{11} + \cdots + \lambda_{lm} o_{1l} + \cdots + \lambda_{pm} o_{1p} = \sum_{l=1}^{p} \lambda_{lm} o_{1l}$$

$$s_{gh} = \lambda_{1h} o_{g1} + \cdots + \lambda_{lh} o_{gl} + \cdots + \lambda_{ph} o_{gp} = \sum_{l=1}^{p} \lambda_{lh} o_{gl}$$

$$s_{N1} = \lambda_{11} o_{N1} + \cdots + \lambda_{l1} o_{Nl} + \cdots + \lambda_{p1} o_{Np} = \sum_{l=1}^{p} \lambda_{l1} o_{Nl}$$

$$s_{Nm} = \lambda_{1m} o_{N1} + \cdots + \lambda_{lm} o_{Nl} + \cdots + \lambda_{pm} o_{Np} = \sum_{l=1}^{p} \lambda_{lm} o_{Nl}$$

(6)

즉 g번째 개체 O_g는 s_{g1}에서 s_{gm}까지의 m개의 인자점수를 갖게 된다.

4. 실험 및 결과

본 논문의 실험을 위하여 미국의 50개 주(state)에서 7종류의 범죄에 대해 10만 명당 범죄율을 조사한 데이터를 이용하였다. 입력 변수는 살인(X1), 강간(X2), 약탈(X3), 폭행(X4), 가택침입(X5), 절도(X6) 그리고 자동차 절도(X7)이다[7]. 본 논문의 제안 방법을 적용하기 위하여 원 데이터에서 Massachusetts 주의 자동차절도 변수 값을 1140.1에서 2140.1로 바꾸어 이상치 판단기준을 만족시키는 이상치를 발생시켰다. 7개의 입력 변수(X1-X7)에 대한 상관계수 행렬을 구하여 인자분석을 수행한 결과가 표 1과 같이 나타났다. 3.2절의 인자개수 결정 방안으로부터 고유치가 1보다 큰 인자의 개수인 2개를 결정하였다. 2개의 인자는 전체 모형의 약 71%정도를 설명하는 것으로 나타났다.

표 1. 인자개수 결정을 위한 고유치와 설명력

	고유치	설명력(%)
F1	3.7571	53.67
F2	1.2075	70.92
F3	0.8612	83.23
F4	0.5106	90.52
F5	0.2589	94.22
F6	0.2281	97.48

F7	0.1765	100

2개의 각 인자로 7개의 원래 변수의 선형 결합을 나타낼 수 있는 인자 패턴을 구한 결과는 표 2에서 보여준다. 이 인자 패턴을 통하여 각 개체에 대한 인자점수를 구할 수 있다.

표 2. 인자 패턴(Factor pattern)

	Factor1	Factor1
X1	0.5817	-0.6972
X2	0.8952	0.0406
X3	0.7728	-0.0614
X4	0.8173	-0.3005
X5	0.8781	0.1012
X6	0.6876	0.5824
X7	0.3291	0.5256

표 3은 원 데이터에 대하여 인자분석을 실시하지 않고 7개의 원래변수(X1-X7)를 사용하여 K-mean 방법[6]에 의해 군집 분석을 수행한 결과이다. 군집 3의 개체수와 state를 보면 이상치로 판정된 Massachusetts 주가 한 개의 군집을 형성하고 있는 것을 볼 수 있다.

표 3. 원 변수에 의한 군집화 결과

군집	개체수	State
1	13	Arizona,…,Washington
2	26	Alaska,…,Wyoming
3	1	Massachusetts
4	10	Alabama,…,West_Virginia

다음은 7개의 원래 변수들에 대해 인자분석을 실시한 결과로서 얻어지는 인자점수를 이용한 군집분석 결과가 표 4에 나타나 있다. 실험 결과에서 한 개의 이상치 개체가 한 개의 군집을 형성하지는 않는 것으로 나타났다.

표 4. 인자 점수에 의한 군집화 결과

군집	개체수	State
1	8	Alabama,…,Tennessee
2	25	Arkansas,…,Wyoming
3	4	Delaware,Hawaii,. Massachusetts,Rhode_island
4	13	Alaska,…, Washington

본 실험은 인자의 개수를 두 개로 결정하고 이에 따른 인자점수를 이용하였지만 인자의 개수를 바꾸어 군집분석을 수행할 수 있다. 또한 군집화의 결과도 다르게 나올 수 있다. 결론적으로 이상치가 있는 데이터는 원래의 데이터를 그대로 분석하는 것보다는 원 변수들에 대한 인자분석에 의한 인자점수를 사용하는 것이 이상치에 의한 영향을 덜 받아서 군집화 결과가 더 만족스럽게 된다.

5. 결론 및 향후 연구과제

본 논문에서는 이상치를 포함하는 데이터에 대한 군집화 전략을 제안하였다. 제안된 방법에 의해 이상치의 영향에서 벗어나는 군집 결과를 얻을 수 있었다. 변수들에 대한 차원 축소는 인자분석을 포함하는 은닉변수(latent variable) 모형을 적용할 수 있다. 또한 인자분석에 의한 인자점수를 자기조직화 지도(Self Organizing Maps)[5] 알고리즘과 같은 자율학습(unsupervised learning) 신경망의 입력변수들에 대해 적용하여도 본 논문과 같은 수준의 군집화 결과가 기대된다고 볼 수 있다. 인자점수가 신경망을 포함한 기계학습(machine learning) 알고리즘에 적용되는 문제는 향후 과제로 남긴다.

감사의 글 : 본 연구는 과학기술부 주관 뇌신경정보학 사업에 의해 지원되었음.

6. 참고문헌

[1] T. W. Anderson, " An Introduction to Multivariate Statistical Analysis" , John Wiley & Sons 1992

[2] Barnett, Vic, Lewis, Toby, " Outliers in Statistical data" , John Wiley 1994

[3] Dillon, William R., Goldstein Matthew, " Multivariate Analysis Methods and Applications" , John Wiley

[4] Johnson, Richard A., Wichern, Dean W., " Applied Mutivariate Statistical Analysis" , Prentice Hall 1992

[5] Teuvo Kohonen, " Self Organizing Maps" , Springer 1997

[6] Tom M. Mitchell, " Machine Learning" , WCB McGraw-Hill 1997

[7] 송문섭, 이영조, 조신섭, 김병천, " SAS를 이용한 통계자료 분석" , 자유아카데미 1992