*Journal of the Korean*
*Data & Information Science Society*
*2006, Vol. 17, No. 2, pp. 421 ~ 436*

# A Modified Grey-Based k-NN Approach for Treatment of Missing Value[1)]

Young M. Chun[2)] · Joon W. Lee[3)] · Sung S. Chung[4)]

## 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 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 relational grade, Grey system theory, Imputation method, Incomplete information system, Root mean square error

## 1.서론

그레이 시스템 이론(Grey System Theory)은 1982년 Deng(1982)에 의해 제안되었다. 불확실한 시스템의 행동을 추정하기 위해 한정된 데이터를 이용하는 이 이론은

---

1) This research was supported by the Program for the Training of Graduate Students in Regional Innovation which was conducted by the Ministry of Commerce Industry and Energy of the Korean Government.

2) Graduate Course, Department of Statistical Informatics, Chonbuk National University, Jeonju, Korea
E-mail : zzari@chonbuk.ac.kr

3) Professor, Division of Electronics and Information Engineering, Chonbuk National University, Jeonju, Korea
E-mail : chlee@chonbuk.ac.kr

4) Corresponding Author : Professor, Division of Mathematics and Statistical Informatics, Chonbuk National University, Jeonju, Korea
E-mail : sschung@chonbuk.ac.kr

중국에서 시작되어 중국과 대만에서 활성화되고 있고, 최근에는 여러 나라 사람들이 관심을 갖고 시도하는 분야가 되어가고 있으며, Pawlak(1982)에 의해 제안된 러프 셑 이론(Rough Set Theory)과의 접목을 시도하는 연구들이 이루어지고 있다(Wu(2005), Zhang(2004)). 그레이 시스템 이론을 통계학과 퍼지이론에 비교한다면, 먼저 자료의 개수에 있어서 통계학은 데이터의 개수가 되도록이면 많을수록 좋고, 퍼지이론은 데이터의 개수를 주로 경험에 의존하게 되는데 반해 그레이 시스템 이론은 단지 4개의 데이터를 가지고도 전개해 나갈 수 있다. 두 번째로 통계학은 자료의 분포가 전형적인 분포를 따른다는 가정 하에서 이론을 전개하고, 퍼지 이론은 소속함수에 이론적 기반을 두는데 반해, 그레이 시스템 이론은 어떤 형태의 분포도 가정하지 않는다는 차이점이 있다.

통계학 뿐만 아니라 다양한 학습 알고리즘들은 대부분 완전한 데이터를 기반으로 고안되고 발전되어 왔다. 하지만 실제 상황에서는 결측값이 포함된 불완전한 데이터를 처리해야 하는 경우가 많이 있다. 결측값을 처리하는 전통적인 방법은 결측값이 포함된 케이스를 제거하거나 무시하는 방법을 주로 사용하여 왔다. 최근에는 결측값을 포함한 케이스를 제거하지 않고 학습알고리즘을 적용하는 사례들도 있으며, 결측값을 새로운 값으로 대체하는 대체 방법을 개발하여 사용하기도 한다. 이런 무응답 대체 방법은 지금까지 상당히 다양한 분야에서 다양한 종류의 방법들이 개발되어 사용되고 있다. 무응답 대체 방법의 가장 큰 구분은 단일 대체법(single imputation method)과 다중 대체법(multiple imputation method)이다. 단일 대체법은 각 결측값에 대해 단 하나의 값으로 대체하는 방법으로써 사용하기는 쉬우나 분산이 감소한다는 단점이 있으며, 다중 대체법은 각 결측값을 하나 이상의 여러 개로 표현할 수 있는 것으로써 결과는 좋으나 여러 번 반복해야 한다는 단점이 있다.

우선 단일대체법에 해당하는 것으로서 가장 오래된 평균대체법(mean imputation method)이 있는데 이 방법은 분산을 과소추정하는 큰 단점이 있다. 분산의 과소추정 문제를 해결하기 위해 Cohen(1996)은 전체 자료를 크기순으로 정렬하여 큰 값을 갖는 집단과 작은 값을 갖는 집단으로 나누어 대체함으로써 과소분산추정의 문제를 해결하려는 시도를 하였다. hot deck imputation 방법(1980)과 cold deck imputation 방법은 통계적 배경이 부족한 사람들에게 이용되는 것으로써 상황에 따라 효율이 많이 좌우되는 단점이 있다. 비율대체법(ratio imputation method)은 보조변수에 강한 상관관계를 갖는 것으로써 두 변수사이에 강한 상관관계가 있을 때 효과적인 방법이다. 회귀대체법(regression imputation method)은 결측값이 포함된 변수를 반응변수로 놓고 회귀분석을 이용하여 결측값을 추정하는 방법이다. Little 등(2002)에 의해 제안된 EM algorithm은 결측값에 대하여 MLE(최대우도추정) 방법을 반복하여 추정하는 방법이다.

다중대체법은 Rubin(1976, 1987)에 의해 제안된 것으로서 결측값에 대하여 여러 개의 대체값을 찾은 후 각각에 대해 통계분석을 실시하여 최적의 결과를 찾는 방법으로 지금까지 알려진 것 중에서 가장 성능이 좋은 것으로 알려져 있다.

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.

불완전한 정보 시스템에 존재하는 결측값의 처리에 대한 문제는 많은 사람들의 관

심사이다. 결측값을 대체하기 위해 사용된 알고리즘의 수행속도가 얼마나 걸리느냐는 컴퓨터 하드웨어의 발달로 인해 어느 정도 해소가 되었다고 볼 수 있다. 지금은 얼마나 성능이 우수한 알고리즘을 사용하느냐가 더 중요한 문제이다. 또한 비슷한 성능의 알고리즘이라면 좀 더 속도가 빨라야 하고, 속도가 비슷하다면 좀 더 성능이 좋아야 할 것이다. 지금까지 상당히 다양한 분야에서 여러 가지의 알고리즘들이 쏟아지고 있으며 각각 나름대로의 타당성을 주장하고 있고, 자료의 종류와 각각의 상황에 따라 서로 다른 알고리즘들이 더 우수한 성능을 보이고 있다. 따라서 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.

본 논문의 구성은 다음과 같다. 2절에서는 그레이 시스템 이론을 소개하고 본 연구에서 관심을 갖고 있는 그레이 생성(grey generating)과 그레이 관계 분석(grey relational analysis)에 대해 알아볼 것이다. 3절에서는 Huang의 논문에서 사용된 방법을 살펴보고, 본 연구에서 제안하는 서로 다른 네 가지 종류의 방법들을 소개할 것이다. 4절에서는 두 가지 예제 자료를 이용하여 실험 결과를 보여줄 것이고, 5절에서는 결론 및 향후 연구 방향에 대해 소개할 것이다.

## 2. 그레이 시스템 이론

그레이 시스템 이론에서, “black”이라는 용어는 정보에 대해 아무것도 알려져 있지 않음을 의미하고, “white”라는 용어는 정보에 대해 완벽하게 알고 있음을 의미하며, “grey”라는 용어는 정보에 대해 부분적으로 알고 있음을 의미한다. 따라서 세상에 있는 모든 정보 시스템을 다음과 같이 세 가지로 나눌 수 있다. 첫 번째는 전체 시스템에 대한 정보가 분명한 “white system”이고, 두 번째는 전체 시스템에 대한 정보가 전혀 알려져 있지 않은 “black system”이고, 세 번째는 전체 시스템에 대한 정보가 불분명하고 불완전한 정보(incomplete information)를 갖는 “grey system”이다. 그레이 시스템의 분야를 나누는데 있어서 저자마다 약간 차이가 있으나 Wen(2004)의 구분에 의하면 그레이 생성(grey generating), 그레이 관계 분석(grey relational analysis), 그레이 모형(grey model), 그레이 예측(grey prediction), 그레이 의사 결정(grey decision making), 그리고 그레이 제어(grey control) 등 여섯 가지로 나누고 있다.

본 연구에서는 여섯 가지의 세부 분야 중에서 그레이 생성과 그레이 관계 분석에 대해 자세히 알아보고 어떻게 사용하였는지를 알아볼 것이다.

### 2.1 그레이 생성(grey generating)

그레이 시스템 이론의 가장 첫 번째 분야인 그레이 생성은 통계학의 관점에서 볼 때, 변수변환과 보간법 등 크게 두 가지로 나눌 수 있다.

### 2.1.1 변수변환

$x_i(l)$가 $i(i=0,1,\cdots,n)$번 째 케이스(sequence, case)의 $l(l=1,2,\cdots,m)$번 째 변수(item, variable)의 값이라고 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 하는데, 어떤 값을 목표로 하느냐에 따라 <표 1>과 같이 다르게 사용된다.

<표 1> 변수 변환 방법 비교

| | Hsia's method | Chang's method |
|---|---|---|
| Larger the better | $z_i(l) = \dfrac{x_i(l) - \min\limits_{all\ i} x_i(l)}{\max\limits_{all\ i} x_i(l) - \min\limits_{all\ i} x_i(l)}$ | $z_i(l) = \dfrac{x_i(l)}{\max\limits_{all\ i} x_i(l)}$ |
| Smaller the better | $z_i(l) = \dfrac{\max\limits_{all\ i} x_i(l) - x_i(l)}{\max\limits_{all\ i} x_i(l) - \min\limits_{all\ i} x_i(l)}$ | $z_i(l) = \dfrac{-x_i(l)}{\min\limits_{all\ i} x_i(l)} + 2$ |
| where | $z_i(l)$ : 생성된 자료, $\min\limits_{all\ i} x_i(l)$ : 최소값, $\max\limits_{all\ i} x_i(l)$ : 최대값 | |

### 2.1.2 보간법(Interpolation)

전술했듯이 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(Wen, 2004)이 있지만 여기에서는 생략하겠다. 이 방법을 이용하여 구한 결과들은 일반적으로 구간값을 갖게 되는데 그 형태가 $[x, \infty), (\infty, x]$, 또는 $(\infty, \infty)$와 같은 경우에 그 값을 제대로 잘 활용할 수 없게 된다.

## 2.2 그레이 관계 분석(Grey Relational Analysis)

### 2.2.1 그레이 관계분석의 개념

그레이 관계 분석(grey relational analysis)은 예측(prediction)과 더불어 그레이 시스템 이론 중에서 연구(research) 분야의 가장 중요한 부분이다. 그리고 그레이 관계 분석의 가장 중요한 역할은 두 개의 서로 다른 케이스들 사이의 관계를 측정한다는 것이다. 이 때 사용되는 측정도구가 바로 그레이 관계 등급(grey relational grade) 또는 그레이 관계 계수(grey relational coefficient)이다.

$P(X)$를 한 가지 주제(theme)라고 하고, $Q$를 하나의 관계라고 가정할 때, $\{P(X);Q\}$를 그레이 관계 등급에서 인자 공간(factor space)이라고 한다.

이 때, $x_i(l) = (x_i(1), x_i(2), \cdots, x_i(m))$, $i(i = 0, 1, 2, \cdots, n)$, $l(l = 1, 2, \cdots, m)$이고 다음과 같은 세 가지 조건을 만족하면 'comparability'를 갖는다고 말한다.

(1) Non-dimensional : 인자들은 차원을 갖지 않는 방향으로 진행된다
(2) Scaling : 각 케이스 $x_i$의 값 $x_i(l)$은 같은 order에 속한다.
(3) Polarization : 인자 공간에서 각 케이스의 배열은 같은 방향에 존재해야 한다.

위의 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.

### 2.2.2 LGRG(Localization Grey Relational Grade)

LGRG는 기준 케이스(reference case)가 따로 존재할 때, 그 기준 케이스와 다른 케이스 사이의 관계를 의미한다. 기준 케이스는 각 변수(item)의 기준값 또는 목표값(target value)이 속해있는 케이스를 의미하는 것으로서 일반적으로 첫 번째 행에 위치한다. 이에 해당하는 대표적인 GRG(그레이 관계 등급)들로는 Deng의 GRG(1989)와 Wen의 GRG(2004) 등이 있다. Deng이 제안한 GRG를 구하기 위해 먼저 GRC(그레이

관계 계수)를 구하면 다음과 같다.

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

여기에서 $j = 1, 2, \cdots, n,\quad l = 1, 2, \cdots, m,$ $x_0$는 기준 케이스, $x_j$는 대상 케이스(inspected case)이고, $\Delta_{\min} = \min_{\forall j} \min_{\forall l} |x_0(l) - x_j(l)|$, $\Delta_{\max} = \max_{\forall j} \max_{\forall l} |x_0(l) - x_j(l)|$ 이며, $\Delta_{0j} = |x_0(l) - x_j(l)|$ 로써 $x_0$와 $x_j$사이의 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 $x_0$와 $x_j$가 서로 비슷한 값을 가지면 1에 가까운 값을 갖게 되고, $x_0$와 $x_j$가 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 $j = 1, 2, \cdots, n,\quad l = 1, 2, \cdots, m,$ $\overline{\Delta}_{0j} = \frac{1}{m} \sum_{l=1}^{m} [\Delta_{0j}(l)]$이다.

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

그리고 Wen의 GRG는 0.5와 1사이에 존재하고, 그 값의 의미는 전체 케이스 중에서 기준 케이스와 조사하는 케이스 사이의 관계 정도이다. 따라서 이 값이 1에 가까우면 가까울수록 조사한 케이스가 기준 케이스와 유사하다는 것을 의미한다.

그 외에 다른 GRG를 구하는 방법들로는 GGRG(globalization grey relational grade), Fuzzy를 이용한 GRG, Entropy를 이용한 GRG(1998) 등이 있다. 먼저 GGRG는 케이스 내에 기준 케이스가 따로 존재하지 않을 때 케이스들 사이의 관계를 알아보고자 할 때 사용하는 방법이다. 따라서 각 케이스가 모두 기준 케이스가 될 수 있다. Fuzzy를 이용한 GRG는 퍼지이론의 소속함수를 이용하여 GRG를 구하는 방법이고, Entropy를 이용한 GRG는 GRC에 Entropy를 이용한 가중치를 부여하는 방법을 말하는 것이다.

## 3. A Modified Grey-Based Nearest Neighbor Approach

우리는 실생활에서 불완전하거나 또는 불분명한 정보시스템을 갖게 된다. 정보의 습득과정에서 정확한 자료의 습득이 불가능하거나 또는 자료의 처리과정에서의 실수로 인해 불분명 또는 부정확한 정보를 갖게 되고 이에 따라 부정확한 자료를 처리해야만 하는 경우가 발생한다. 부정확한 자료가 포함된 케이스 자체를 제거하고 자료를 활용하기도 하지만 때에 따라서는 그 케이스가 전체 정보시스템에서 상당히 중요한 것이어서 제거를 하는 것이 올바른 방법이 아닐 수도 있다. 따라서 해당 케이스를 제거하지 않고 그 자료를 이용함으로써 오히려 성능이 더 뛰어난 결과를 보여줄 수도 있다. 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

전체적인 흐름을 수식과 함께 살펴보도록 하자. 먼저 원래 자료의 구성은 다음과 같다.

$$
\begin{aligned}
x_0 &= (\, x_0(1),\ x_0(2),\ \cdots,\ x_0(m)) \\
x_1 &= (\, x_1(1),\ x_1(2),\ \cdots,\ x_1(m)) \\
&\vdots \\
x_n &= (\, x_n(1),\ x_n(2),\ \cdots,\ x_n(m))
\end{aligned}
$$

GRG를 구하기 위하여 일반적으로 원래 자료를 전처리(preprocessing) 과정을 통해 변환하게 되는데 본 연구에서는 Huang 등에서 사용했던 것과 똑같이 Hsia의 방법을 사용하였으며 $z_i(i=0,1,2,\cdots,n)$ 형태로 다음과 같이 나타내었다.

$$\begin{matrix} z_0(1) & z_0(2) & \cdots & z_0(l) & \cdots & z_0(m) \\ z_1(1) & z_1(2) & \cdots & z_1(l) & \cdots & z_1(m) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ z_n(1) & z_n(2) & \cdots & z_n(l) & \cdots & z_n(m) \end{matrix}$$

첫 번째 행은 기준 케이스를 나타내는 것이다. 이 때, 만일 첫 번째 케이스의 l번째 변수에서 결측값이 발생한다면 다음과 같은 형태가 된다.

$$\begin{matrix} z_0(1) & z_0(2) & \cdots & z_0^*(l) & \cdots & z_0(m) \\ z_1(1) & z_1(2) & \cdots & z_1(l) & \cdots & z_1(m) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ z_n(1) & z_n(2) & \cdots & z_n(l) & \cdots & z_n(m) \end{matrix}$$

이 때 $z_i^*(l)$이 바로 결측값을 나타내는 것이다. 이제 GRG를 구하기 위하여 결측값이 발생한 목록을 제거한 다음과 같은 완전한 자료를 이용한다.

$$\begin{matrix} z_0(1) & z_0(2) & \cdots & z_0(l-1) & z_0(l+1) & \cdots & z_0(m) \\ z_1(1) & z_1(2) & \cdots & z_1(l-1) & z_1(l+1) & \cdots & z_1(m) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ z_n(1) & z_n(2) & \cdots & z_n(l-1) & z_n(l+1) & \cdots & z_n(m) \end{matrix}$$

위의 자료를 이용하여 i번 째 케이스와 j번 째로 높은 그레이 관계 등급인 GRG를 구하게 되는데, 대부분의 경우에 첫 번째 행을 기준 케이스로 하여 각각의 케이스들이 기준 케이스와 얼마나 비슷한지를 알아보게 되는데 본 연구에서는 Wen의 GRG를 구하는 방법을 이용하였다. 이 때, Deng의 GRG를 구하는 과정에서는 변수 개수만큼의 GRC를 구한 후에 GRC의 평균을 구하여 GRG를 계산하는 두 단계의 과정이 있다. 그러나 Wen의 GRG를 구하는 과정은 기준 케이스와 대상 케이스 사이의 평균 거리를 이용하여 GRG를 구하게 되므로 한 단계가 생략된 형태로 값을 구할 수 있게 된다.

$$\Gamma_0^j(j=1,2,\cdots,n)$$

이 때, j는 순위를 나타내는 것이다. Huang 등의 연구에서는 GRG를 근접 이웃의 개수를 구하는 데에만 사용했으나 본 연구에서 구해진 GRG는 근접 이웃의 개수인 k를 정하는 데 뿐만 아니라 다음과 같이 결측값을 대체하는데 있어서 평균에 대한 가중값으로도 사용한다.

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

이 때, $z_i^k(l)(i=1,2,\cdots,n,\ k=1,2,\cdots,n)$은 $\Gamma_0^k$를 갖는 자료의 값을 의미한다. 따라서 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
3. 근접 이웃(nearest neighbor)의 수 k 결정
4. 결측값 대체(imputation)
   1) 산술평균 이용
   2) 가중평균 이용

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 같음을 알 수 있다.

## 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.

<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.

<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 |

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. 방법들간의 성능 비교를 위하여 다음과 같은 RMSE(Root Mean Square Error)를 사용하였다.

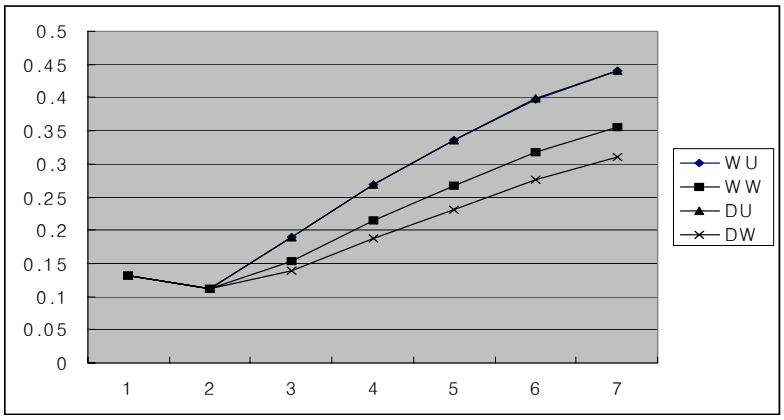
$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(e_i-\widehat{e_i})^2},$$

여기에서 $e_i$는 원래 자료값이고, $\widehat{e_i}$는 추정값을 의미한다.

<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)




<그림 1> 근접 이웃의 개수($k$)에 따른 RMSE 변화

<표 3>과 <그림 1>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 있다. 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.

이번에는 예제 자료에서 가장 성능이 좋은 것으로 나타난 DW 방법을 기존에 사용되던 평균 대체법, 회귀 대체법과 성능을 비교하기 위하여 RMSE를 구하여 비교하였다.

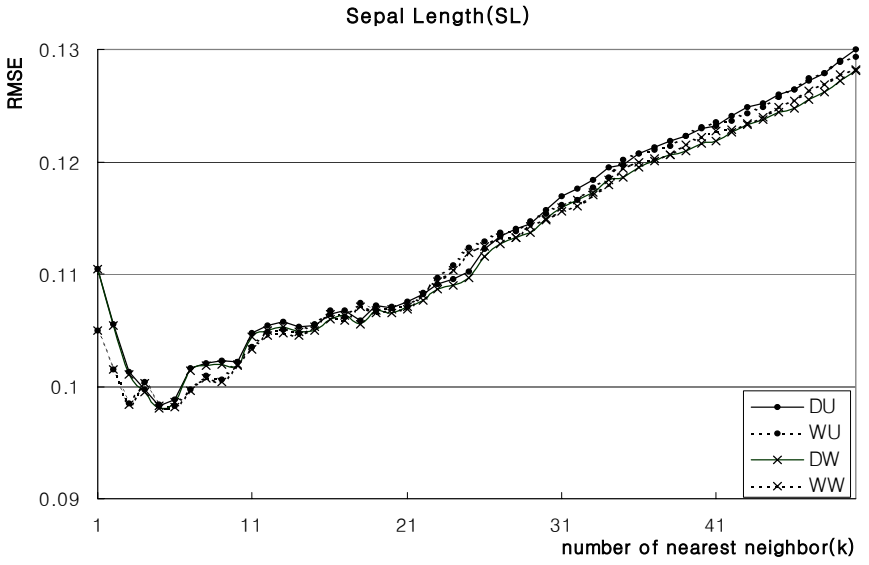
<표 4> DW 방법과 기존 방법들의 RMSE 비교

| 방법 | 평균 대체법 | 회귀 대체법 | DW(k=1) | DW(k=2) |
|---|---|---|---|---|
| RMSE | 0.385746 | 0.226591 | 0.131053 | 0.111515 |

<표 4>에서 보는 바와 같이 근접 이웃의 수를 하나로만 제한하였을 경우에도 평균 대체법과 회귀대체법보다 성능이 좋은 것으로 나타났으며, 근접 이웃의 수를 2로 하였을 경우에는 훨씬 좋은 성능을 갖는 것으로 나타났다.

예제 데이터의 케이스의 수가 적기 때문에 Deng의 방법을 통한 GRG와 Wen의 방법을 통한 GRG의 순위가 동일하게 나타나는 경우가 있고 통계적 일치성(consistency)을 충분히 지지하기 어렵다고 판단하여 붓꽃 데이터를 이용하여 실험을 실시하였다. 붓꽃 데이터는 SL(sepal length), SW(sepal width), PL(petal length), PW(petal width) 등의 네 개의 변수로 이루어져 있고, 150개의 케이스로 이루어져 있다. 실험 방법은 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.

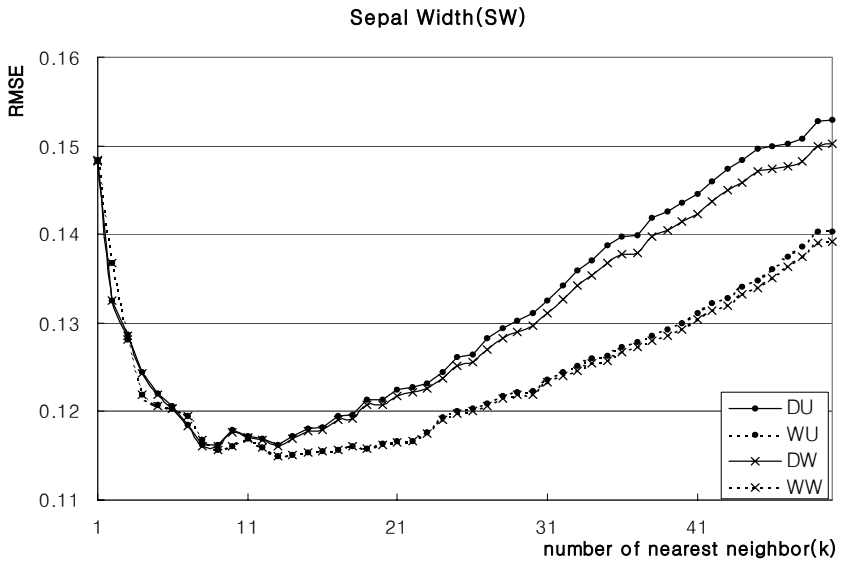
또한 근접 이웃의 개수를 1에서부터 50까지 변화를 주며 비교하였다.




<그림 2> 붓꽃 자료의 SL 변수에 대한 실험 결과

<그림 2>는 sepal length의 경우에 근접이웃의 개수에 따라 RMSE의 변화를 비교한 것이다. 그림에서 보는 바와 같이 근접 이웃의 개수가 적을 경우에는 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. 그리고 sepal length의 경우에는 근접이웃의 개수를 6으로 하였을 경우에 가장 작은 RMSE 값인 0.0981을 갖는 것으로 나타났다.

본 연구에서 실험한 결과값과 Huang 등의 결과값이 서로 다른 이유는 GRG를 구하는 과정에서 같은 순위에 속하는 서로 다른 케이스의 처리 때문인데, 본 연구에서는 동일 순위가 발생하였을 경우에 랜덤하게 결정하는 방법을 사용하였다.

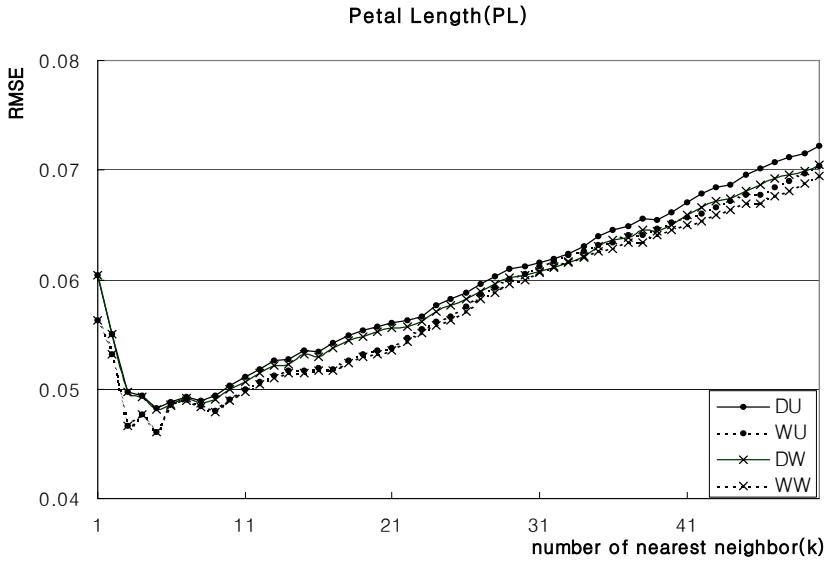



<그림 3> 붓꽃 자료의 SW 변수에 대한 실험 결과

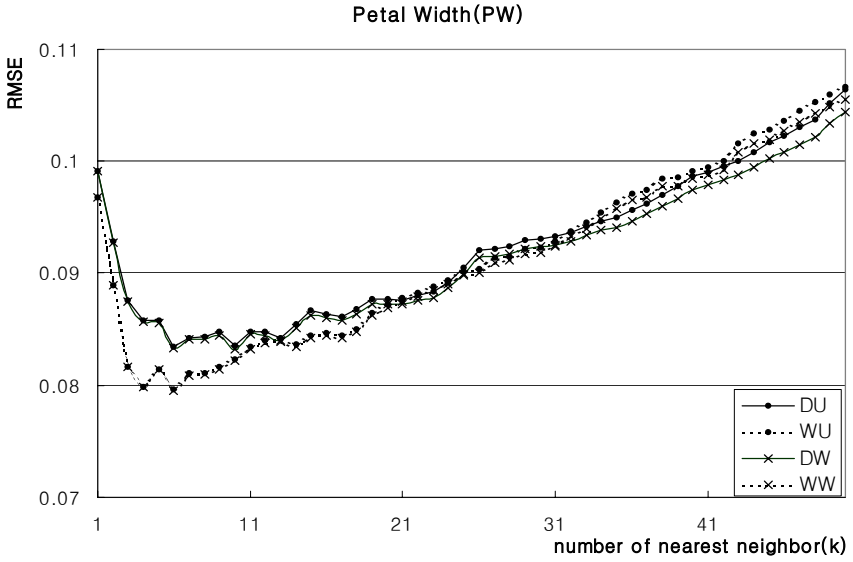
<그림 3>은 sepal width의 경우에 근접이웃의 개수에 따라 RMSE의 변화를 비교한 것이다. 그림에서 보는 바와 같이 근접 이웃의 개수가 적을 경우에는 네 가지 방법들이 서로 비슷한 성능을 갖는 것으로 보이나 전체적으로 Wen의 방법을 이용하여 GRG를 계산한 WU와 WW 방법이 Deng의 방법을 이용하여 GRG를 계산한 DU와 DW 방법보다 작은 RMSE를 갖는 것으로 나타났다. 또한 같은 GRG 계산 방법을 사용한 경우에는 sepal length의 경우와 마찬가지로 가중평균을 사용했을 경우에 가중평균을 사용하지 않은 경우보다 성능이 약간 좋은 것으로 나타났다. 그리고 sepal width의 경우에는 근접이웃의 개수를 13으로 하였을 경우에 가장 작은 RMSE 값인 0.1150을 갖는 것으로 나타났다.

<그림 4>는 petal length의 경우에 근접이웃의 개수에 따라 RMSE의 변화를 비교한 것이다. 그림에서 보는 바와 같이 근접 이웃의 개수와 상관없이 전체적으로 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 앞의 경우와 마찬가지로 가중평균을 사용했을 경우에 가중평균을 사용하지 않은 경우보다 성능이 약간 좋은 것으로 나타났다. 그리고 petal length의 경우에는 근접이웃의 개수를 5로 하였을 경우에 가장 작은 RMSE 값인 0.0461을 갖는 것으로 나타났다.

<그림 5>는 petal width의 경우에 근접이웃의 개수에 따라 RMSE의 변화를 비교한 것이다. 그림에서 보는 바와 같이 근접 이웃의 개수가 적을 경우에는 Wen의 방법을 이용하여 GRG를 계산한 WU와 WW 방법이 Deng의 방법을 이용하여 GRG를 계산한 DU와 DW 방법보다 작은 RMSE를 갖는 것으로 나타났으며 근접이웃의 개수가 많은 경우에는 DU와 DW 방법이 WU와 WW 방법보다 작은 RMSE를 갖는 것으로 나타났다. 하지만 이 경우에도 같은 GRG 계산 방법을 사용한 경우에는 앞의 경우와 마찬가지로 가중평균을 사용했을 경우에 가중평균을 사용하지 않은 경우보다 성능이 약간 좋은 것으로 나타났다. 그리고 petal width의 경우에는 근접이웃의 개수를 6으로 하였을 경우에 가장 작은 RMSE 값인 0.0796을 갖는 것으로 나타났다.




<그림 4> 붓꽃 자료의 PL 변수에 대한 실험 결과




<그림 5> 붓꽃 자료의 PW 변수에 대한 실험 결과

## 5. 결론 및 향후 연구 방향

그레이 시스템 이론은 불확실한 정보가 있을 경우에 자료를 다루는 효과적인 방법이다. 퍼지와 러프셀과 같은 방법론들과 접근 방법에 차이가 존재하지만 전체적인 개념은 비슷한 부분이 많이 있다. 따라서 세 가지 방법은 여러 가지 방법으로 장점들을 접목하는 시도들이 다양하게 이루어지고 있다. 본 연구에서는 그레이 시스템 이론 중에서 그레이 관계 분석 부분에 관심을 가지고 결측값을 다루어 보았다. 특히 Deng의 방법을 이용하는 경우보다 Wen의 방법을 이용하는 경우에 계산 속도도 빠르고 RMSE도 적은 것으로 나타났다. 따라서 이미 개발된 다양한 종류의 GRG 방법들을 이용한 연구도 필요할 것으로 보인다. 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.

추후 연구에서는 자료의 표준화 방법을 통한 자료의 전처리 방법을 사용할 것이고, 여러 가지 다양한 GRG 방법들을 이용하여 근접이웃의 개수 뿐만 아니라 가중값을 결정하는 방법을 사용할 것이다. 또한 퍼지의 alpha-cut과 같은 개념을 도입하여 그레이 관계 등급의 하한값을 제한하는 방법도 좋은 방법이 될 것이라고 생각한다.

## 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. Cohen, M. P. (1996). A new approach to imputation, *American Statistical Association Proceedings of the Section on Survey Research Methods*, 293-298.
4. Cox, B. G. (1980). The weighted sequential hot deck imputation procedure, *Proceedings of the American Statistical Association Section on Survey Research Methods*, 721-726.
5. Deng J. (1982). Control problems of grey systems, *Systems and Control Letters*, 5, 288-294.
6. Deng J. (1989). The basic couse of grey system theory, *HUST Publisher*.
7. 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.
8. Hu, M. X. (2001). A Study of Imputation Algorithm, *NCES Working paper*.
9. Little, R. J. A. and Rubin, D .B. (2002). Statistical Analysis with Missing Data, Second Edition, New York : *Wiley*.
10. Pawlak, Z. (1982). Rough Sets, *International Journal of Computer and Information Sciences*, 11 (1), 341-356.
11. Rubin, D. B. (1976). Inference and Missing Data, *Biometrika*, 63, 581-592.
12. Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys, New York : *Wiley*.
13. Scheffer, J. (2002). Dealing with missing data, *Research Letters Information Mathematical Sciences,* 3, 153-160.
14. Wen, K.-L. (2004). Grey systems : modeling and prediction, *Yang's Scientific Press*.
15. Wen, K.-L., Chang, T.-Ch., You, M.-L. (1998). Grey entropy and its application in weighting analysis, *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, 2, 1842-1844.
16. 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.
17. Zhang, Q. and Chen, G. (2004). Rough grey sets, *Kybernetes*, 33, 2, 446-452.

[ 2006년 4월 접수, 2006년 5월 채택 ]