Adaptive Nearest Neighbors를 활용한 결측치 대치

전 명 식[1] 정 형 철[2]

요 약

비모수적 결측치 대치 방법으로 널리 사용되는 k-nearest neighbors(KNN) 방법은 자료의 국소적(local) 특징을 고려하지 않고 전체 자료에 대해 균일한 이웃의 개수 k를 사용하는 단점이 있다. 본 연구에서는 KNN의 대안으로 자료의 국소적 특징을 고려하는 adaptive nearest neighbors(ANN) 방법을 제안하였다. 나아가 microarray 자료의 경우에 대하여 결측치 대치를 통해 KNN과 ANN의 성능을 비교하였다.

주요용어 : adaptive nearest neighbors, k-nearest neighbors, 결측치 대치, microarray 자료

1. 소개

자료에 결측치가 발생했을 때, 이를 대치하는 대표적인 비모수적 방법으로 k-nearest neighbors(KNN) 방법이 있다. KNN은 가장 가까운 k 개의 이웃을 택한 후, 이들 k 개의 관찰치들을 사용하여, 결측치를 추정하는 방법이다. 그런데, KNN은 전체 자료에 대하여 고정된 k를 사용하므로 각각의 결측치의 위치가 지니는 국소적 특징이 고려되지 않는다. 따라서 자료의 국소적 특징을 고려하여 각 결측치의 위치에 따라 이웃의 개수 k를 변화시키는 adaptive nearest neighbors(ANN) 방법을 제안하고자 한다. 나아가, microarray 자료의 결측치 대치에 있어 제안된 ANN 방법이 KNN보다 통계적으로 효율적이고 강건함을 실증적으로 보이고자 한다. 2장에서는 ANN 대치방법을 소개하였고, 3장에서는 자료분석을 다루었다.

2. ANN 대치방법

2.1 ANN 알고리즘

p개의 변수와 n개의 개체들로 구성된 $n \times p$ 자료행렬 $X=(x_{ij})_{n\times p}$가 주어졌다고 하자. 여기서 x_{ij}는 i번째 개체의 j번째 변수의 관찰치를 나타낸다. 적어도 하나 이상의 결측치가 포함되어 있는 r개의 개체들에 대한 자료를 $X_m=(x_{ij})_{r\times p}$이라 놓고, 결측치가 없는 자료행렬을 $X_c=(x_{ij})_{(n-r)\times p}$라 놓자. 이에 따라 $X=(X_m', X_c')$ 로 표기하기로 한다. 이제 m_j개의 결측치를 지닌 j번째 개체 $x_j \in X_m$를 결측부분 $x_{m(j)}=(x_{1j}, x_{2j}, \cdots, x_{m_j j})$와 관측부분 $x_{o(j)}=(x_{1j}, x_{2j}, \cdots, x_{o_j j})$로 표기하자. 여기서, $p=m_j+o_j$이며, 결측치의 총 개수는 $M=\sum_{j=1}^{r} m_j$이다. 또한 $R=(r_{ij})$을 x_{ij}의 관찰치가 있으면 $r_{ij}=0$, 관찰치가 없으면 $r_{ij}=1$인 결측치 지시행렬(missing indicator matrix)이라 놓자.

1) (136-701) 서울시 성북구 안암동 고려대학교 통계학과 교수
2) (450-701) 경기도 평택시 용이동 평택대학교 정보통계학과 부교수

이제, ANN 알고리즘은 다음과 같이 주어진다.

[단계1] 자료행렬 X 를 X_m과 X_c로 나눈다.

[단계2] $x_j \in X_m$에 대해, x_j와 $x_k \in X_c$와의 가중 유클리디안 거리 d_{jk}를 계산한다.

$$d_{jk} = d(x_j, x_k) = \left\{ n_{jk}^{-1} \sum_{i=1}^{n} r_{ij} r_{ik} (x_{ik} - x_{ij})^2 \right\}^{1/2} , \; x_k \in X_c$$

여기서 $n_{jk} = \sum_{i=1}^{n} r_{ij} r_{ik}$로 x_j와 x_k간 이용 가능한 관찰치의 수다.

[단계3] $d_{jk}^* = [d_{jk} + median(d_{jk}, k = 1, \cdots, c)]$를 계산한 후, 크기 순으로 정렬된 거리 벡터 $d_{j(k)}^*$에 대해, $d_{j(k)}^* / d_{j(1)}^*$ 값이 일정 증분 이하인 개체 k_j개를 x_j의 이웃으로 선택한다.

[단계4] 선택된 k_j개의 이웃을 C_v^*라 하고, v번째 결측치 x_{vj}에 대해 가중평균 $\hat{x}_{vj}$을 계산한다.

$$\hat{x}_{vj} = \sum_{k_j \in C_v^*} w_{k_j} x_{vk_j} = \sum_{k_j \in C_v^*} \frac{1}{d_{jk_j} \left(\sum_{k_j \in c_v^*} d_{jk_j}^{-1} \right)} x_{vk}$$

여기서, $\sum w_k = 1$로 w_k는 유사성 가중치다.

[단계5] 모든 $x_k \in X_c$에 대해 [단계2]부터 [단계4]를 반복하여 결측치를 대치한다.

2.2 평가 방법

결측치 대치에 대한 평가로 쉽게 생각할 수 있는 것은 결측치와 실제값과의 차이를 전체 결측치에 대해 계산하는 방법이다. 즉, 실제값을 x_{vj}, 대치된 값을 $\hat{x}_{vj}$라 놓으면, 전체 평균오차는

$$\sqrt{\sum_{v \in m_j, j \in X_m} |x_{vj} - \hat{x}_{vj}|^s / c_j \, M} \; , s = 1, 2, \cdots ,$$

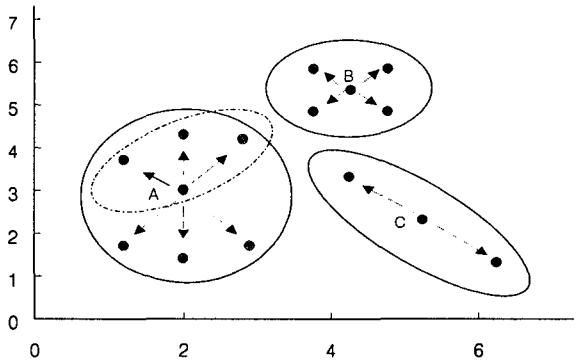
로 계산된다. 여기서, c_j는 표준화 상수로, 해당 개체나 전체 자료의 평균 등을 고려할 수 있다. Root mean square(RMS)는 위의 식에서 $s = 2$인 경우이며, root mean absolute(RMA)는 $s = 1$인 경우이다. RMA는 본질적으로 개개의 absolute error (AE)를 평균한 것으로 개개의 AE는 $e_i = |x_{vj} - \hat{x}_{vj}| / c_j, \; i = 1, \cdots, M$ 이다.

위의 RMS나 RMA는 결측치 전체를 평가하는 것으로 각 개체가 지니는 고유한 특징을 무시할 수 있다. 그러므로 RMS로 전체적인 특징을 살피고, AE로 개개의 특징을 살피는 것도 의미가 있다. 본 연구에서는 개개의 개체가 지니는 특징을 살피는 방법으로 각 개체의 지레값(leverage)을 고려하였다. 지레값은 $H = X(X'X)^{-1}X'$의 대각요소인데, 이는 i 번째 개체가 표본 평균벡터에서 얼마나 떨어져 있는가를 나타내는 척도("Mahalanobis" 거리와 비례함)라 하겠다. 우리는 m_j개의 결측치를 지니는 j번째 개체의 평균 AE인 $MAE = \sum_{i=1}^{m_j} e_i / m_j$ 를 각 개체의 지레값 별로 계산하여 추정값의 특징을 살펴보았다.

2.3 ANN의 간단한 예

15개의 개체를 지닌 세 개의 변수 X_1, X_2, X_3에 대해, $X_3 = X_1 - X_2$의 관계가 있으며, 변수 X_3

에만 결측치가 있다고 가정하자. [그림 1]은 주어진 15개의 개체들에 대해 X_1, X_2 를 그린 그림이며, 3개의 A, B, C 위치에서 X_3가 결측되어 있다. KNN과 ANN 대치 결과에 대한 RMS는 [표 1]에 제시되었다.



[그림 1] A, B, C 는 각각 결측치의 위치를 나타내며, 가로축은 X_1, 세로축은 X_2이다.

KNN	k	1	2	3	4	5	6	7
	RMS	1.443	0.863	0.772	0.907	1.070	1.220	1.449
ANN	증분비(%)	100	110	120	130	140	150	160
	RMS	0.866	0.645	0.004	0.004	0.004	0.791	1.222

[표 1] 주어진 자료에 대한 KNN과 ANN의 RMS

[표 1]은 증분비로 이웃의 개수 k를 자료의 특징에 따라 다르게 선정하는 ANN 방법의 장점을 보여주고 있다. 즉, 자료의 특징에 따라 적절한 증분비를 선정하여 이웃을 다르게 주는 방법이 전체자료에 대하여 고정된 이웃의 개수를 선택하는 KNN보다 RMS가 낮음을 볼 수 있다.

3. 자료분석

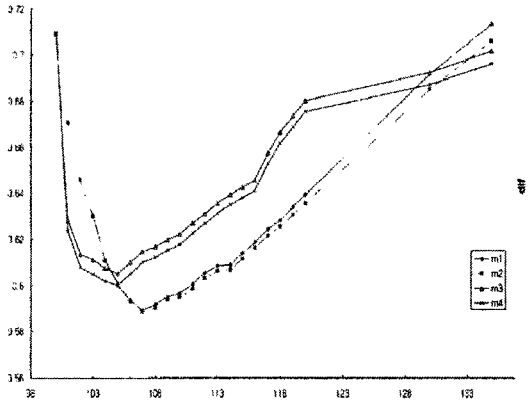
본 장에서는 microarray 발현정보에 대한 자료분석을 통해 ANN과 KNN의 특징을 살펴보도록 하겠다. 분석에 사용된 자료는 로그변환된 cDNA microarray 자료와 Oligonucleotide 칩에 의한 자료이다. 대치를 위해 결측치를 완전한 자료행렬에서 임의로 발생시킨 후, 실제값과 대치값의 차이를 계산하였다.

3.1 Melanoma data(cDNA microarray 자료)

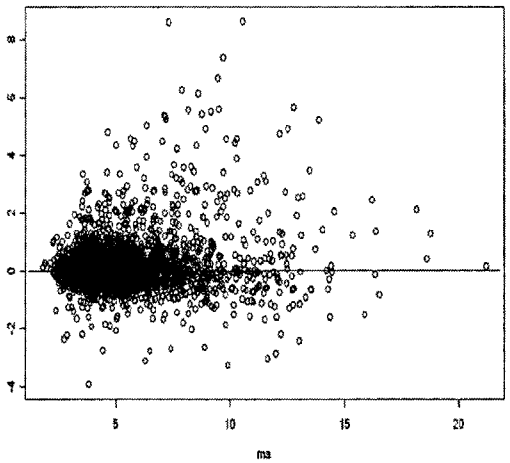
Melanoma 자료는 8150개의 인간 유전자에 대해 피부 흑색종 mRNA 표본과 reference mRNA 표본을 각각 Cy5, Cy3 처리하여 얻은 cDNA 자료이다. 동일한 조건에 대해서 31반복이 되어 있기에, 자료의 질을 고려하여 측정이 잘된 3613개의 유전자에 대해서 분석하였다. 자료는 http://www.nhgri.nih.gov/DIR/Microarray/Melanoma_Supplement/index.html에서 얻을 수 있다. 이 자료에 대해 결측확률을 0.1로 놓고, 독립반복시행 100회를 실시하였으며, 각 시행마다 얻어진 RMS의 평균 average root mean square(ARMS)를 계산하였다.

[그림 2]는 100회의 반복시행에 의한 ARMS를 보여준다. m1은 ANN을 이용한 평균예측, m2는 ANN을 이용한 가중평균예측, m3는 KNN을 이용한 평균예측, m4는 KNN을 이용한 가중평

균예측 결과이다. 편의상 ANN과 KNN의 ARMS를 겹쳐서 나타내었는데, 가로축은 ANN의 증분비율을 나타낸다. 여기서 ANN의 (m1, m2)와 KNN의 (m3, m4)를 같이 놓고 비교하는 것은 다소 문제가 있다. KNN의 (m3, m4)에 대해서는 가로축이 k를 나타내야 하기 때문이다. 하지만, ANN의 증분비율에 비례하여 KNN의 cutoff 인 k역시 증가하므로, 대략 비슷한 위치에 그림을 겹쳐놓아서 전체적인 경향을 서로 비교하고자 한다.



[그림 2] ARMS 그림



[그림 3] KNN과 ANN의 AMAE 차이

[그림 2]에서 ANN이나 KNN 모두 가중평균을 이용하여 예측하는 것이 평균을 이용한 예측보다 ARMS가 작음을 볼 수 있다. 또한 KNN과 ANN의 ARMS의 최소값을 비교하면, KNN은 k=15에서 ARMS가 0.601, ANN은 107%에서 0.587로 ANN의 ARMS가 작게 계산되었다. [그림 3]은 k=15에서 KNN의 MAE와 증분비 107%에서 ANN의 MAE의 차이를 각 유전자 개체의 지레값 별로 나타낸 결과이다. [그림 3]에서 세로축은 $KNN(AMAE) - ANN(AMAE)$, 가로축은 결측치와 평균간의 지레값의 함수인 "Mahalanobis" 거리를 의미한다. 여기서 세로축이 0보다 크면, KNN의 MAE가 ANN의 MAE보다 큼을 의미한다. [그림 3]에서 보면, 0보다 큰 개체가 전체의 70%이상을 차지하고 있다. 또한 평균에 가까운 유전자에서 결측치가 발생하면 KNN의 MAE와 AKNN의 MAE가 큰 차이가 없으나 평균으로부터 거리가 멀어질수록 KNN의 MAE가 좀더 커짐을 볼 수 있다.

3.2 Oligonucleotide data

여기서는 1000개의 유전자에 대해 같은 조건하에서 8반복 실험한 Tusher *et al.* (2001)의 oligonucleotide chip 자료로 결측치 대치를 비교하고자 한다. 주어진 자료에 대해 결측확률 0.1에 대하여 독립반복시행을 100회 실시하여 ARMS를 계산하였다. [표 2]는 KNN과 ANN의 ARMS를 보여준다.

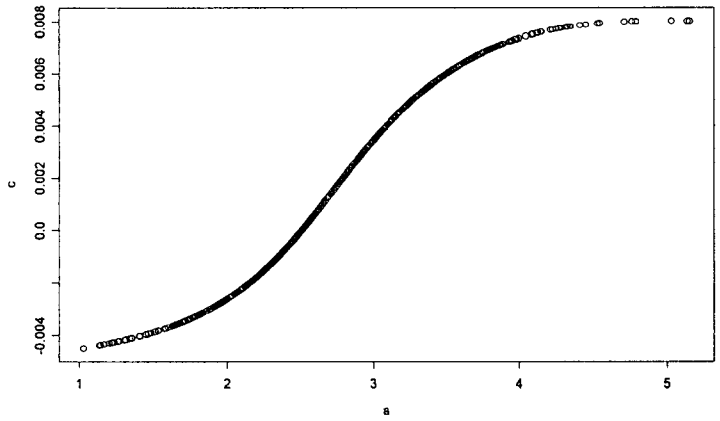
[표 2]의 결과 역시 가중평균예측 방법이 평균예측 방법보다 ARMS가 작음을 볼 수 있다. 그런데, KNN의 ARMS와 ANN의 ARMS의 차이는 극히 미미한 수준임을 볼 수 있다. 이는 임의로 발생된 결측치 별로 일정한 크기의 이웃들이 존재하거나, 표준화된 자료가 상당히 조밀한

(dense) 구조임을 짐작할 수 있다.

방법	최적 cutoff	평균예측	분산	가중평균예측	분산
KNN	56 개	1.086169	0.034336	1.085238	0.034164
ANN	130 %	1.086902	0.032836	1.084342	0.032896

[표 2] KNN과 ANN에 대해 평균예측과 가중평균예측방법으로 계산한 ARMS와 분산

각 개체의 지레값별로 KNN과 ANN의 특징을 파악하기 위해 [표 2]에서 제시된 최적의 조건에서 $KNN(AMAE) - ANN(AMAE)$를 계산하였다. 여기서, 두 방법의 특징을 파악하기 위해 $KNN(AMAE) - ANN(AMAE)$에 대한 평활결과를 [그림 4]에 제시하였다. [그림 4]에서 가로축은 결측치와 평균간의 "Mahalanobis" 거리를 의미한다. [그림 4]는 본 자료에 대한 ANN 대치와 KNN대치의 특징을 잘 보여준다고 하겠다. 즉, 평균으로부터 "Mahalanobis" 거리가 2.5이하에서는 평활결과가 음수(-), 2.5 이상에서는 양수(+)임을 볼 수 있다. 본 자료의 평균 "Mahalanobis" 거리는 2.7이다. 우리는 이로부터 다음의 사실을 추론할 수 있다. 즉, 위 자료에 있어, 평균에 가까이 있는 유전자에서 결측치가 발생했을 때는 KNN의 MAE가 ANN의 MAE보다 작으나 그 차이는 상대적으로 작은 수준이며, 평균에서 멀리 떨어진 유전자에서 결측치가 발생했을 때는 ANN의 MAE가 KNN의 MAE보다 작다는 사실이다. 또한 본 자료는 평균 "Mahalanobis" 거리 2.7이하에 약 63%의 유전자가 위치하는데, 결측치 발생이 임의이므로, KNN과 ANN의 ARMS에는 큰 차이가 나타나지 않았다고 유추된다.



[그림 4] "Mahalanobis" 거리에 따른 $KNN(AMAE) - ANN(AMAE)$ 평활 결과

4. 결론 및 논의

본 연구에서는 결측치 대치 방법으로 널리 사용되는 KNN에 대한 대안으로 ANN 대치 방법을 제안하였고, microarray 자료에 대한 사례를 다루었다. 특히, microarray 자료에 대한 결측치 대치에 대한 연구로 Troyanskaya *et al.* (2001), Oba *et al.* (2003), Zhou *et al.* (2003), 그리고 Nguyen *et al.* (2004) 등을 들 수 있다. 이들의 연구는 크게 KNN 대치방법과 베이지안 방법으로 유의한 유전자를 선택한 후 회귀분석(OLS 혹은 PLS) 등을 통한 대치방법이다. 그런데, microarray와 같은 대용량의 자료에서 KNN 대치방법은 다른 대치 방법에 비하여 계산이 간편

하며, 관련 생물학자 들이 개념적으로 쉽게 활용할 수 있는 통계적으로 강건한 방법이라 하겠다. 그러나, KNN은 이웃의 개수 k를 모든 결측치에 대해 동일하게 사용하는 단점이 있다. 그러므로 KNN의 단점을 극복하기 위해 대치결과 다시 활용하는 다중 KNN 방법이 고려된다 (Caruana, 2002). 그런데, ANN은 KNN의 단점을 보완하여 이웃을 결측치 별로 다르게 선정하므로 KNN에 비해 효율적이라 하겠다. 특히 ANN은 KNN과 비교하여 계산의 효율성에서 거의 차이가 나지 않는 방법이며, 다중 KNN 대치 방법과 같은 KNN에서 활용되는 모든 방법을 그대로 적용할 수 있는 장점이 있다. 즉, 다중 ANN 대치방법으로 ANN 대치방법을 개선할 수 있다.

추후 ANN 대치 방법을 모수적 방법인 EM 알고리즘을 통한 대치와 다중대치방법(schafer, 1997), 그리고 다중 KNN 대치 등과 모의실험을 통한 비교로 특징을 살펴보고자 한다.

참고문헌

Caruana, r. (2002) A non-parametric EM-style algorithm for imputing missing values. Center for Automated Learning and Discovery, Carnegie-Mellon University.

Nguyen, D., Wang, N. and Carroll, R. (2004) Evaluation of missing value estimation for microarray data. *Journal of Data Science*. (in press)

Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K. and Ishii, S. (2003) A Bayesian missing value estimation method for gene expression profile data, *Bioinformatics*, Vol. 19, 2088-2096.

Schafer, J.L. (1997) *Analysis of Incomplete Multivariate Data*, Chapman & Hall, London

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, Vol. 17, 520-525.

Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response, *Proceedings of the National Academy of Sciences*, Vol. 98, 5116-5121.

Zhou, X., Wang, X. and Dougherty, E. (2003) Missing value estimation using linear and non-linear regression with Bayesian gene selection, *Bioinformatics*, Vol. 19, 2302-2307.