Proceedings of Joint Conference of Korean Data And Information Science Society and The Korean Data Analysis Society, April 28 - 29, 2006, pp. 55-62

A Study on the Treatment of Missing Value using Grey Relational Grade and k-NN Approach*

Young-Min Chun[1], *Sung-Suk Chung*[2]

Abstract

Huang proposed a grey-based nearest neighbor approach to predict accurately missing attribute value in 2004. Our study proposes which way to decide the number of nearest neighbors using not only the deng's grey relational grade but also the wen's grey relational grade. Besides, our study uses not an arithmetic(unweighted) mean but a weighted one. Also, GRG is used by a weighted value when we impute a missing values. There are four different methods - DU, DW, WU, WW. The performance of WW(Wen's GRG & weighted mean) method is the best of any other methods. It had been proven by Huang that his method was much better than mean imputation method and multiple imputation method. The performance of our study is far superior to that of Huang.

Keywords : Grey System Theory, Grey Relational Grade, Imputation Method, Root Mean Square Error

1.서론

지금까지 불분명하고 불확실한 정보를 다루기 위한 다양한 이론들이 제안되어 사용되고 있는데, 그중에서 가장 널리 사용되는 것은 Zadeh(1965)에 의해 제안된 퍼지 이론이다. 그 이후에 Pawlak(1982)에 의해 제안된 러프 셑 이론(Rough Set Theory)과 Deng(1982)에 의해 제안된 그레이 시스템 이론(Grey System Theory)이 있다. 그 중에서 그레이 시스템 이론은 국내에 가장 덜 알려진 이론으로써, 중국과 대만 등지에서 많이 연구되는 분야이고 러프셑 이론과의 접목을 시도하는 연구들도 다양하게 이루어지고 있다(Wu, 2005; Zhang, 2004).

불분명하고 불확실한 정보를 다루는 이론들에서도 불완전한 정보를 다루는 문제에 대해서는 각 이론적 특성에 맞게 다양한 방법들이 연구되고 있으며, 이론적 배경은 다르지만 유사한 방법을 공

*이 논문은 산업자원부의 지역혁신 인력양성사업의 연구결과로 수행되었음.

[1](First Author) Instructor, Department of Statistical Informatics, Chonbuk National University, Jeonju, Korea. E-mail : zzari@chonbuk.ac.kr

[2]Professor, Division of Mathematics and Statistical Informatics, Chonbuk National University, Jeonju, Korea. E-mail : sschung@chonbuk.ac.kr

유하여 사용하기도 한다. 통계학을 포함한 여러 분야에서 결측값을 처리하는 부분에 대해 연구된 내용들을 살펴보면 결측값이 포함된 케이스를 제거하거나 무시하는 방법, 변수의 평균을 결측값으로 대체하는 평균대체법(mean imputation method), 전체 자료를 크기순으로 정렬하여 큰 값을 갖는 집단과 작은 값을 갖는 집단으로 나누어 대체함으로써 과소분산추정의 문제를 해결하려는 시도를 한 코헨대체법(Cohen imputation method), 유사한 직전 자료 또는 같은 자료의 값을 임의로 채워 넣는 핫덱대체법(hot-deck imputation method)과 콜드덱대체법(cold-deck imputation method), 보조변수와 강한 상관관계를 가질 때 이용하는 비율대체법(ratio imputation method), 결측값이 포함된 변수를 반응변수로 놓고 회귀분석을 이용하여 결측값을 추정하는 회귀대체법(regression imputation method), 결측값에 대하여 MLE(최대우도추정) 방법을 반복하여 추정하는 EM 알고리즘, 결측값에 대하여 여러 개의 대체값을 찾은 후 각각에 대해 통계분석을 실시하여 최적의 결과를 찾는 방법을 사용하는 다중대체법(multiple imputation method)등이 가장 대표적이고, 여기에서 파생되거나 변형된 매우 많은 대체 방법들이 존재한다.

Huang 등(2004)은 그레이 시스템 이론 중에서 그레이 관계 분석의 그레이 관계등급(GRG)을 이용하여 결측값과 가장 가까운 k개의 케이스를 찾는 방법을 통해 결측값을 대체하는 방법을 사용하였다. 실제로 이 방법은 평균대체법이나 다중대체법 보다 우수한 성능을 갖는 것으로 나타났다. 본 연구에서는 국내에는 잘 알려져 있지 않은 그레이 시스템 이론을 소개하고 Huang 등이 제안한 결측값 처리 방법을 소개하고자 한다. 그리고 Huang 등의 방법을 개선하기 위하여 다른 종류의 그레이 관계등급인 Wen의 그레이 관계등급을 이용하여 계산 속도를 빠르게 할 뿐만 아니라 Huang 등의 방법보다 더 좋은 성능을 나타내는 것을 보여줄 것이다. 또한 Huang 등이 사용했던 산술평균이 아닌 가중평균 방법을 이용하여 결측값을 대체함으로써 성능을 향상시켰다.

2. 그레이 시스템 이론

전체 시스템에 대한 정보가 불분명하고 불완전한 상태를 "grey system"이라고 한다. 그레이 시스템의 분야를 Wen(2004)은 grey generating, grey relational analysis, grey model, grey prediction, grey decision making, 그리고 grey control 등 여섯 가지로 나누고 있다.

본 연구에서는 여섯 가지의 세부 분야 중에서 grey generating과 grey relational analysis에 대해 자세히 알아보고 어떻게 사용하였는지를 알아볼 것이다.

2.1 grey generating

grey generating은 변수변환과 보간법의 두 가지로 나눌 수 있다. $x_i(l)$가 $i(i=0,1,\cdots,n)$번 째 케이스(sequence, case)의 $l(l=1,2,\cdots,m)$번 째 변수(item, variable)이라고 하자. 이 때 원래의 자료는 다음과 같다.

$$x_0 = (x_0(1), x_0(2), \cdots, x_0(m))$$
$$x_1 = (x_1(1), x_1(2), \cdots, x_1(m))$$
$$\vdots$$
$$x_n = (x_n(1), x_n(2), \cdots, x_n(m))$$

이 때, 전통적인 방법을 수정한 Hsia의 방법(1998)과 Chang(2000)의 방법을 주로 사용하여 변수변환을 하게 되는데, 어떤 값을 목표로 하느냐에 따라 <표 1>과 같이 다르게 사용된다.

<표 1> 변수 변환 방법 비교

	Hsia's method	Chang's method
최대값 기준	$z_i(l) = \frac{x_i(l) - \min_{all\ i} x_i(l)}{\max_{all\ i} x_i(l) - \min_{all\ i} x_i(l)}$	$z_i(l) = \frac{x_i(l)}{\max_{all\ i} x_i(l)}$
최소값 기준	$z_i(l) = \frac{\max_{all\ i} x_i(l) - x_i(l)}{\max_{all\ i} x_i(l) - \min_{all\ i} x_i(l)}$	$z_i(l) = \frac{-x_i(l)}{\min_{all\ i} x_i(l)} + 2$
단,	$z_i(l)$: 생성된 자료, $\min_{all\ i} x_i(l)$: 최소값, $\max_{all\ i} x_i(l)$: 최대값	

보간법에 대해 살펴보면, 그레이 시스템 이론은 연속된 자료의 개수가 4개 이상이면 자료의 해석이 가능하다. 따라서 연속된 네 개의 자료 사이에 결측값이 발생했을 경우에 그 값을 대체할 수 있다. 어떤 임의의 연속된 자료에 대한 일반적인 결측형태는 $a, \otimes, c, d$, $a, b, \otimes, d$와 같이 두 가지 경우로 나눌 수 있다. $\otimes$ 표시가 결측을 나타내는 것으로서, 이 값을 구하는 소프트웨어와 계산과정(2004)이 있지만 여기에서는 생략하겠다. 이 방법을 이용하여 구한 결과들은 일반적으로 구간값을 갖게 되는데 그 형태가 $[x, \infty), (\infty, x]$, 또는 (∞, ∞)와 같은 경우에 그 값을 제대로 잘 활용할 수 없게 된다.

2.2 그레이 관계 분석(Grey Relational Analysis)

그레이 관계분석의 가장 중요한 역할은 두 개의 서로 다른 케이스들 사이의 관계를 측정한다는 것이다. 이 때 사용되는 측정도구가 그레이 관계 등급(grey relational grade)이다.

$P(X)$를 한 가지 주제(theme)라고 하고, Q를 하나의 관계라고 가정할 때, $P(X); Q$를 그레이 관계 등급에서 인자 공간(Factor space)이라고 한다. 이때, $x_i(l) = (x_i(1), x_i(2), \cdots, x_i(m))$, $i(i = 0, 1, 2, \cdots, n)$, $l(l = 1, 2, \cdots, m)$이고, Non-dimensional, Scaling, Polarization의 세 가지 조건을 만족하면 'comparability'를 갖는다고 말하며, 세 가지 조건을 만족하는 경우의 공간을 그레이 관계 공간(grey relational space)이라고 하고 $P(X); \Gamma$로 표현하며 다음과 같은 네 가지 공리(axioms)를 갖는다.

(1) Normality : $0 \le \Gamma(x_i, x_j) \le 1, \forall_i, \forall_j$

(2) Duality Symmetric : 두 개의 케이스들 사이에서, $\Gamma(x_i, x_j) = \Gamma(x_j, x_i)$

(3) Wholeness : 세 개 이상의 케이스들 사이에서, $\Gamma(x_i, x_j) \overset{often}{\neq} \Gamma(x_j, x_i)$

(4) Closeness : $|x_i(l) - x_j(l)|$이 $\Gamma(x_i, x_j)$를 결정하는 중요한 요소이다.

따라서 comparability를 만족하면서 네 가지 공리를 따르는 $\gamma(x_i, x_j) \in \Gamma$가 존재할 때, $\gamma(x_i, x_j)$를 그레이 관계 계수(grey relational coefficient)라고 한다.

대표적인 GRG(그레이 관계 등급)로는 Deng의 GRG(1989)와 Wen의 GRG(2004)가 있다. Deng이 제안한 GRG를 구하기 위해 먼저 GRC(그레이 관계 계수)를 구하면 다음과 같다.

$$\gamma_{0j} = \gamma(x_0(l), x_j(l) = \frac{\Delta_{\min} + \zeta \Delta_{\max}}{\Delta_{0j} + \zeta \Delta_{\max}},$$

여기에서 $j = 1, 2, \cdots, n$, $l = 1, 2, \cdots, m$, x_0는 기준 케이스, x_j는 대상 케이스(inspected case)이고,

$$\Delta_{\min} = \min_{\forall_j} \min_{\forall_l} \left| x_0(l) - x_j(l) \right|, \Delta_{\max} = \max_{\forall_j} \max_{\forall_l} \left| x_0(l) - x_j(l) \right|$$

이며, $\Delta_{0j} = |x_0(l) - x_j(l)|$로써 x_0와 x_j사이의 거리(norm)를 의미한다. 또한 $\zeta \in [0, 1]$는 일반적으로 0.5를 사용한다. GRG는 GRC들의 산술평균을 의미하는 것으로 다음과 같다.

$$\Gamma_{oj} = \frac{1}{m} \sum_{l=1}^{m} \gamma(x_0(l), x_j(l))$$

여기에서 $j = 1, 2, \cdots, n$ 이다. GRG는 0과 1 사이에 존재하게 되는데 x_0와 x_j가 서로 비슷한 값을 가지면 1에 가까운 값을 갖게 되고, x_0와 x_i가 비슷하지 않은 값을 가지면 0에 가까운 값을 갖게 된다.

한편 Wen이 제안한 GRG를 구하는 수식은 다음과 같다.

$$\Gamma_{0j} = \frac{\Delta_{\min} + \Delta_{\max}}{\overline{\Delta}_{0j} + \Delta_{\max}}$$

여기에서 $j = 1, 2, \cdots, n$, $l = 1, 2, \cdots, m$, $\overline{\Delta}_{0j} = \frac{1}{m} \sum_{l=1}^{m} \left[\Delta_{0j}(l) \right]$이다.

Wen의 GRG는 Deng의 GRG와 비교할 때, GRC를 구하는 과정이 생략되어 있고 ζ를 사용하지 않기 때문에 Deng의 방법보다 속도가 빠른 장점이 있다.

그리고 Wen의 GRG는 0.5와 1사이에 존재하고, 그 값의 의미는 전체 케이스 중에서 기준 케이

스와 조사하는 케이스 사이의 관계 정도이다. 따라서 이 값이 1에 가까우면 가까울수록 조사한 케이스가 기준 케이스와 유사하다는 것을 의미한다.

3. A Modified Grey-Based Nearest Neighbor Approach

Huang 등은 GRG를 이용하여 불완전한 정보 자료를 처리하는 방법에 대해 살펴보았다. Huang 등은 GRG를 근접 이웃(nearest neighbor)을 찾는데 국한하여 사용하였는데, 본 연구에서는 GRG 자체를 근접 이웃을 찾는데 사용할 뿐만 아니라 결측값을 구하는데 가중치를 부여하는 방법으로 이용하였다. 또한 Huang 등이 사용한 Deng의 GRG 뿐만 아니라 또 다른 GRG인 Wen의 GRG를 이용하여 실험을 실시하였다.

3.1 A Grey-Based Nearest Neighbor Approach

어떤 케이스에 결측값이 발생했을 경우에 해당 변수에 해당하는 모든 값을 제거한 자료를 이용하여 해당 케이스와 가장 높은 그레이 관계 계수를 갖는 케이스들의 산술평균을 구하여 그 값을 결측값에 채워 넣는 방법으로 Huang 등이 2004년에 제안하였으며 그 절차는 다음과 같다.

step

1. Hsia의 방법을 이용한 자료의 전처리(data preprocessing or data generating)
2. 그레이 관계 계수(grey relational coefficient) 계산
3. 그레이 관계 등급(grey relational grade)을 계산한 후에 내림차순으로 정렬
4. 근접 이웃(nearest neighbor)의 수 k 결정
5. 산술평균을 이용하여 결측값 대체(imputation)

3.2 A Modified Grey-Based Nearest Neighbor Approach

본 연구에서 제안한 절차를 간단히 정리하면 다음과 같다.

step

1. Hsia의 방법을 통한 자료의 전처리
2. 그레이 관계 등급(grey relational grade)을 계산한 후에 내림차순으로 정렬
    1) deng의 방법
    2) wen의 방법
        - Wen의 방법을 사용하면 절차가 줄어드는 효과 발생
3. 근접 이웃(nearest neighbor)의 수 k 결정
4. 결측값 대체(imputation)

1) 산술평균 이용

2) 가중평균 이용

가중평균은 다음과 같은 공식을 사용하게 된다.

$$z_i^*(l)=\frac{\Gamma_0^1}{\Gamma_0^1+\Gamma_0^2+\cdots+\Gamma_0^k}\cdot z_i^1(l)+\cdots+\frac{\Gamma_0^k}{\Gamma_0^1+\Gamma_0^2+\cdots+\Gamma_0^k}\cdot z_i^k(l)$$

이 때, Γ_0^k는 GRG를 나타내며 $z_i^k(l)(i=1,2,\cdots,n,\ k=1,2,\cdots,n)$은 Γ_0^k를 갖는 자료의 값을 의미한다.

따라서 어떤 종류의 그레이 관계 등급과 평균을 사용했느냐에 따라 ① DU(Deng & Unweighted) 방법, ② DW(Deng & Weighted) 방법, ③ WU(Wen & Unweighted) 방법, ④ WW(Wen & Weighted) 방법과 같이 네 가지로 나눌 수 있다. 이 때 DU 방법은 Huang 등이 제안한 방법과 같다.

4. Simulation

본 연구에서 제안하는 방법을 확인하기 위하여 Huang 등의 논문에서 사용한 예제 데이터와 붓꽃(Iris) 데이터를 이용하여 실험을 실시하였다. Huang 등은 그들의 논문에서 이미 벌써 그들의 방법이 평균 대체법이나 다중 대체법보다 우수하다고 밝혔으므로 본 연구에서는 본 연구에서 제안하는 방법이 Huang 등의 방법보다 우수한 것을 밝히는 데에 중점을 두고 실험을 실시하였다.

실험방법은 leave one out cross validation을 사용하였는데 이 방법은 첫 번째 케이스부터 마지막 케이스까지 순차적으로 결측값을 발생시킨 후에 결측값이 발생한 케이스를 제외한 나머지 케이스의 정보를 이용하여 결측값을 대체하는 것을 말하는 것이다. 또한 근접 이웃의 개수를 1에서부터 50까지 변화를 주며 비교하였으며 방법들간의 성능 비교를 위하여 RMSE(Root Mean Square Error)를 사용하였다.

<표 2>에 나와 있는 자료는 Huang 등의 논문에서 예제로 다루었던 것인데, 보는 바와 같이 5개의 변수와 8개의 케이스로 이루어졌다. 자료는 미리 전처리 과정을 통해 가공되었으며 0에서 1사이의 값을 갖고 있다.

만일 A변수의 첫 번째 케이스에 있는 값인 0.92가 결측되었다고 가정하자. 이 값을 대체하기 위하여 Huang 등이 사용한 방법(DU)과 본 연구에서 제안한 방법들(DW, WU, WW)을 사용하였다.

<표 3>에서 보는 바와 같이 Huang 등의 논문에서 제안한 방법인 DU 방법보다 본 연구에서 제안한 방법들이 성능이 우수하거나 동일한 것으로 나타났다. 산술평균을 사용하는 것보다는 가중평균을 사용했을 경우에 성능이 좋은 것을 알 수 있으며, DU 방법과 WU 방법의 RMSE 값이 똑같이 나오는 이유는 케이스의 수가 적으면서 deng의 방법과 wen의 방법을 통한 GRG를 구했을 때, 기준 케이스에 대한 각 케이스의 GRG 값의 순위가 서로 같기 때문이다. 한편 이 자료의 경우에는 네

가지 방법 모두 근접 이웃의 개수를 2로 하였을 때 RMSE가 가장 작은 것으로 나타났다.

<표 2> example data

Cases	Items				
	A	B	C	D	E
x0	0.92*	0.94	0.25	0.07	0.84
x1	0.00	0.17	0.81	1.00	0.15
x2	0.86	1.00	0.00	0.23	1.00
x3	0.23	0.21	1.00	0.99	0.00
x4	0.85	0.82	0.21	0.00	0.93
x5	1.00	0.88	0.14	0.14	0.87
x6	0.96	0.95	0.09	0.13	0.85
x7	0.18	0.00	0.91	0.98	0.09

<표 3> 네 가지 방법들(DU, DW, WU, WW)의 RMSE 비교

	k(number of nearest neighbor)						
	1	2	3	4	5	6	7
DU	0.131053	0.111552	0.189788	0.269232	0.335926	0.398361	0.440853
DW	0.131053	0.111515	0.139557	0.18774	0.230315	0.275731	0.309985
WU	0.131053	0.111552	0.189788	0.269232	0.335926	0.397643	0.440853
WW	0.131053	0.111530	0.154194	0.215369	0.266383	0.318092	0.356476

DU(deng & unweighted mean), DW(deng & weighted mean),
WU(wen & unweighted mean), WW(wen & weighted mean)

붓꽃 자료의 경우에도 Wen의 방법을 이용하여 GRG를 계산한 WU와 WW 방법이 Deng의 방법을 이용하여 GRG를 계산한 DU와 DW 방법보다 작은 RMSE를 갖는 것으로 나타났다. 또한 같은 GRG 계산 방법을 사용한 경우에는 가중평균을 사용했을 경우에 가중평균을 사용하지 않은 경우보다 성능이 약간 좋은 것으로 나타났다.

5. 결론

실험 결과를 살펴보면, Deng의 방법을 이용하는 경우보다 Wen의 방법을 이용하는 경우에 계산 속도도 빠르고 RMSE도 적은 것으로 나타났다. 또한 결측값을 대체하는데 있어서 가중평균을 사용하는 경우에 그렇지 않은 경우에 비해 RMSE가 전반적으로 낮은 것으로 나타났다. Huang의 논문에서는 기준 케이스와 다른 케이스들 사이의 그레이 관계등급을 구하였으나, 근접 이웃을 결정하는 데에만 사용하였을 뿐 그 수치가 가지는 정보를 제대로 활용하지 못하였다. 하지만 본 연구에서는 그 수치를 가중값으로 사용하여 결측값에 대한 대체의 성능을 향상시키는데 사용하였다.

참고문헌

[1] Huang, C.-C. and Lee, H.-M. (2004). A Grey-Based Nearest Neighbor Approach for Missing Attribute Value

Prediction, *Applied Intelligence,* 20, 239-252.

[2] Chang, W. C. (2000). A Comprehensive study of grey relational generating, *Journal of Chinese Grey System,* 3, 53-62.

[3] Deng J. (1982). Control problems of grey systems, *Systems and Control Letters,* 5, 288-294.

[4] Deng J. (1989). *The basic couse of grey system theory,* HUST Publisher.

[5] Hsia, K. H. and Wu, J. H. (1998). A study on the data preprocessing in grey relational analysis, *Journal of Chinese Grey System,* 1, 47-54.

[6] Pawlak, Z. (1982). Rough Sets, *International Journal of Computer and Information Sciences,* 11 (1), 341-356.

[7] Wen, K.-L. (2004). *Grey systems : modeling and prediction,* Yang's Scientific Press.

[8] Wu, S., Liu, S., and Li, M. (2005). Study of integrate models of rough sets and grey systems, *Lecture Notes in Artificial Intelligence* (Subseries of Lecture Notes in Computer Science) 3613 (PART I), 1313-1323.

[9] Zhang, Q. and Chen, G. (2004). Rough grey sets, *Kybernetes,* 33, 2, 446-452.