The Korean Journal of Applied Statistics (2020) DOI: http://dx.doi.org/10.5351/KJAS.2020.33.3.269
33(3), 269–280

Missing values imputation for time course gene expression data using the pattern consistency index adaptive nearest neighbors

Heyseo Shin[a] · Dongjae Kim[a,1]

[a]Department of Biomedicine · Health Science, The Catholic University of Korea

(Received February 26, 2020; Revised April 11, 2020; Accepted May 6, 2020)

Abstract

Time course gene expression data is a large amount of data observed over time in microarray experiments. This data can also simultaneously identify the level of gene expression. However, the experiment process is complex, resulting in frequent missing values due to various causes. In this paper, we propose a pattern consistency index adaptive nearest neighbors as a method of missing value imputation. This method combines the adaptive nearest neighbors (ANN) method that reflects local characteristics and the pattern consistency index that considers consistent degree for gene expression between observations over time points. We conducted a Monte Carlo simulation study to evaluate the usefulness of proposed the pattern consistency index adaptive nearest neighbors (PANN) method for two yeast time course data.

Keywords: missing values imputation, adaptive nearest neighbors, pattern consistency index, time course gene expression data

1. 서론

정확한 진단이나 적절한 치료방법의 개발을 위해 질병들의 양상을 분자적인 수준에서 살펴볼 수 있는 마이크로어레이 기술을 통해 유전자 발현 양상에 따라 재분류하게됨으로써 많은 연구들이 진행되어왔다 (Park과 Lee, 2002). 마이크로어레이 자료는 광학현미경 슬라이드에 연구할 대상인 특정 세포나 특정 조직에서 얻어진 cDNA 타겟을 혼합시키면 이들이 서로 화학적 반응을 통해 빨간색과 녹색으로 유전자의 발현정도를 나타내며 이를 로그비로 변환한 자료이다. 전체 구성 유전자의 동시분석을 가능하게 하여 생물현상의 규명에 있게 되어 명확한 정보를 제공하는 마이크로어레이 기술은 이전보다 실험자의 시간과 노력을 덜어주었고 다양한 임상적 응용 가능성에 청신호를 보여주었다 (Park과 Lee, 2002). 마이크로어레이 유전자 실험을 시간의 흐름에 따라서 연속적으로 수행되면서 얻어지는 자료를 마이크로어레이 시간경로 유전자 발현자료라고 하며 유전자 발현수준이 시간에 따라 어떻게 변화하는지를 파악할 수 있게 된다 (Son과 Baek, 2005). 많은 생물학적 시스템은 동적 시스템으로 이루어져 있어 시간경로자료의 분석은 주어진 생물학적 과정이 전개되면서 유전자의 발현수준이 시간에 따라 변화의 정도를 파악할 수 있게 한다 (Son과 Baek, 2005). 또한 복잡한 여러 단계의 실험 과정을 거쳐 스폿팅 오류나 합성의

[1]Corresponding author: Department of Biomedicine · Health Science, The Catholic University of Korea, 222, Banpo-daero, Seocho-gu, Seoul 06591, Korea. E-mail: djkim@catholic.ac.kr

실패 및 불충분한 해상력, 긁힘 또는 지문, 이미지 훼손, 중도절단된 자료, 유전자 발현수준이 매우 낮아 유전적 정보를 나타내지 못하는 경우 등 다양한 원인으로 인해 결측값이 포함된다 (Kim 등, 2008). 하지만 마이크로어레이 자료의 연구 초기에는 흔히 결측 자료를 무시하고 분석하거나 실험을 다시 수행하는 방법을 취해 왔으나, 재실험을 통해 추가 자료를 얻는 방법은 매우 많은 비용과 시간이 소요되므로 현실적으로 어려움이 크다 (Kim 등, 2008). 대부분의 마이크로어레이 자료에서 발생하는 결측값의 개수는 전체 자료의 크기에 비해 유전자 단위로 비교했을 경우 비율이 높게 나타난다. 이는 결측값을 가지고 있는 전체 유전자들 가운데서 하나의 유전자에 하나씩 결측값이 있는 유전자의 비율이 높게 발생하기 때문이다 (Kim 등, 2008). 따라서 시간의 흐름에 따라 관측된 자료의 특성에 맞는 결측값을 추정하는 방법을 사용하고 결측값을 대체하여 주어진 자료의 유용성과 분석한 결과의 질을 높여야 한다.

완전 자료인 전체에 대해서 공분산 구조를 사용하여 결측값을 대체하는 방법인 전체적 방법과 전체 자료를 이용하기 보다 어느 부분의 강한 유사성을 띄는 자료들만을 이용하여 결측값을 대체하는 국소적인 방법으로 나눌 수 있다. 전체적인 자료를 활용하는 기존 방법들은 singular value decomposition (SVD) 방법, Bayesian principal component analysis (BPCA) 방법, partial least squares (PLS) 방법 등이 있고, 유사성을 지닌 일부분만을 사용하는 기존 방법들은 0대체인법, 평균대체법, 핫덱대체법, k-최근접 이웃(k-nearest neighbors; KNN) 방법, 적응 최근접 이웃(adaptive nearest neighbors; ANN) 방법 등이 있다. SVD 방법은 행 평균 대체값을 적용하여 완전 자료를 만들고 설명력이 높은 k개의 고유 유전자를 추출한다. 그리고 이들과 결측 유전자 간의 회귀모형을 적합시켜 추정된 회귀계수를 이용하여 결측값을 추정한다. 마지막으로 EM 알고리즘을 통해 기준 임계치에 도달할 때까지 반복하여 최종 추정치를 산출한다 (Kim 등, 2008). BPCA방법은 주성분 회귀분석을 이용하여 얻어진 주성분과 인자점수를 이용하여 베이지안 방법으로 결측값을 추정하고 PLS방법은 전체 유전자를 대상으로 유의한 주성분들을 추출하여 차원을 축소한 다음 회귀모형에 적합시켜 추정하는 방법이다 (Kim 등, 2008). 기존에 사용하던 국소적인 방법으로 0대체법은 모든 결측값을 0으로 대체하는 방법이고 평균대체법은 결측값을 제외한 관측값으로 평균을 구하여 대체하는 방법으로 사용이 편리하나 표준오차가 작게 추정되는 문제가 있다 (Kim과 Kim, 2018). 핫덱 대체법은 관측값 중에 임의로 하나를 선택하여 결측값을 대체하는 방법이지만 표준오차를 구하기 어렵다는 단점이 있다 (Kim과 Kim, 2018). Troyanskaya 등 (2001)이 제안한 비모수적 방법인 KNN 방법은 결측이 발생한 위치에서 가장 가까운 k개의 이웃을 활용하여 결측값을 추정하고 대체한다. 그러나 결측값을 추정하기 위한 이웃의 개수를 고정시켜 위치에 따른 특성을 무시하는 단점이 있다. 이러한 단점을 보완한 방법으로 Jhun 등 (2007)이 제안한 ANN 방법은 모수적 모형이 만족되지 않을 때도 강건성을 유지하며 계산 알고리즘이 간단하고 결측값마다 이웃의 개수를 다르게 정하여 추정하기 때문에 효율적이다. 하지만 기존의 방법은 반복적인 실험조건인 시간경로 유전자 발현자료에서 관측 시점간에 유전자의 발현수준이 변화하면서 발생하는 상관성을 간과하여 독립적인 변수로 다루어 왔었기에 결측값 추정의 정확성이 떨어질 것으로 예상된다.

본 논문은 이러한 문제점을 보완한 결측값 대체 방법으로 패턴 적응 최근접 이웃(pattern consistency index adaptive nearest neighbors; PANN) 방법을 제안하고자 한다. 이 방법은 결측값이 있는 위치에서 가장 가까운 거리에 있는 유전자들을 유동적으로 선택하는 ANN 방법의 장점과 시간의 흐름에 따라 변화하는 각 유전자의 발현정도가 일치하는지 파악할 수 있는 Son과 Baek (2005)이 제안한 패턴일치지수를 결합시켰으므로 결측값을 정확히 추정하고 대체할 수 있을 것으로 기대한다. 2장에서는 제안한 PANN 방법에 사용되는 유사성 거리와 패턴일치지수를 소개하였고 3장에서 예제를 통해 계산하는 방법을 자세히 설명하였다. 4장에서 기존 방법인 ANN 방법과 제안한 PANN 방법의 성능을 비교하기 위하여 두가지 실제 시간경로 유전자 발현 자료를 이용하여 모의실험을 실시하였다. 마지막으로 5장에서 결론 및 고찰로 마무리하였다.

2. 제안 방법

새로 제안한 PANN 방법은 결측값의 위치에 따른 이웃의 개수를 조정할 수 있는 ANN 방법에 Son과 Baek (2005)이 제안한 패턴일치지수를 이용하여 결측값을 추정하여 대체하는 방법이다.

결측값이 포함되어있는 시간경로 유전자 발현 자료는 p개의 유전자와 n개의 관측 시점으로 반복 측정한 자료이다. 이는 $p \times n$ 자료행렬인 $X = (x_{i,t_j})_{p\times n}$, $i = 1, 2, \ldots, p$, $j = 1, 2, \ldots, n$으로 표현할 수 있다.

여기서 x_{i,t_j}는 i번째 유전자의 시점 t_j에서 관찰된 값을 의미한다. x_{i,t_j}가 관측값이면 $r_{i,t_j} = 1$이고 결측값이면 $r_{i,t_j} = 0$으로 나타낸다. 적어도 하나 이상의 결측값을 가진 m개의 유전자 행들의 자료행렬을 $M = (x_{i,t_j})_{m\times n}$이라 하고, 결측값이 없는 $p - m$개의 완전한 유전자 행들의 자료를 $C = (x_{i,t_j})_{(p-m)\times n}$라고 표기한다. 자료행렬 M에 속하는 a번째 유전자 행 $x_a = (x_{a,t_1}, x_{a,t_2}, \ldots, x_{a,t_n})$와 자료행렬 C에 속하는 b번째 유전자 행 $x_b = (x_{b,t_1}, x_{b,t_2}, \ldots, x_{b,t_n})$의 거리를 유사성 거리 $d_{a,b}$라고 정의하고 x_a에서 결측값을 제외한 나머지 관측값만으로 계산한다.

유사성 거리를 측정하는 함수로 가중 유클리디안 거리(weighted Euclidian distance)와 피어슨 상관계수 거리(correlation distance)를 이용한다. 가중 유클리디안 거리는 유클리디안 거리에서 관측값에 가중치를 두어 두 유전자의 단순 거리를 계산하는 거리함수이고 피어슨 상관계수 거리는 피어슨 상관계수를 사용하여 두 유전자간 상관성을 거리로 나타내는 측도이다.

$$\text{Weighted Euclidian distance}: \; d_{a,b} = \sqrt{n_a^{-1} \sum_{j=1}^{n} r_{a,t_j} \left(x_{a,t_j} - x_{b,t_j}\right)^2}, \quad n_a = \sum_{j=1}^{n} r_{a,t_j},$$

$$\text{Correlation distance}: \; d_{a,b} = 1 - c_{a,b}, \quad a \neq b = 1, 2, \ldots, p$$

$$\left(c_{a,b} = \frac{\sum_{j=1}^{n} \left(x_{a,t_j} - \bar{x}_a\right)\left(x_{b,t_j} - \bar{x}_b\right)}{\sqrt{\sum_{j=1}^{n} \left(x_{a,t_j} - \bar{x}_a\right)^2} \sqrt{\sum_{j=1}^{n} \left(x_{b,t_j} - \bar{x}_b\right)^2}} \right).$$

가중 유클리디안 거리와 피어슨 상관계수 거리를 이용하여 유전자 간의 유사성을 거리로 측정하였다면 다음으로 시간에 따라 변화하는 유전자의 발현정도가 비슷한 양상을 띄는지 알 수 있는 측도로 패턴일치지수를 사용하여 고려하였다. 시간경로자료에서 사용되는 Son과 Baek (2005)이 제안한 패턴일치지수는 시간의 흐름에 따른 유전자들의 상승-하강-정체의 패턴과 최소-최대발현값을 나타내는 시점의 일치하는 정도를 수량화 하였다. 결측값이 포함되어 있는 자료행렬 M에 속하는 유전자 행 x_a에서 시점 t_j에서의 관측값인 x_{a,t_j}와 다음 시점인 t_{j+1}에서의 관측값 $x_{a,t_{j+1}}$는 적어도 두개 이상이어야 하며 두 점을 지나가는 직선의 기울기는

$$\text{slope}\,(a, t_j, t_{j+1}) = \frac{x_{a,t_{j+1}} - x_{a,t_j}}{t_{j+1} - t_j}, \quad a = 1, 2, \ldots, p, \; j = 1, 2, \ldots, n-1$$

이고, 직선의 상승-하강-정체의 정보를 가지는 함수는

$$L_{a,t_j,t_{j+1}} = \begin{cases} 1, & \text{slope}\,(a, t_j, t_{j+1}) > 0, \\ 0, & \text{slope}\,(a, t_j, t_{j+1}) = 0, \\ -1, & \text{slope}\,(a, t_j, t_{j+1}) < 0, \end{cases} \quad a = 1, 2, \ldots, p, \; j = 1, 2, \ldots, n-1$$

과 같이 정의한다.

위의 두 함수를 이용해 유전자 행 x_a와 x_b의 상승-하강-정체 패턴의 일치도 $A_{a,b}$는

$$A_{a,b} = \frac{1}{n-1}\sum_{j=1}^{n-1} I\left(L_{a,t_j,t_{j+1}} = L_{b,t_j,t_{j+1}}\right), \quad a \neq b = 1,2,\ldots,p,\ 0 \leq A_{a,b} \leq 1$$

로 여기서 지시함수 $I(L_{a,t_j,t_{j+1}} = L_{b,t_j,t_{j+1}})$는 자료행렬 C에 속하는 유전자 행 x_b의 $L_{b,t_j,t_{j+1}}$를 구하여 $L_{a,t_j,t_{j+1}}$와 같으면 1, 다르면 0을 갖는다. 결측값을 제외한 유전자 행 x_a와 x_b의 최댓값, 최솟값의 시점에 대한 일치도 $M_{a,b}$는

$$M_{a,b} = \begin{cases} 1, & T_a^{\min} = T_b^{\min} \text{ and } T_a^{\max} = T_b^{\max}, \\ 0.5, & T_a^{\min} = T_b^{\min} \text{ or } T_a^{\max} = T_b^{\max}, \\ 0, & T_a^{\min} \neq T_b^{\min} \text{ and } T_a^{\max} \neq T_b^{\max}, \end{cases} \qquad a \neq b = 1,2,\ldots,p$$

으로 정의하였고 $A_{a,b}$와 $M_{a,b}$를 이용한 유전자 행 x_a와 x_b의 패턴일치지수는

$$P_{a,b} = w_1 * A_{a,b} + w_2 * M_{a,b}, \quad a \neq b = 1,2,\ldots,p$$

로 정의한다. 여기서 $P_{a,b}$는 0과 1사이의 값으로 w_1과 w_2는 합이 1인 음이 아닌 실수로서, $P_{a,b}$에서 $A_{a,b}$와 $M_{a,b}$가 차지하는 비중이다.

두 가지의 유사성 거리와 패턴일치지수를 이용해 결측값을 대체하는 방법인 PANN 방법을 제안한다. 관측값으로만 계산하며 다음과 같은 단계를 따른다.

(단계 1) p개의 유전자와 n개의 관측시점을 가지고 있는 원자료 행렬 X를 결측값이 있는 행렬 M과 결측값이 없는 행렬 C로 나눈다.

$$X = \begin{pmatrix} x_{1,t_1} & x_{1,t_2}^{**} & x_{1,t_3} & \cdots & x_{1,t_n} \\ x_{2,t_1} & x_{2,t_2} & x_{2,t_3}^{**} & \cdots & x_{2,t_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{p,t_1} & x_{p,t_2} & x_{p,t_3} & \cdots & x_{p,t_n} \end{pmatrix}, \quad x_{i,t_j}^{**} : \text{missing value } (i = 1,\ldots,p,\ j = 1,\ldots,n),$$

$$M = \begin{pmatrix} x_{1,t_1} & x_{1,t_2}^{**} & x_{1,t_3} & \cdots & x_{1,t_n} \\ x_{2,t_1} & x_{2,t_2} & x_{2,t_3}^{**} & \cdots & x_{2,t_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{m,t_1} & x_{m,t_2} & x_{m,t_3} & \cdots & x_{m,t_n} \end{pmatrix}, \quad C = \begin{pmatrix} x_{1,t_1} & x_{1,t_2} & x_{1,t_3} & \cdots & x_{1,t_n} \\ x_{2,t_1} & x_{2,t_2} & x_{2,t_3} & \cdots & x_{2,t_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{p-m,t_1} & x_{p-m,t_2} & x_{p-m,t_3} & \cdots & x_{p-m,t_n} \end{pmatrix}.$$

(단계 2) 자료행렬 M에서의 a번째 유전자 행 x_a와 자료행렬 C에서의 b번째 유전자 행 x_b의 각 행들 간의 거리를 계산한다. 이 거리는 $d_{a,b}$로 가중 유클리디안 거리와 피어슨 상관계수 거리이다. 단, 결측값을 제외한 관측값만 계산한다.

(단계 3) 수정된 거리행렬 $d_{a,b}^*$과 최솟값으로 나눈 거리행렬 $d_{a,b}^{**}$

$$d_{a,b}^* = d_{a,b} + \text{median}(d_{a,b}),$$
$$d_{a,b}^{**} = \frac{d_{a,b}^*}{d_{a,(1)}^*}, \quad a \neq b = 1,2,\ldots,p$$

을 계산한 후, $d_{a,b}^{**}$의 값이 임계치 $q\,(>1)$ 보다 같거나 작으면 자료행렬 M에서의 유전자 행 x_a에 속하는 결측값을 대체할 수 있는 이웃집단 k_a로 정의하고 자료행렬 C에서의 x_b로 구성된다.

Table 3.1. Example 1

X	i	Time				$d_{a,b}$			$d^*_{a,b}$			$d^{**}_{a,b}$		
		0	5	10	15	$d_{1,b}$	$d_{2,b}$	$d_{3,b}$	$d^*_{1,b}$	$d^*_{2,b}$	$d^*_{3,b}$	$d^{**}_{1,b}$	$d^{**}_{2,b}$	$d^{**}_{3,b}$
	1	x^{**}_{1,t_1}	3	6	9									
M	2	10	2	x^{**}_{2,t_3}	4									
	3	1	9	8	x^{**}_{3,t_4}									
	4	4	9	5	7	3.697	5.598	2.449	7.256	9.157	6.996	1.658	1.731	1.000
	5	6	8	4	1	5.568	4.509	3.742	9.127	8.068	8.288	2.086	1.525	1.185
	6	8	3	4	2	4.203	1.732	5.802	7.762	5.291	10.348	1.774	1.000	1.479
	7	3	2	5	9	0.816	4.967	4.546	4.376	8.526	9.092	1.000	1.611	1.300
C	8	9	5	3	4	3.559	1.826	5.916	7.118	5.385	10.462	1.627	1.018	1.496
	9	9	8	6	5	3.697	3.559	4.796	7.256	7.118	9.342	1.658	1.345	1.335
	10	8	7	5	6	2.944	3.317	4.546	6.503	6.876	9.092	1.486	1.299	1.300
	11	7	1	9	6	2.708	2.160	5.802	6.267	5.719	10.348	1.432	1.081	1.479
	12	1	4	6	7	1.291	5.598	3.109	4.850	9.157	7.655	1.108	1.731	1.094

(단계 4) 이웃집단 k_a에 해당하는 $P_{a,b}$와 $d_{a,b}$를 이용하여 가중치 $W_{a,b}$는 $P_{a,b}$를 $d_{a,b}\times\sum_{b\in k_a} P_{a,b}/d_{a,b}$로 나누어 구한다.

(단계 5) 단계 4에서 구한 가중치 $W_{a,b}$와 행렬 C에서 이웃으로 선택된 관측값 x_{b,t_j}를 각각 곱하여 가중평균값을 구하고, 이 값으로 행렬 M의 결측값을 대체한다.

이웃을 선택하는 기준이 되는 임계치 q는 수정된 거리 $d^*_{a,b}$와 수정된 거리들의 최솟값 $d^*_{a,(1)}$의 비로 이웃집단을 결정하는데 수정된 거리를 사용하는 의미는 다음과 같다. 가중 유클리디안 거리를 $d_{a,b}$라고 하면, 이 거리들의 최솟값 $d_{a,(1)}$으로 나눈 두 거리의 비를 최대로 허용하는 한계값이 임계치 q가 되고 이보다 같거나 작으면 결측값의 이웃집단으로 구성하게 된다. 이는 이웃을 정할 때 결측값의 위치에 따른 국소적 특징을 반영한다. 두 거리의 비에서 분모인 거리들의 최솟값 $d_{a,(1)}$이 0에 가까워지면 거리 비는 무한히 커져서 이웃을 선택할 수 없게 되므로 통상적으로 거리행렬의 중앙값을 더한 수정된 거리행렬을 사용한다.

3. 예제

다음은 유전자의 개수 $p = 12$, 관측 시점 $n = 4$로 인위적인 자료를 생성하여 PANN 대치방법을 적용한 예제이며, 임계치 q는 1.3으로 설정하였고 유사성의 거리함수는 가중 유클리디안 거리를 사용하였다.

Table 3.1에서는 랜덤으로 추출한 데이터 값과 가중 유클리디안 거리인 $d_{a,b}$, 수정된 거리인 $d^*_{a,b}$, 수정된 거리 중에서 가장 작은 값으로 나눈 거리의 비 $d^{**}_{a,b}$를 나타내었다. 원자료 행렬 X에서 결측값이 포함되어 있는 행렬 M과 그렇지 않은 완전행렬 C로 나눈다. 행렬 M에서 3개의 결측값을 추정하기 위해서 각 열에 해당하는 결측값을 제외한 관측값을 사용하여 행렬 C에서의 유전자들과의 가중 유클리디안 거리를 계산한다. 그리고 가중 유클리디안 거리행렬에서 중앙값을 구하여 수정된 거리행렬을 구한다.

Table 3.1에서 계산된 값을 보면 x_1의 거리행렬 $d_{1,b}$에서 중앙값은 3.559, x_2의 거리행렬 $d_{2,b}$에서 중앙값은 3.559, x_3의 거리행렬 $d_{3,b}$에서 중앙값은 4.546이다. 수정된 거리행렬에서 $d^*_{1,b}$의 최솟값 $d^*_{1,(1)}$은 4.376, $d^*_{2,b}$의 최솟값 $d^*_{2,(1)}$은 5.291, $d^*_{3,b}$의 최솟값 $d^*_{3,(1)}$은 6.996이다. 수정된 거리행렬에서 최솟값을 구한 뒤 나눈 값으로 임계치 $q = 1.3$보다 같거나 작은 이웃집단을 구한다. 이웃집단 k_1에는 x_7, x_{12}로, k_2에는 x_6, x_8, x_{10}, x_{11}로, k_3에는 x_4, x_5, x_7, x_{10}, x_{12}로 이웃이 선택되었다.

Table 3.2. Example 2

X	i	$A_{a,b}$			$M_{a,b}$			$P_{a,b}$			$W_{a,b}$		
		$A_{1,b}$	$A_{2,b}$	$A_{3,b}$	$M_{1,b}$	$M_{2,b}$	$M_{3,b}$	$P_{1,b}$	$P_{2,b}$	$P_{3,b}$	$W_{1,b}$	$W_{2,b}$	$W_{3,b}$
	1												
M	2												
	3												
	4			0.667			1			0.833			0.449
	5			0.667			0.5			0.583			0.206
	6		0.333			0.5			0.417			0.306	
	7	0.667		0.333	1		0	0.833		0.167	0.693		0.048
C	8		0.333			0.5			0.417			0.290	
	9												
	10		0.333	0.333		0.5	0		0.417	0.167		0.160	0.048
	11		0.333			0.5			0.417			0.245	
	12	0.667		0.667	0.5		0.5	0.583		0.583	0.307		0.248

Table 3.2는 이웃집단 k_a에 속하는 상승-하강-정체 패턴의 일치도 $A_{a,b}$와 최소 및 최대값 시점의 일치도 $M_{a,b}$의 값들로 계산하여 패턴일치지수인 $P_{a,b}$를 구하였고 가중치 $W_{a,b}$까지 계산한 결과를 나타내었다. 패턴일치지수에서 상승-하강-정체 패턴의 일치도 $A_{a,b}$를 구하기 위해 직선의 기울기는 결측값을 제외하고 관측치만 고려한다. x_{1,t_j} 직선의 기울기 slope$(1, t_2, t_3)$과 slope$(1, t_3, t_4)$는 0.6이고 0보다 큰 값으로 L_{1,t_2,t_3}과 L_{1,t_3,t_4}은 1이 된다. 이와 같이 x_{7,t_j}의 직선기울기 slope$(7, t_2, t_3)$는 0.6이고 0보다 큰 값으로 L_{7,t_2,t_3}은 1, slope$(7, t_3, t_4)$는 0.8이고 L_{7,t_3,t_4}은 1이 된다. 일치도 $A_{1,7}$를 구하면 $(1+1) \times 1/(4-1) = 2/3$이다. 최댓값과 최솟값의 일치도 $M_{1,7}$를 구하면 x_{1,t_j}에서 결측값이 있는 시점은 제외한 최댓값의 시점 j는 6이고 최솟값의 시점 j는 2이고 x_{7,t_j}에서도 최댓값의 시점과 최솟값의 시점은 같으므로 1이다. 패턴일치지수 $P_{1,7}$은 $w_1 = w_2 = 0.5$로 두고 계산하면 $0.5 \times (2/3) + 0.5 \times 1 = 0.833$이 되고 같은 방법으로 $P_{1,12}$도 구하면 $0.5 \times 0.667 + 0.5 \times 0.5 = 0.583$이다. 다음으로 가중치 $W_{a,b}$를 계산하면 $W_{1,7}$은 $P_{1,7}$를 $d_{1,7} \times (P_{1,7}/d_{1,7} + P_{1,12}/d_{1,12})$로 나누어 계산하여 $0.833/(0.816 \times 1.472) = 0.693$이고 같은 방법으로 $W_{1,12}$은 $0.583/(1.291 \times 1.472) = 0.307$이다. 이웃집단에 속한 관측값을 각각 곱해준 후 모두 더한 가중평균값은 $(0.693 \times 3) + (0.307 \times 1) = 2.386$이 되고 x^{**}_{1,t_1}를 대체한다. 이와 동일한 방법으로 x^{**}_{2,t_3}은 5.095, x^{**}_{3,t_4}은 5.542로 대체할 수 있다.

4. 모의실험 계획 및 결과

본 논문에서 제안한 PANN 방법과 기존의 ANN 방법의 성능을 비교하기 위해 두 가지의 실제 시간경로 유전자 발현 자료를 사용하여 모의실험을 시행하였다. 첫 번째 시간경로 자료는 DeRisi 등 (1997)에 의한 유전자 발현의 대사 및 유전자 조절에 관한 연구로 6,400개의 유전자를 7개의 시점에서 반복하여 측정한 자료이며 유전자간 상관성이 높은 형태이다 (Kim 등, 2008). 기능이 알려진 유전자들에 대해서 동질성을 가지고 있지 않고 인식된 기능이 없는 400개 이상의 유전자도 포함되어 있기에 자료의 효율성을 고려하여 사전 클러스트된 6,068개의 유전자에 대해서 분석하였다. 해당 자료는 yeast functional genomics database (YFGdb) (http://yfgdb.princeton.edu/)에서 얻을 수 있다. 두 번째 시간경로 자료는 Spellman 등 (1998)에 의해 18개의 시점에서 6,178개의 유전자를 실험하는 연구 가운데 alpha-factor 실험 부분에 해당하고 관측 시점 간 상관성이 높은 형태이다 (Kim 등, 2008). 자료는 stanford genomic resources (SGR) (http://genome-www.stanford.edu/cellcycle/)에서 얻을 수 있다. 첫 번째

시간경로 자료를 A, 두 번째 시간경로 자료를 B로 부르기로 하겠다.

모의실험에 필요한 임계치 q는 1 과 2 사이의 값에서 0.1의 동일한 간격으로 선택하였고 전체 유전자 중에 랜덤으로 500개의 유전자를 추출하였다. 원(raw) 자료의 크기에 따라 결측률을 다르게 하여 A 자료에는 1%, 5%, 10%로, B 자료에는 5%, 10%, 20%로 설정하여 임의로 결측값을 생성하였다. 이와 같은 과정을 독립적으로 100회 반복하였으며 SAS 프로그램 9.4버전에서 균등분포를 이용하여 몬테카를로 모의실험을 시행하였다. 결측값을 추정한 후에 실제 자료의 참 값에 대한 적합 수준을 알 수 있는 정규화 제곱근 평균 제곱오차(normalized root mean square error; NRMSE)를 사용하였다.

$$\text{NRMSE} = \frac{1}{x'_{\max} - x'_{\min}} \left\{ \sum \frac{\left(x_{ij} - x'_{ij}\right)^2}{N} \right\}^{\frac{1}{2}},$$

여기서 x_{ij}는 실제값, x'_{ij}는 추정값, N은 결측값 수, $x'_{\max}$은 추정값 중 최댓값, $x'_{\min}$은 추정값 중 최솟값을 나타낸다. 추정된 값이 실제값과 유사하면 NRMSE가 0에 가깝고 그렇지 않다면 1에 가깝다. NRMSE가 작은 값을 기준으로 ANN 방법과 PANN 방법을 비교 평가하였다.

Table 4.1은 7개의 관측 시점으로 유전자간 상관성이 높은 자료 A의 모의실험 결과를 정리한 표이다. 결측률 1%일 경우에 가중 유클리디안 거리를 사용한 ANN 방법과 PANN 방법을 비교해보면, 임계치 q가 1.1일 때 PANN 방법의 NRMSE가 0.159로 ANN 방법의 NRMSE가 0.160으로 보다 작게 나타났다. 피어슨 상관계수 거리를 사용한 방법들의 비교에서는 임계치 q가 1.1일 때 ANN 방법의 NRMSE가 0.194로 PANN 방법의 NRMSE보다 낮아 높은 성능을 보였고 임계치 q가 1.2 이상부터 2.0 이하까지 PANN 방법이 높은 성능을 보였다. 결측률 5%일 경우에 가중 유클리디안 거리를 사용한 방법들 중에는 임계치 q가 1.1일 때 ANN 방법의 NRMSE가 가장 낮은 값인 0.097로 성능이 높았고 그 외 임계치 q가 1.2 이상 2.0 이하는 PANN 방법의 성능이 높았다. 피어슨 상관계수 거리를 사용한 방법들에서는 PANN 방법이 ANN 방법보다 각각 임계치 q에 NRMSE가 모두 작게 나타났다. 결측률 10%일 경우에는 가중 유클리디안 거리를 사용한 ANN 방법과 PANN 방법에서 보면 임계치 q가 1.1일 때 ANN 방법의 NRMSE가 0.086으로 성능이 높게 나왔고 나머지 임계치 q에서는 PANN 방법의 성능이 높았다. 피어슨 상관계수 거리를 사용한 ANN 방법과 PANN 방법도 비교하면 임계치 q에 상관없이 모두 PANN 방법이 정확하게 결측값을 추정하였다.

ANN 방법에서 가중 유클리디안 거리와 피어슨 상관계수 거리를 비교하였을 때 결측률 1%인 경우에 임계치 q가 1.1 이상 1.3 이하에서 가중 유클리디안 거리의 NRMSE가 0.160, 0.172, 0.193으로 작았고 그 외 임계치 q가 1.4 이상부터는 피어슨 상관계수 거리의 NRMSE가 0.212으로 가장 작게 나타났다. 결측률 5%인 경우에 임계치 q가 1.1 이상 1.4 이하에서 가중 유클리디안 거리가, 임계치 q가 1.5 이상부터는 피어슨 상관계수 거리가 높은 성능을 나타냈다. 결측률 10%인 경우도 살펴보면 결측률 5%인 경우와 같은 결과로 나타났다. PANN 방법에서는 가중 유클리디안 거리와 피어슨 상관계수 거리를 비교하면 결측률 1%일 경우 임계치 q가 1.1 이상 1.3 이하에서 가중 유클리디안 거리를 사용했을 때가 NRMSE가 0.159, 0.171, 0.189로 작게 나왔으며 임계치 q가 1.4 이상부터 2.0 이하까지 피어슨 상관계수 거리를 사용했을 때가 작게 나타났다. 결측률 5%인 경우에 임계치 q가 1.1 이상 1.4 이하에 가중 유클리디안 거리를 사용한 PANN 방법이 피어슨 상관계수 거리를 사용한 PANN 방법보다 NRMSE가 작게 나타났고 임계치 q가 1.5 이상 2.0 이하에서는 이와 반대로 피어슨 상관계수 거리를 사용한 PANN 방법이 가중 유클리디안 거리를 사용한 PANN 방법보다 NRMSE가 작게 나타났다. 결측률 10%인 경우에도 결측률 5%인 경우와 같은 결과로 임계치 q가 1.4를 기준으로 이하일 경우에 가중 유클리디안 거리를 사용한 PANN 방법이, 임계치 q가 1.5 이상일 경우 피어슨 상관계수 거리를 사용한 PANN 방법이 정확하게 결측값을 추정하였다.

Table 4.1. Average of the NRMSE based on 100 independent trials for dataset A

Data	Missing	q	ANN_{we}	PANN_{we}	ANN_{corr}	$\text{PANN}_{\text{corr}}$
A	1%	1.1	0.1600905	0.1599793	0.1942201	0.1951570
		1.2	0.1721296	0.1713490	0.1999636	0.1990262
		1.3	0.1935777	0.1892632	0.2070380	0.2043416
		1.4	0.2277207	0.2177844	0.2126082	0.2081953
		1.5	0.2719134	0.2495282	0.2171625	0.2110475
		1.6	0.3288821	0.2850590	0.2205119	0.2131564
		1.7	0.3960509	0.3195623	0.2227946	0.2145745
		1.8	0.4578736	0.3470732	0.2251912	0.2160261
		1.9	0.5006314	0.3644719	0.2273691	0.2171936
		2.0	0.5318479	0.3776816	0.2298040	0.2183950
	5%	1.1	0.0970414	0.0973832	0.1373338	0.1350324
		1.2	0.1047278	0.1041604	0.1452475	0.1411755
		1.3	0.1188049	0.1168676	0.1503938	0.1449282
		1.4	0.1430338	0.1371297	0.1543000	0.1473419
		1.5	0.1784134	0.1625820	0.1575158	0.1492523
		1.6	0.2268719	0.1921594	0.1597119	0.1504505
		1.7	0.2859140	0.2224621	0.1610825	0.1511577
		1.8	0.3442446	0.2469324	0.1622432	0.1517017
		1.9	0.3912024	0.2639776	0.1632404	0.1520931
		2.0	0.4270783	0.2753323	0.1642983	0.1524970
	10%	1.1	0.0862599	0.0867285	0.1278920	0.1250162
		1.2	0.0933886	0.0931591	0.1346938	0.1302767
		1.3	0.1069533	0.1058553	0.1387351	0.1332807
		1.4	0.1275454	0.1229089	0.1415877	0.1350285
		1.5	0.1589941	0.1462222	0.1438053	0.1363086
		1.6	0.2044010	0.1743621	0.1453593	0.1371387
		1.7	0.2597852	0.2030322	0.1462871	0.1375938
		1.8	0.3153270	0.2258667	0.1444784	0.1363865
		1.9	0.3604612	0.2409153	0.1477483	0.1382134
		2.0	0.2918951	0.2000486	0.1484910	0.1384746

ANN_{we} = ANN method using weighted Euclidian distance; PANN_{we} = PANN method using Weighted Euclidian distance; ANN_{corr} = ANN method using correlation distance; $\text{PANN}_{\text{corr}}$ = PANN method using correlation distance. ANN = adaptive nearest neighbors; PANN = pattern consistency index ANN.

따라서 유전자간 상관성이 높은 자료 A의 형태에서는 결측률이 10%이고 임계치 q가 1.1인 가중 유클리디안 거리를 사용한 ANN 방법이 NRMSE가 0.0863으로 가장 정확하게 결측값을 추정하였다. 그 다음은 가중 유클리디안 거리를 사용한 PANN 방법이었고 임계치 q가 증가할수록 NRMSE가 상대적으로 작은 값이 많이 나타났던 방법은 피어슨 상관계수 거리를 사용한 PANN 방법으로 나타났다.

Table 4.2는 18개의 관측 시점이고 관측 시점간 상관성이 높은 자료 B의 모의실험 결과를 정리한 표이다. 유사성 거리에 따른 ANN 방법과 PANN 방법을 살펴보면, 결측률 5%인 경우에서 가중 유클리디안 거리를 사용한 방법들을 비교하면 임계치 q가 1.1일 때 ANN 방법이 NRMSE가 0.160으로 가장 낮았고 임계치 q가 1.2 이상부터 2.0 이하까지 PANN 방법이 가장 낮게 나타났다. 피어슨 상관계수 거리를 사용한 ANN 방법과 PANN 방법을 비교하면 각각 임계치 q에 모두 PANN 방법의 NRMSE가 가장 작게 나타났다. 결측률 10%인 경우에서는 가중 유클리디안 거리를 사용한 ANN 방법과 PANN 방법을

Table 4.2. Average of the NRMSE based on 100 independent trials for dataset B

Data	Missing	q	ANN_{we}	PANN_{we}	ANN_{corr}	$\text{PANN}_{\text{corr}}$
B	5%	1.1	0.1606068	0.1612996	0.1336137	0.1314843
		1.2	0.2187549	0.2131827	0.1767360	0.1743168
		1.3	0.3727564	0.3445421	0.2251798	0.2176609
		1.4	0.5933437	0.5227984	0.2778186	0.2615469
		1.5	0.7766546	0.6551642	0.3342540	0.3054670
		1.6	0.8698007	0.7045806	0.3935685	0.3483216
		1.7	0.9264490	0.7264574	0.4165836	0.3603672
		1.8	0.9521834	0.7357322	0.5025179	0.4241438
		1.9	0.9656366	0.7401881	0.5526067	0.4560491
		2.0	0.9742520	0.7441330	0.5963664	0.4826802
	10%	1.1	0.1729225	0.1716602	0.1208910	0.1203675
		1.2	0.2312899	0.2229378	0.1495823	0.1466418
		1.3	0.3515736	0.3256458	0.1923286	0.1812998
		1.4	0.5373986	0.4664071	0.2419629	0.2189178
		1.5	0.7191658	0.5906185	0.2927219	0.2563348
		1.6	0.7164307	0.5598820	0.3484710	0.2920384
		1.7	0.8681574	0.6566053	0.4022318	0.3234311
		1.8	0.8958268	0.6656022	0.4544919	0.3504229
		1.9	0.9093846	0.6697370	0.5000507	0.3736015
		2.0	0.9177620	0.6718329	0.5374765	0.3907386
	20%	1.1	0.2182253	0.2167994	0.2060591	0.2072207
		1.2	0.2707421	0.2567761	0.2040654	0.2038443
		1.3	0.3581828	0.3113008	0.2179578	0.2138045
		1.4	0.4529762	0.3527422	0.2370688	0.2274182
		1.5	0.5275085	0.3797595	0.2615423	0.2430067
		1.6	0.5774985	0.3919490	0.2886470	0.2568702
		1.7	0.6006464	0.3961863	0.3161781	0.2704558
		1.8	0.6186407	0.3982756	0.3384917	0.2803963
		1.9	0.6301866	0.3991489	0.3602372	0.2870282
		2.0	0.6351775	0.4000494	0.3728618	0.2905353

ANN_{we} = ANN method using weighted Euclidian distance; PANN_{we} = PANN method using Weighted Euclidian distance; ANN_{corr} = ANN method using correlation distance; $\text{PANN}_{\text{corr}}$ = PANN method using correlation distance. ANN = adaptive nearest neighbors; PANN = pattern consistency index ANN.

살펴보면 임계치 q가 1.1에서 PANN 방법의 NRMSE가 0.171로 ANN 방법보다 낮게 나타났으며 이외의 임계치 q에서 모두 PANN 방법이 낮게 나타났다. 피어슨 상관계수 거리를 사용한 방법들 간 비교에서도 동일한 결과로 임계치 q가 1.1에서 PANN 방법의 NRMSE가 0.1203으로 ANN 방법보다 낮았고 이외의 임계치 q에서도 모두 결측값을 정확하게 추정하였음을 알 수 있다. 결측률 20%인 경우에서 가중 유클리디안 거리를 사용한 ANN 방법과 PANN 방법을 비교했을 때 임계치 q가 1.1에서 PANN 방법의 NRMSE가 0.216으로 가장 낮았고 피어슨 상관계수 거리를 사용한 방법간에서는 임계치 q가 1.1일 경우에 ANN 방법의 NRMSE가 0.206으로 PANN 방법의 NRMSE보다 낮았다. 하지만 이 경우를 제외한 나머지 임계치 q에서 피어슨 상관계수 거리를 사용한 PANN 방법이 우수하였다.

ANN 방법안에서 두 가지 유사성 거리를 비교하면, 결측률 5%인 경우에서 각각 임계치 q에서 피어슨 상관계수 거리가 가중 유클리디안 거리를 사용했을 때보다 NRMSE가 낮게 나타났고 결측률 10%와

20%인 경우에서도 동일한 결과로 피어슨 상관계수 거리를 사용한 ANN 방법이 가중 유클리디안 거리를 사용한 ANN 방법보다 정확하게 결측값 추정을 하였다. PANN 방법에서 유사성의 거리를 비교해보면 ANN 방법안에서 유사성의 거리를 비교했을 때와 동일한 결과로 결측률 5%, 10%, 그리고 20%인 경우에서도 각각 임계치 q에서 모두 피어슨 상관계수를 사용한 PANN 방법이 피어슨 상관계수를 사용한 ANN 방법보다 NRMSE가 작아 우수한 성능을 보였다.

따라서 관측 시점간 상관성이 높은 자료 B에서는 가중 유클리디안 거리를 사용하는 방법보다 피어슨 상관계수 거리를 사용한 방법이 많은 경우에 NRMSE가 작게 나타났다. 결측률 10%이고 임계치 q가 1.1일 때 피어슨 상관계수 거리를 사용한 PANN 방법이 NRMSE가 0.1203로 가장 정확하게 결측값을 추정하였고 그 다음으로 피어슨 상관계수 거리를 사용한 ANN 방법이며 임계치 q에 따라 상대적으로 NRMSE가 작은 값이 많았던 방법은 피어슨 상관계수 거리를 사용한 PANN 방법으로 성능을 보였다.

두 자료의 전반적인 공통점으로 기존 방법인 ANN 방법과 새로운 방법인 PANN 방법에서 피어슨 상관계수 거리를 사용한 PANN 방법이 임계치 q가 1.5 이상일 때 다른 방법보다 우수한 결과를 얻었다. 가중 유클리디안 거리를 사용하였을 때는 임계치 q가 증가함에 따라 NRMSE가 변화하는 폭이 넓었으나 피어슨 상관계수 거리를 사용하였을 때 상대적으로 NRMSE가 변화하는 폭이 좁았고 안정적으로 나타났다. 이는 유전자간 상관성을 고려하는 피어슨 상관계수 거리는 가중 유클리디안 거리보다 가깝게 측정되었기 때문이라고 할 수 있다.

5. 결론 및 고찰

본 논문에서는 시간경로 유전자 발현 자료에 대해서 전처리 과정으로 결측값을 추정하는데 발생하는 문제를 다루었고 이를 보완하여 자료의 유용성을 높이기 위한 새로운 방법을 제안하였다. 결측값의 위치에 따라 이웃의 개수를 고정하지 않고 유동적으로 조정할 수 있는 ANN 방법의 장점과 시간경로 자료의 특징인 관측 시점간 유전자 발현의 일치 정도를 고려한 패턴일치지수를 활용한 PANN 방법을 제안하였다. 모의실험은 실제 시간경로 유전자 발현 자료인 유전자간 상관성이 높은 자료과 관측 시점간 상관성이 높은 자료를 사용하여 ANN 방법과 PANN 방법을 비교함과 동시에 유사성 거리의 적합 수준도 비교하였다. 자료의 크기에 따라 20% 이하로 결측을 임의로 발생시켰고 결측값 추정에 대한 적합 수준은 NRMSE로 상대적으로 작은 값을 중점으로 방법들을 비교하였다.

모의실험의 결과로 유전자간 상관성이 높은 자료에서 결측률 1%인 경우에 임계치 q가 1.1 이상 1.3 이하에는 가중 유클리디안 거리를 사용한 PANN 방법이 NRMSE가 가장 낮았다. 결측률 5%와 10%인 경우에 임계치 q가 1.1에서는 가중 유클리디안 거리를 사용한 ANN 방법의 NRMSE가 0.097, 0.086으로 가장 낮았고 임계치 q가 1.2 이상 1.4 이하에는 가중 유클리디안 거리를 사용한 PANN 방법이 NRMSE가 낮게 나타났다. 그리고 임계치 q가 1.5 이상에는 피어슨 상관계수 거리를 사용한 PANN 방법이 NRMSE가 가장 낮게 나타났다. 결측률에 관계없이 임계치 q를 1.5 이하로 설정하면 피어슨 상관계수 거리보다 단순 거리를 측정하는 가중 유클리디안 거리를 사용한 방법들이 결측값을 정확하게 추정하였다. 임계치 q를 1.5 이상으로 설정하면 유전자간 상관성을 고려하는 피어슨 상관계수 거리가 가중 유클리디안 거리보다 가깝게 측정되었다고 할 수 있다.

관측 시점간 상관성이 높은 자료에서는 결측률 5%와 10%인 경우에 임계치 q가 1.1에서 피어슨 상관계수 거리를 사용한 PANN 방법의 NRMSE가 0.131, 0.120으로 가장 낮게 나타났다. 이외의 다른 임계치 q에서도 피어슨 상관계수 거리를 사용한 PANN 방법의 NRMSE가 낮게 나타났지만 결측률 20%인 경우 임계치 q가 1.1에서는 피어슨 상관계수 거리를 사용한 ANN 방법의 NRMSE가 0.206으로 낮게 나타났다. 이외의 다른 임계치 q에서는 피어슨 상관계수 거리를 사용한 PANN 방법의 NRMSE가 가장 낮

게 나타났다. 이는 유전자간 상관성을 고려한 피어슨 상관계수 거리가 가중 유클리디안 거리보다 작게 측정되었다고 볼 수 있으며, 관측 시점간 유전자 발현의 일치 정도를 패턴일치지수로 두어 결측값을 추정하였기에 다른 방법보다 우수한 성능을 보인것으로 여겨진다. 하지만 본 논문에서 제시한 방법은 결측값이 시간에 따라 연속적으로 되어 있는 경우에도 사용이 가능하나 연달아 있는 두 시점에는 관측값이 적어도 두개 이상 있어야 한다.

새롭게 제안한 PANN 방법은 기존의 ANN 방법에서 결측값의 위치에 따라 이웃의 개수를 정하는 장점과 관측 시점간 유전자 발현의 일치성을 가중치로 두는 패턴일치지수를 사용하였으므로 모의실험에서 기존 ANN 방법보다 정확한 추정 가능성을 보였다. PANN 방법을 시간경로 유전자 발현자료의 특성에 맞게 사용하고 유사성의 거리도 고려한다면 특히 관측 시점간 상관성이 높은 특징을 가진 자료에 대해서는 피어슨 상관계수 거리를 사용한 PANN 방법을 사용하는 것이 기존 방법들보다 개선된 결과를 보일 것으로 예상된다.

References

DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale, *Science*, **278**, 680–686.

Jhun, M., Jeong, H., and Koo, J. (2007). On the use of adaptive nearest neighbors for missing value imputation, *Communications in Statistics: Simulation and Computation*, **36**, 1275–1286.

Kim, K., Oh, M., and Son, Y. (2008). Missing values estimation for tine course gene expression data using the sequential partial least squares regression fitting, *The Korean Journal of Applied Statistics*, **21**, 275–290.

Kim, S. and Kim, D. (2018). Imputation method for missing data based on clustering and measure of property, *The Korean Journal of Applied Statistics*, **31**, 29–40.

Park, J. and Lee, I. (2002). Utilization of BioInforMetics with high efficiency array biotech, *News & Information for Chemical Engineers*, **20**, 431–440.

Son, Y. and Baek, J. (2005). A pattern consistency index for detecting heterogeneous time series in clustering time course gene expression data, *The Korean Journal of Applied Statistics*, **18**, 371–379.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle- regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, *Molecular Biology of the Cell*, **9**, 3273–3297.

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

시간경로 유전자 발현자료에서 패턴일치지수와 적응 최근접 이웃을 활용한 결측값 대치법

신혜서[α] · 김동재[α,1]

[α]가톨릭대학교 의생명 · 건강과학과

(2020년 2월 26일 접수, 2020년 4월 11일 수정, 2020년 5월 6일 채택)

요 약

시간경로 유전자 발현 자료는 마이크로어레이 실험을 시간에 따라 관측한 대용량의 자료로 유전자 발현 수준을 동시에 파악할 수 있다. 하지만 실험 과정이 복잡하여 다양한 원인들에 의해 결측값이 자주 발생한다. 본 논문에서는 시간경로 유전자 발현 자료에 대한 결측값을 추정하는 방법으로 패턴 적응 최근접 이웃(pattern consistency index adaptive nearest neighbors; PANN) 방법을 제안하였다. 이 방법은 국소적 특징을 반영하는 적응 최근접 이웃(adaptive nearest neighbors; ANN) 방법과 관측 시점간 유전자 발현의 일치 정도를 고려하는 패턴일치지수를 결합시킨 것이다. 제안한 PANN 방법의 효능을 평가하기 위하여 두 가지의 실제 시간경로 자료들을 사용하여 몬테카를로 모의실험(Monte Carlo simulation study)을 시행하였다.

주요용어: 결측값 대치법, 적응 최근접 이웃, 패턴일치지수, 마이크로어레이 자료, 시간경로 유전자 발현 자료, 국소적

[1]교신저자: (06591) 서울특별시 서초구 반포대로 222, 가톨릭대학교 의생명 · 건강과학과.
E-mail: djkim@catholic.ac.kr