2001년 한국정보처리학회 추계 학술발표논문집 제8권 제2호

보건 데이터 활용에 관한 연구(II)

임기영, 조은희
한밭대학교 전기전자제어 공학부
원광대학교 의무행정과
e-mail : limgy@hanbat.ac.kr
e-mail : ehjo@sky.wkhc.ac.kr

A study of the Health Data Application

Gi-Young Lim, Hee-Eun Cho
Dept. of Control & Instrumentation Engr, Hanbat National University.
Deptment of Medical Adminisration, Wonkwang Health Science College.

요 약

정규분포 등의 가정이 곤란한 복잡한 밀도 분포에 대해 데이터의 선험적인 지식 없이 해석하기 위해 다수의 항목이 되고 복잡한 밀도 분포를 가진 데이터를 보다 소수의 단순한 밀도 분포가 되는 그룹으로 분류하는 방법을 나타내었고 데이터를 그룹으로 분류하는데 표본에 의한 분류와 항목에 의한 분류를 할 수 있다. 선험지식을 사용하지 않고 데이터를 분류하면 Parzen의 창함수에 의한 추정과 대수우도에 의한 평가함수를 사용하는 것으로 복잡한 형상을 가진 밀도분포도 선험지식 없이 해석이 가능하다. 표본의 밀도분포와 항목의 밀도분포를 나타내기 위하여 다수의 밀도 분포의 합과 곱의 형으로 전개하는 방법을 보였고 제안하는 방법을 의도적으로 생성한 데이터에 적용하여 원래의 밀도분포에 따라 분류결과를 얻을 수 있었다.

1. 서론

데이터에서 새로운 지식을 추출하기 위해서는 가공되지 않은 전체의 데이터에서 선험 지식을 사용하지 않고 해석하는 것이 중요하다. 그 때문에 다차원공간상에서 복잡한 밀도분포를 가진 데이터에 대해 적용할 필요가 있다. 그러나 종래의 해석방법에서는 데이터에 대한 선험지식이 필요하고 또 정규분포의 가정을 필요로 하기 때문에 데이터에서의 새로운 지식추출에는 적용이 곤란하였다.

데이터의 선험적인 지식을 사용하지 않고 또 복잡한 밀도분포를 가지는 데이터에서도 밀도분포를 추정하기 위해서는 Parzen의 창 함수에 의한 추정방법을 사용한다.

통계적인 평가기준을 정의하기 위해 데이터가 발생하는 과정을 모델화 한다.

하나의 모집단에서 몇 개의 표본이 추출되는 경우를 생각한다. 모집단이 여러 개 있고 각 모집단에서의 표본을 모두 모아 전체 데이터를 구성한다고 하자. 이와 같은 모델에서 복잡한 밀도분포를 다수의 밀도분포의 합으로써 나타낼 수 있으며 군집화를 하는 것으로 다수의 단순한 형의 밀도분포로 전개 할 수 있다.

같은 방법으로 항목을 분할 할 수 있다. 항목의 분할에서는 모집단에서 몇 개의 검사값을 얻을 수 있다. 여러 개의 모집단에서 검사 값을 모아 한사람의 데이터를 얻을 수 있다고 할 때 밀도분포는 다수의 밀도분포의 곱으로 나타내게 된다.

의료분야의 데이터에서는 다수의 검사항목이 되어 각각 Kg, m, Hb등 처럼 단위나 스케일이 다르다. 이 때문에 밀도분포의 추정에 있어서는 검사항목간의 단위, 스케일의 차이를 고려할 필요가 있다. 일반적으로는 평균 분산에 정규화하는 방법이 사용되지만 정규분포의 가정이 필요하게 된다. 이것에 대해 Parzen의 창함수에 의한 추정법에서는 창함수를 검사항목에 맞춰 결정하는 것으로 단위, 스케일의 차이를 보정하는 것이 가능하다. 이때 항목에 가장 적합한 창 함수를 어떻게 결정하는가 라는 문제가 있다. 항목의 분할에 있어서는 항목간의 함수를 통계적인 방법으로 가장 적합한 창함수를 정하는 방법을 연구한다.

표본의 밀도분포와 항목의 밀도분포를 같이 나타내는 것으로 다수의 항목이 되는 복잡한 밀도분포를 다

수의 밀도분포의 합과 곱의 형으로 전개할 수 있다. 이것의 각각의 밀도분포는 다변량 해석법을 적용 할 수 있다.

또 표본과 항목의 분할의 과정을 계층적으로 나타내는 것으로 표본의 그룹과 항목과의 함수를 동시에 나타내게 하고 이것에 의해 계층의 tree를 찾는 것으로 질병과 항목과의 함수나 질병의 식별에 유용한 항목 등의 정보를 얻을 수 있다. 표본의 분할에서는 모집단의 계층적인 구조를 2 계층으로 나타낸다. 또 항목의 분할에서도 마찬가지로 2계층으로 나타낸다. 이 때문에 2개를 합치기 위해서는 2개의 2계층을 어떻게 편성하는지가 문제가 된다. 그래서 데이터의 밀도분포에 맞춰 어느 쪽의 분할을 우선적으로 행하여 표본 및 항목의 분할을 하나의 2계층으로 나타낸다.

2. 표본의 계층적 분류

진찰 진단을 받은 사람, 각종의 검사를 받은 사람에 대해서 검사결과를 표본이라 부르고 다수의 사람에 대해 표본의 집단을 데이터라 부르기로 한다. 표본이 N개의 검사항목이 될 때, 각 검사항목은 검사 값을 고리(원형)로 하는 N차원공간으로 투영할 수 있다. 한편 심전도, 심근도, 뇌파와 같은 파형데이터, X선 화상, 내시경화상, CT화상과 같은 화상 데이터의 경우에도 마찬가지로 다 차원 공간에 투영할 수 있다. 여기에서는 데이터가 N개의 점으로 샘플링 되어 있다면 각 샘플링 점을 고리로 하는 차원공간에 투영하는 방법을 사용한다. 표본을 다차원공간에서의 분포에 따라 분류하는데는 다차원공간에서의 밀도 분포를 고려할 필요가 있다. 본 연구에서는 정규분포 등의 근사가 곤란한 복잡한 밀도분포에 대해 데이터의 선험적인 지식 없이 밀도분포를 추정하기 위해 Parzen의 창 함수방법을 사용한다.

2.1. 데이터의 발생 모델과 추정

데이터의 발생에 대해 아래와 같은 모델을 생각한다. 데이터가 복수개로 있는 모집단에서 무작위로 추출된 표본의 집합이라고 한다. 복수인 모집단 중에서 유사한 성질을 가진 모집단의 집합을 정리해 하나의 모집단이라 간주한다. 이것을 반복하는 것으로 계층적인 구조를 가진 모집단의 집합을 생각할 수 있다.

지금, 표본 집합의 실현 값으로써 관측된 데이터가 주어짐으로써 데이터를 근거로 모집단을 추정하는 것을 생각한다. 이 때문에 우선 각 데이터가 그 모집단에서 표본인지를 가정하고 가정에 따라 모집단의 밀도분포를 추정한다. 추정한 밀도분포가 관측된 데이터에 대해 어느 정도 타당한지는 각 데이터를 어느 모집단에서의 표본이라고 가정하는가에 따른다.

그래서 추정한 밀도분포가 어느 정도 타당한지를 평가하기 위해 모집단의 밀도분포와 관측 데이터와 우도를 정의한다. 우도는 데이터와 모집단과의 대응에서 한 뜻으로 구할 수 있다. 가장 타당한 밀도분포를 구하기에는 우도가 최대가 되도록 데이터와 모집단과의 대응을 구한다. 여기에서는 데이터가 계층적인 구조를 가진 복수개의 모집단에서 무작위 표본이라는 모델을 생각해 우도로 주어지는 평가함수를 근거로 데이터에서 모집단의 계층구조를 추정해 데이터의 계층적인 군집화를 구한다. 이상을 정리하면 군집화를 아래의 3가지 스텝에서 구한다.

1) 샘플링 되어 다수의 샘플 값, 또는 화소 값이 되는 데이터를 다차원 공간에 투영한다.

2) 데이터의 발생 원으로서 모집단을 가정해 그 밀도분포를 대수우도를 원으로 한 평가함수에 의해 추정한다. 추정된 모집단의 밀도분포에서 데이터가 어느 쪽의 모집단에서 발생했는지를 판단한다. 같은 모집단에서 발생한 데이터의 집합을 하나의 군집이라 한다.

3) 데이터를 2개의 군집에 분할하는 조작을 차례 차례로 반복한다. 분할의 가정은 2분목을 사용해 나타낸다.

3. 평가함수

대수우도를 근거로 한 평가함수를 정의한다. 제안하는 계층적 군집화에서는 데이터는 평가함수에 의해 2개의 군집에 2분할된다. 각각의 군집의 데이터는 같은 평가함수에 의해 더욱더 2분할되어 이것을 반복한다. 여기에서는 임의 군집 C를 생각해 이것을 분할 $\{X_1, X_2, \ldots, X_M\}$하는 평가함수를 정의한다. 군집에 M개 데이터가 속한다고 한다. 여기에서 M개의 벡터가 그 모집단에서의 확률 변수라 하고, $\{x_1, x_2, \ldots, x_M\}$가 그 관측 값이라 한다. 이때 $X_j(j=1\sim M)$는 확률 밀도분포 함수 $P_c(x)$에 따른다. 여기에서 M개의 벡터가 각각 모집단에서 독립적으로 얻어진 표본이라 하면, $\{x_1, x_2, \ldots, x_M\}$의 동시 밀도 함수는

$$f(x_1, x_2, \ldots, x_M) = P_c(x_1)P_c(x_2)\ldots P_c(x_m), \quad (3\text{-}1)$$

로 주어진다. 다음에 M개의 표본이 2개의 모집단

$\rho_1(x), \rho_2(x)$에서 무작위표본이라 한다. 이때 $\rho_1(x)$에 따라 모집단에서의 랜덤 벡터의 집합을 S_1로 나타내고, $\rho_2(x)$에 따라 모집단에서의 랜덤 벡터의 집합을 S_2로 나타낸다. 이때 $\{x_i \mid X_i \in S_1\}$되는 관측값의 집합은 군집 C_1를 구성하고, $\{x_j \mid X_j \in S_2\}$되는 관측 값의 집합은 군집 C_2를 구성한다.

군집화의 평가함수는 관측값 $\{x_1, x_2, \cdots x_M\}$이 얻어졌을 때 p_1, p_2, S_1 및 S_2에 대한 추정치 $\hat{p}_1, \hat{p}_2, \hat{S}_1$ 및 $\hat{S}_2$에 대해 대수우도로 정의한다. 대수우도는

$$l(\hat{p}_1, \hat{p}_2, \hat{S}_1, \hat{S}_2) = \log f(x_1, x_2, \ldots, x_M \mid \hat{p}_1, \hat{p}_2, \hat{S}_1, \hat{S}_2) = \sum_{i=1}^{M} \log f(x_i \mid \hat{p}_1, \hat{p}_2, \hat{S}_1, \hat{S}_2)$$

$$f(x_k \mid \hat{p}_1, \hat{p}_2, \hat{S}_1, \hat{S}_2) = \begin{cases} p_1(x_k) \text{ for } x_k \mid X_k \in \hat{S}_1 \\ p_2(x_k) \text{ for } x_k \mid X_k \in \hat{S}_2 \end{cases} \quad (3-3)$$

2개의 집합에 $\hat{C}_1, \hat{C}_2$에 분할 할 수 있고 $\hat{P}_1, \hat{P}_2$를 추정할 수 있다.

이것에서 $\hat{P}_1, \hat{P}_2$은 $\hat{C}_1, \hat{C}_2$ 의 함수로서 나타낼 수 있고,

$$\hat{P}_1(x) = \frac{1}{m} \sum_{i \mid x_i \in C_1} W(x, x_i)$$

$$\hat{P}_2(x) \frac{1}{n} \sum_{j \mid x_j \in C_2} W(x, x_j) \quad (3-4)$$

여기에서 m, n는 각 군집에 속한 데이터 개수이다. 이 식을 사용하는 것으로 모집단을 추정하는 것은 $\hat{C}_1, \hat{C}_2$ 의 추정으로 바꿔둘 수 있다.

$$l(\hat{p}_1, \hat{p}_2, \hat{C}_1, \hat{C}_2) = \sum_{k \mid x_k \in C_1} \log\left[\frac{1}{m} \sum_{i \mid x_i \in C_1} W(x_k, x_i)\right] + \sum_{l \mid x_l \in C_2} \log\left[\frac{1}{n} \sum_{j \mid x_j \in C_2} W(x_l, x_j)\right]$$

대수우도는 $\hat{C}_1, \hat{C}_2$의 함수로서,

(3-5) 라 쓸 수 있다.

이것에 의해, $X_j (j = 1 \sim M)$가 군집 $\hat{C}_1, \hat{C}_2$로 분할되면, (3-5)식에서 대수우도를 구할 수 있고 이것을 최대로 하는 듯한 $\hat{C}_1, \hat{C}_2$를 가장 적합한 군집이라 한다.

(3-5)식을 최대라 하는 가장 적합한 분할을 해석적으로 구할 수 없기 때문에 수치계산에 의해 구한다.

여기에서는 simulated annealing을 사용하는 경우에 대해 말한다.

4. 고찰

최초로 3개의 정규분포가 되는 데이터에 대해 적용한다 그림 3. 5(a)은 데이터의 분포를 나타낸다. 데이터는 3개의 같은 법인 2차원 정규 분포에 따른 군집이 된다. 각각의 군집 2 차원 정규분포는

$$\rho(x, y) = \frac{1}{2\pi \times A^2} \exp\left[\frac{(x - B_x)^2 + (y - B_y)^2}{2 \times A}\right], \quad (3-7)$$

로 나타낸다.

데이터를 계층적으로 군집화를 하는 방법을 제안했다. 데이터에 대해 전처리나 실험 지식을 사용하지 않게 군집화를 행하기 때문에, 대수우도를 근거로 평가함수를 정의했다. 평가함수를 구한 후에, Parzen의 창함수에 의해 추정한 밀도분포를 사용한다. 평가 함수를 최대라 하는 데이터의 2분할을 차례로 구하는 것으로, 계층적인 군집화를 얻는다.

제안하는 방법을 의도적으로 생성한 데이터에 적용하는 것으로, 원래의 밀도분포에 따라 분류결과를 얻을 수 있다. 종래의 방법에서는 선험 지식이 없다고 바르게 분류되지 않은 듯한 예에 대해서도 본 방법에서는 선험지식 없이도 분류를 할 수 있었다.

본 방법의 유효성을 확인하기 위해, 심전도에 대해 적용했다. 그 결과, 계층을 명한 것에 의해 다른 타입의 이상파형을 분류하는 것을 확인할 수 있었다.

5. 결 론

본 연구에서는 정규분포 등의 근사가 곤란한 복잡한 밀도 분포에 대해 데이터의 선험적인 지식 없이 해석하여 새로운 지식을 얻기 위해 다수의 항목이 되고 복잡한 밀도 분포를 가진 데이터가 주어졌을 때 보다 소수의 단순한 밀도 분포가 되는 그룹으로 분류하는 방법을 제안하고 의도적으로 생성한 데이터에 적용하여 원래의 밀도분포에 따라 분류결과를 얻을 수 있었다.

참고문헌

1) 김기영, 전명식, SAS 군집분석, 1991
2) 한국보건정보 교육학회, 보건정보학 개론, 2000
3)허명회, 다변량자료분석, 자유아카데미, 1999
4) 김우철,김재수외, 일반통계학, 영지문화사, 1997
5)염준근, 선형회귀분석, 자유아카데미, 1993

6) Stephen Polgar, Shane A, Tomas, "Introduction to research in the health science", Churchill Livingstone, 1995

7) M.A. Kraaijveld,"A parzen classifierwith an improved robusteness against deviations between training and test data",Pattern Recognition Letters, vol 17. 1996

8)K.Rose, E. Gurewitz and G. Fox, " A deterministic annealing approach to clustering" Pattern Recognition Letters, vol 11. no 9, 1990

9)K.Fukunaga, Introduction to Statistical Pattern Recognition, Academic press, INC. 1993

10)G.A. Babich and O.I.Camps, "Weighted Parzen Windows for Pattern Classification" IEEE Trans. Patt. Anal.Mach. Intell.,vol.18, .no5,1966

11) W.W. Daniel, Biostatistics, John Wiley & sons, New York, 1987

12) Engel G, The need for a new medical model, Science, 1977

13) G.J Mclachlan, Cluster analysis and related technicque in medical research, 1992