모델트리의 결측치 처리 방법에 따른 콜레스테롤수치 예측의 성능 변화

정 용 규* · 원 재 강** · 신 성 철***

목 차

요약
1. 서론
2. 관련연구
2.1 결측치(missing value)
2.2 모델트리(Model Tree)
3. 실험 데이터
4. 실험 결과
5. 결론
참고문헌
Abstract

요약

데이터 마이닝은 특정분야에서만 관심을 갖는 분야가 아니라 현재 우리주변 여러 분야에서 많이 사용되고 응용되고 있다. 즉, 수많은 데이터 가운데 숨겨져 있는 유용한 상관관계를 발견하여, 미래에 실행 가능한 정보를 예측하여 추출해 내고 추후에 의사 결정에 이용하는 과정을 말한다. 하지만, 일부 데이터 집합에서는 매우 많은 결측치를 포함하는 변수들이 존재한다. 다시 말해서 다수의 레코드에서 측정치가 존재하지 않는 데이터 집합이 존재한다. 그래서 본 논문에서는 Cholesterol 값을 예측하기 위한 결측치 처리에 따른 모델트리 알고리즘을 적용하고, 실험을 통해서 각 처리방식에 대한 성능을 분석한다. 또는 이 결과를 통하여 결측치 대체방법에 대한 효율적인 적용사례를 제시한다.

표제어: 결측치, 모델트리, 회귀트리, 콜레스테롤

접수일(2012년 8월 30일), 수정일(1차: 2012년 9월 5일), 게재확정일(2012년 9월 20일)
* 을지대학교, 의료IT마케팅학과 교수, ygjung@eulji.ac.kr
** 경기대학교, 컴퓨터과학과 강사, 교신저자, 06240604@hanmail.net
*** 한국후지쯔(주) 헬스케어솔루션부 부장, scsihn@kr.fujitsu.com

1. 서론

데이터 마이닝은 "대량의 데이터에서 새롭고 유용한 지식을 창출하는 것"으로서, 데이터 더미에서 일반적인 사실을 의미하는 데이터가 아니라 의사 결정에 도움이 되는 유용한 정보를 포함한 지식을 추출하는 것이므로 '데이터 캐기(Data mining)'란 용어는 잘못된 용어라 볼 수 있다. 그런 관점에서는 '데이터에서 지식을 캐기(knowledge discovery from data: KDD)'라는 용어가 흔히 사용된다. 데이터 마이닝은 데이터처리, 데이터 요약, 기계학습, 패턴인식, 시각화 기술, 통계학, 지식추출 기술 등 다양한 분야의 다학제적 기술을 필요로 한다[2].

한편, 특별한 경우를 제외하고는 항상 결측치(missing value)가 존재하고, 결측치의 적절한 값으로 채워 넣는 결측치 대체(imputation)에 대한 연구가 활발히 진행되고 있다. 일반적으로 결측치가 있으면 해당되는 관측치(observation, record)는 제거하고 분석을 수행한다. 하지만 결측치를 대체해야만 하는 경우에는 여러 가지 기법들을 사용한다[4]. 전역상수(global constant), 속성(attribute) 평균, 최빈값 등은 가장 기본적으로 고려되는 결측치 대체 방법들이다.

본 논문에서는 콜레스테롤수치 예측을 위한 데이터를 사용하고, 결측치 처리방법으로 결측치가 있는 인스턴스를 삭제하거나 특정값으로 대체하여 실험 분석한다.

2. 관련연구

2.1 결측치(missing value)

일반적으로 일부 레코드들은 결측치(missing values)를 포함한다. 결측치를 갖는 레코드의 수가 적다면 그 레코드는 제외될 수 있다. 그러나 변수의 수가 많은 경우 결측치의 비율이 적다하더라도 많은 레코드에 영향을 미칠 수 있다. 단지 30개의 변수들에 대해서 만약 그 변수값 중 5%가 결측치라고 한다면 (그 결측치는 사례와 변수들 사이에서 무작위적이며 독립적으로 퍼져 있다면), 거의 80%의 레코드들은 분석대상에서 제외되어야 할 것이다(주어진 레코드에서 결측치를 갖지 않을 가능성은 $0.95^{30} = 0.215$).

결측치를 갖는 레코드를 처리하는 하나의 대안은 변수의 결측치를 다른 레코드의 값들을 토대로 계산된 대체값으로 교체하는 것이다. 예를 들어 30개의 변수들 중 가구소득이 특정 레코드에서 결측되어 있다면, 전체 레코드의 평균 소득금액으로 대체될 수 있다. 물론 이렇게 한다고 해서 가구소득이 성과변수에 얼마나 영향을 미치는지에 대한 정보가 추가되는 것은 아니다. 대체값을 교체하는 것은 단지 분석을 계속 수행하게 하는 역할을 하며, 나머지 29개 변수의 해당 레코드에 포함된 정보를 사용할 수 있게 한다. 이러한 기법을 사용하면 데이터 집합의 변동성은 상대적으로 낮게 평가된다는 점에 유의해야 한다. 그러나 평가용 데이터를 이용하여 데이터마이닝 기법의 변동성과 성과를 평가할 수 있으므로 이러한 기법이 심각한 문제를 일으킨다고 보기는 어렵다.

일부 데이터 집합에서는 매우 많은 결측치를 포함하는 변수들이 존재한다. 다시 말해서 다수의 레코드에서 측정치가 존재하지 않는 데이터 집합이 존재한다. 이런 경우에 결측치를 갖는 레코드들을 제외한다면 데이터의 손실이 매우 클 것이다. 또한 결측치를 대체값으로 대체하는 방법은 적은 수의 기존 레코드들을 토대로 계산될 경우 결측치 대체에 따른 유용성이 낮아지게 된다. 한 가지 대안은 예측변수의 중요성을 조사하는 것이다. 예측변수가 매우 중요하지 않다면 분석에서 제외될 수 있다. 만약 그 변수가 중요하다면 소량의 결측치를 갖는 대리변수(proxy variable)가 대신 사용될 수 있다. 물론 예측변수가 중요한 변수라고 생각될 때의 가장 좋은 결측치 처리방법은 결측된 데이터값을 찾는 것이다. 이러한 결측치를 처리하기 위한 방법은 다양하게 연구되었는데 기본적으로 다음과 같은 방법들이 있다[3, 6].

① 해당 데이터 개체 또는 속성의 제거: 결측치가

발생한 데이터 개체를 분석 과정에서 제거하거나 해당 속성을 제거하는 것으로 데이터가 충분히 많이 있다면 고려할만한 방법이다. 하지만 데이터 내에 결측치를 가진 데이터나 속성이 많은 경우 대부분의 정보를 제거하게 될 수 있어 실제로는 많이 사용하지 않는 방법이다.

② 결측치의 추정: 일반적으로 많이 사용되는 방법으로 결측치가 발생한 데이터와 유사한 데이터를 사용하여 결측치를 추정하는 방법이다. 이는 결측치를 추정하는 방법에 따라 다양한 형태가 존재하는데, 예를 들어 A개체의 3번째 속성에서 결측치가 발생한 경우, 결측치가 발생하지 않은 다른 속성을 이용하여 다수의 유사한 개체를 선택하고 이들이 가진 3번째 속성 값을 평균하여 A개체의 결측치를 추정할 수 있다.

③ 결측치의 무시: 알고리즘이나 응용에 따라서는 결측치가 발생한 속성을 무시하고 분석을 수행할 수도 있다. 예를 들어 개체들 사이의 유사성 계산에 있어 많은 수의 속성이 있는 경우 이 중 하나의 속성이 없다면 이를 제외하고 유사성을 계산할 수 있도록 알고리즘을 조정하는 것이다. 하나의 속성 값이 없더라도 유사성을 계산하는데 미치는 영향이 크지 않다면 이러한 방법도 적용 가능하다. 하지만 속성이 몇 개 없어 하나의 속성이라도 무시하기 힘든 경우라면 이러한 방법의 적용은 좋지 않다.

2.2 모델트리(Model Tree)

회귀 트리(Regression trees)는 통계적 회귀 분석의 또 다른 방법이다. 회귀트리라는 이름은 통계학자들이 수치 출력 값을 예측하는 모델의 회귀 모델이라고 말하는 것에서 그 이름이 나왔다. 본질적으로 회귀 트리는 결정 트리(decision tree)이며 다만 트리의 터미널 노드(leaf node)의 값이 범주적이 아닌 수치 값이라는 것이다. 각 터미널 노드의 값은 그 터미널 노드를 통해 전달되는 모든 인스턴스에 대하여 출력 어트리뷰트의 수치 평균 값을 취함으로써 계산되어진다[6].

회귀 트리는 데이터가 비선형일 때 선형 회귀 방정식보다 더 정확하다. 하지만, 회귀 트리는 아주 해석하기가 귀찮고 매우 어려울 수 있다. 이러한 이유로 회귀 트리는 때로는 선형 회귀와 함께 합쳐져서 모델트리(model tree)라는 것을 이룬다. 모델트리에서는 부분 회귀 트리의 각 터미널 노드는 평균 어트류리뷰트 값보다는 선형 회귀 방정식을 나타낸다. 선형 회귀를 회귀 트리와 합치면서 정확한 결과를 얻는 데 필요한 트리 레벨의 수가 적어지므로 회귀 트리 구조도 더 간단해질 수 있다.

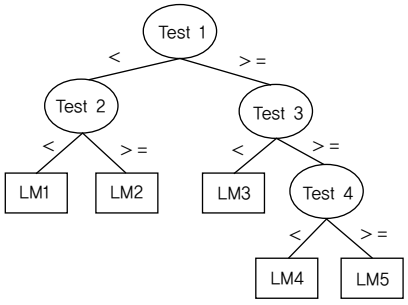


그림 1. 일반적인 모델트리 구조
Fig. 1. General Organization or Model Tree

그림 1은 기본적인 모델트리 구조를 보여준다. 실수 어트리뷰트에 대한 4개의 test가 있고 5개의 터미널 노드가 보여 지고 있는데 각 터미널 노드는 각 선형 회귀 방정식을 나타낸다. 모델트리의 복잡도는 독립 변수와 종속 변수 간에 보여 지는 선형도(degree of linearity)에 따른다[5].

모델트리 기반의 다양한 알고리즘이 제안되고 있으며, 주된 기법으로 M5, RERIS, M5', RegTree 및 HTL 등이 있다. 일반적으로 예측문제에서는 연속적인 입력 변수 및 출력 값을 갖는 데이터들이 대부분을 차지한다. 모델트리는 터미널 노드에 속한 출력값의 평균값을 계산하는 회귀트리와 달리 연속적인 입력값과

출력값을 이용하여 예측 오차값이 최소화되는 계수값을 계산한 후, 계산된 계수값을 이용하여 출력값을 예측한다. 이러한 모델트리도 회귀트리와 같이 데이터를 반복적으로 분리하여 트리 구조를 생성하는 상-하 추론 모델트리(TIMIT: Top-down Induction of Model Tree) 형식을 갖는다[1].

```
MakeModelTree(instances)
{
SD = sd(instances)
for each k-valued nominal attribute
   convert into k-1 synthetic binary attributes
root = newNode
root.instances = instances
split(root)
prune(root)
printTree(root)
}

split(node)
{
if sizeof(node.instances) < 4 or
   sd(node.instances) < 0.05*SD
   node.type = LEAF
else
   node.type = INTERIOR
    for each attribute
     for all possible split positions of the attribute
        calculate the attribute's SDR
        node.attribute = attribute with maximum SDR
    split(node.left)
    split(node.right)
}

prune(node)
{
if node = INTERIOR then
   prune(node.leftChild)
   prune(node.rightChild)
   node.model = linearRegression(node)
if subtreeError(node) > error(node) then
   node.type = LEAF
}
subtreeError(node)
{
l = node.left; r = node.right
if node = INTERIOR then
   return (sizeof(l.instances)*subtreeError(l)
             + sizeof(r.instances)*subtreeError(r))
             /sizeof(node.instances)
else return error(node)
}
```

그림 2. M5' 알고리즘의 Pseudo-code
Fig. 2. Pseudo-code for M5' Algorithm

그림 2는 모델트리 기반의 알고리즘 중에서 M5'에 대한 의사코드이다. 총 5가지의 함수로 이루어졌으며, 메인 함수인 MakeModelTree()는 split()함수를 이용해서 연속적으로 분할하는 노드에 의해 트리를 만들고 prune()함수를 통해 만들어진 트리의 가지치기를 수행한다.

3. 실험 데이터

본 논문에서는 실험을 위한 도구로써 WEKA v3.6 [7, 8]'을 사용한다. WEKA는 Univesity of Waikato(뉴질랜드)의 제품이며 1997년에 최초로 현대적인 형태로 구현되었다. 이 제품은 GNU GPL(General Public License)을 사용한다. 이 소프트웨어는 JavaTM 언어로 쓰여 졌으며, 데이터 파일과의 소통과 시각적 결과물을 생산하기 위해 GUI를 사용한다. 또한 일반 API가 있기 때문에, 다른 라이브러리와 마찬가지로 고유한 애플리케이션에서 자동화된 서버측 데이터 마이닝 작업과 같이 WEKA를 임베디드할 수 있다. 본 실험에서는 Cholesterol 데이터[9]를 사용하였다. 실험 데이터는 독립변수들을 통해서 콜레스테롤수치를 예측하기 위함이다. Cholesterol 데이터 셋은 총 13개의 애트리뷰트로 구성되어있으며 각 애트리뷰트는 표 1과 같다. 실험에 들어가기에 앞서 데이터의 모든 변수를 파악해야한다. 표 1에 나타나있는 변수 중에서 name의 변수의 모든 인스턴스의 값이 name으로 저장되어 있기 때문에 name변수를 제거하고, group 변수도 환자의 생존가능성을 파악하기위한 의미가 없으므로 group 변수도 실험에서 제외한다. 본 논문에서는 결측치(missing value) 처리에 따른 성능을 비교한다. 첫 번째 방법은 결측치가 있는 인스턴스는 제거를 하고 실행하였고, 두 번째는 결측치를 애트리뷰트의 평균값으로 대체하였고, 세 번째에서는 평균값 대신에 최빈값을 사용하였다.

표 1. Cholesterol 데이터셋 속성내역

Tab. 1. Properties of Cholesterol Dataset

속성	데이터형	내역
age	real	age in years
sex	{0, 1}	sex (1 = male; 0 = female)
cp	{1, 4, 3, 2}	chest pain type
		Value 1: typical angina
		Value 2: atypical angina
		Value 3: non-anginal pain
		Value 4: asymptomatic
trestbps	real	resting blood pressure
fbs	{1, 0}	fasting blood sugar > 120 mg/dl(1 = true; 0 = false)
restecg	{2, 0, 1}	resting electrocardiographic results
		Value 0: normal
		Value 1: having ST-T wave abnormality
		Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
thalach	real	maximum heart rate achieved
exang	{0, 1}	exercise induced angina(1 = yes; 0 = no)
oldpeak	real	ST depression induced by exercise relative to rest
slope	{1, 2, 3}	the slope of the peak exercise ST segment
		Value 1: upsloping
		Value 2: flat
		Value 3: downsloping
ca	real	number of major vessels(0~3) colored by flourosopy
thal	{6, 3, 7}	3 = normal; 6 = fixed defect; 7 = reversable defect
num	real	diagnosis of heart diseas
		Value 0: < 50% diameter narrowing
		Value 1: > 50% diameter narrowing
chol	real	serum cholestoral in mg/dl

4. 실험 결과

본 논문에서는 Cholesterol 데이터 셋을 이용하여 값을 예측하기 위한 결측치 처리에 따른 모델트리 알고리즘을 적용하고, 실험을 통하여 각 처리방식에 대한 성능을 분석하였다.

① 결측치 제거

결측치를 제거할 경우에는 오류율은 줄어들지만, 총 데이터의 303개 중에서 62개의 데이터만을 사용하게 되어 버려지는 데이터가 많게 된다. 즉, 데이터의 손실이 매우 크게 되므로 비효율적인 처리방법이다.

```
=== Summary ===

Correctly Classified Instances          60               98.3607 %
Mean absolute error                      0.032
Root mean squared error                  0.1302
Relative absolute error                  7.8577 %
Root relative squared error             28.959  %
Total Number of Instances               61
```

그림 3. 결측치 제거의 성능 결과

Fig. 3. Performance Test after Deleting the Missing Values

② 결측치 대체(애트리뷰트의 평균값)

모델트리는 구조가 간단하기 때문에 그림 3과 같은 트리 결과를 볼 수 있다. sex라는 애트리뷰트값에 따라 2개의 선형회귀모델 방정식을 갖게 된다. 콜레스테롤 값은 그림 4에서 나타난 것처럼 각 선형회귀 방정식에 따라 적용되는 것을 볼 수 있다.

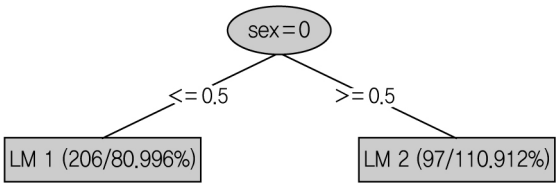


그림 4. Cholesterol 데이터 모델트리
Fig. 4. Data Model Tree of Cholesterol

그림 5는 결측치를 애트리뷰트의 평균값으로 대체해서 처리한 경우의 모델트리의 성능을 보여준다. 예상한 값과 실제 실험 결과가 평균적으로 얼마만큼 떨어졌는가 하는 것과 유사한 것은 47.3253, 오차들의 절대평균값은 36.7682로 나타난 것을 볼 수 있다.

```
M5 pruned model tree:
(using smoothed linear models)

sex=0 <= 0.5 : LM1 (206/80.996%)
sex=0 >  0.5 : LM2 (97/110.912%)

LM num: 1
chol =
          0.0741 * age
          + 1.646 * sex=0
          + 1.0892 * restecg=2,1
          + 0.0158 * thalach
          + 7.4293 * ca
          + 0.9076 * thal=7
          + 226.8278

LM num: 2
chol =
          1.9344 * age
          + 3.2479 * sex=0
          + 27.5845 * restecg=2,1
          + 0.7142 * thalach
          + 39.5547 * thal=7
          + 22.9485
```

그림 5. Cholesterol 모델트리의 결과
Fig. 5. Result of Cholesterol Model Tree

③ 결측치 대체(최빈값)

세 번째 실험은 결측치를 가장 빈번하게 나오는 값을 이용하였다. 그에 대한 실험결과는 그림 6에서 볼 수 있다. 실험결과는 두 번째 애드리뷰트 평균값을 사용한 것과 큰 차이가 없다는 것을 알 수 있다.

```
=== Summary ===

Correlation coefficient                  0.4033
Mean absolute error                     36.7682
Root mean squared error                 47.3153
Relative absolute error                 93.5248 %
Root relative squared error             91.5342 %
Total Number of Instances              303
```

그림 6. 결측치를 애트리뷰트 평균값으로 대체한 성능 결과
Fig. 6. Performance Test after Substituting the Average Value

```
=== Summary ===

Correlation coefficient                  0.4033
Mean absolute error                     36.7682
Root mean squared error                 47.3153
Relative absolute error                 93.5248 %
Root relative squared error             91.5342 %
Total Number of Instances              303
```

그림 7. 결측치를 최빈값으로 대체한 성능 결과
Fig. 7. Performance Test after Substituting the Most Frequency Values

5. 결론

데이터 마이닝을 특정분야에서만 관심을 갖는 분야가 아니라 현재 우리주변 여러 분야에서 많이 사용되고 응용되고 있다. 즉, 수많은 데이터 가운데 숨겨져 있는 유용한 상관관계를 발견하여, 미래에 실행 가능한 정보를 예측하여 추출해 내고 추후에 의사 결정에 이용하는 과정을 말한다. 이러한 데이터에는 다수의 레코드에서 측정치가 존재하지 않는 데이터 집합이 존재한다. 이것을 흔히 결측치(missing value)라고 일컫는다. 이 결측치를 어떻게 대처하느

냐에 따라서 결과예측의 성능이 달라 질 수 있다.

본 논문에서는 Cholesterol 데이터 셋을 이용하여 값 예측하기 위한 결측치 처리에 따른 모델트리 알고리즘을 적용하여 실험하여 각 처리방식에 대한 성능을 분석하였다. 그 결과, 결측치를 무시하면 성능은 좋지만 총 인스턴스 중에서 절반에 해당하는 데이터만 사용하기 때문에 효율적인 방법이 아니었고, 결측치를 최빈값과 애트리뷰트의 평균값으로 대처할 경우, 두 경우의 성능이 비슷하게 나타났다.

참 고 문 헌

〔국내 문헌〕

[1] 박진일, 이대종, 김용삼, 조영임, 전명근 (2008), “상호 노드 정보를 이용한 클러스터 기반 퍼지 모델트리”, 한국지능시스템학회.

[2] 이선미, 박래웅 (2009), “임삼에서의 데이터 마이닝 개념과 원칙”, 대한의료정보학회.

[3] 이승주, 전성해 (2009), “결측치 대체방법에 대한 경험적 비교”, KIIS Spring Conference.

〔국외 문헌〕

[4] Han, J. and Kamber, M. (2001), “Data Mining Conceptand Techniques, Morgan Kaufmann”.

[5] Lan, H. Witten, Eibe Frank, “Data.Mining.Practical.Machine.Learning.Tools.and.Techniques.Second.Editionx”, Morgan.Kaufmann.

[6] PETER, C. BRUCE, GALIT SHMUELI, NITIN R. PATEL (2009), “Data Mining for Business Intelligence: Concepts, Technigues, and App”, l.

〔웹사이트〕

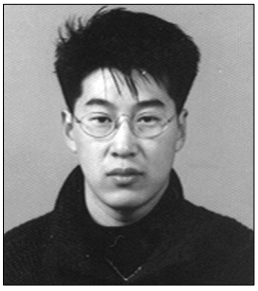
[7] http://www.cs.waikato.ac.nz/~ml/weka/index.html.

[8] http://www.ibm.com/developerworks/kr/library/os-weka1/index.html.

[9] http://archive.ics.uci.edu/ml/.

정 용 규 (Yong Gyu Jung)

서울대학교, 연세대학교, 경기대학교에서 각각 학사, 석사, 박사학위를 취득하였고, 현재 을지대학교 의료IT마케팅학과 교수로 재직 중이다. ISO, UN의 전자문서분야 한국대표위원으로 활동하고 있으며, 의료정보, 전자무역, 해상물류, 금융전산에 Semantic Web, Process Modelling, ebXML 등의 표준기술의 적용에 관심이 많다.



원 재 강 (Jae Kang Won)

강릉대학교, 경기대학교에서 각각 학사, 석사, 박사학위를 취득하였고 현재 을지대학교, 경기대학교에서 강의를 하고 있다. 주요 연구분야로는 워크플로우이며 해저탐사 등의 분야에서 의사결정지원을 위한 다양한 결정요인을 마이닝기법을 통해 실험하고 연구하고 있다.

신 성 철 (Sung Chul Sihn)

한남대학교에서 학사학위를 취득하였고, 현재 연세대학교 보건대학원에 재학 중이며 한국후지쯔(주) 헬스케어솔루션팀장으로 재직 중이다. 헬스케어 분야 IT 솔루션 구축 업무 및 IT 기획 업무를 하고 있으며 헬스케어분야 IT 솔루션 특히 병원정보, 원/내외물류관리, 병원경영에 관심이 많다.

Using Missing Values in the Model Tree to Change Performance for Predict Cholesterol Levels

Yong Gyu Jung* · Jae Kang Won** · Sung Chul Sihn***

ABSTRACT

Data mining is an interest area in all field around us not in any specific areas, which could be used applications in a number of areas heavily. In other words, it is used in the decision-making process, data and correlation analysis in hidden relations, for finding the actionable information and prediction. But some of the data sets contains many missing values in the variables and do not exist a large number of records in the data set. In this paper, missing values are handled in accordance with the model tree algorithm. Cholesterol value is applied for predicting. For the performance analysis, experiments are approached for each treatment. Through this, efficient alternative is presented to apply the missing data.

Keywords: Missing Value, Model Tree, Regression Tree, Cholesterol Levels

* Department of Medical IT and Marketing, Eulji University, ygjung@eulji.ac.kr
** Department of Computer Science, Kyonggi University, Corresponding Author, 06240604@hanmail.net
*** Department of Healthcare Solution, Fujitsu Korea Limited, scsihn@kr.fujitsu.com