데이터 마이닝을 위한 신경망 이용 결측 값 처리 방법

성지애° 류정우 김명원
숭실대학교 컴퓨터 학부
unico96°@korea.com, mkim@computing.ssu.ac.kr

A Method Processing Missing Values for Data Mining based on Artificial Neural Network

Ji-Ae Sung° Joung-Woo Ryu Myoung-Won Kim
School of Computing, Soongsil University

요 약

실세계의 많은 데이터는 결측 값들을 포함하고 있기 때문에 데이터 마이닝 시스템에 완벽한 데이터를 제공하기는 불가능하다. 또한 결측 값이 존재하는 대용량의 데이터를 추천시스템에 적용하여 분석하고자 할 경우, 정확성이 떨어지는 결과를 초래할 수 있다. 따라서 데이터에 결측 값이 존재할 경우 입력 데이터를 사전에 보간하는 전처리 방법이 필요하다. 이러한 기존의 보간 전처리 방법에는 결측 값 속성을 삭제하거나 대치하는 방법이 대표적이나, 삭제 방법은 결측 값이 존재하는 데이터를 제거하는 방법으로 중요 속성 삭제 및 데이터 손실을 유발하는 단점이 있어 일반적으로 결측 값을 다른 값으로 처리하는 대치 방법이 널리 사용된다.

본 논문에서는 전처리 방법 중 결측 값을 처리하는 가장 일반적인 대치 방법과 신경망을 이용한 평가 예측 처리 방법을 소개한다. 또한 신경망을 이용 결측 값을 대치하는 새로운 모델을 제안하고, 각각의 결측 값 처리방법을 비교 분석한다.

1. 서 론

데이터 마이닝(Data Mining)이란 축적된 많은 데이터로부터 유용한 지식을 발견하는 것을 말하며, 최근에는 추천시스템, 마케팅 전략과 같은 넓은 분야에서 다양하게 이용되고 있다. 그러나 실세계의 많은 데이터는 알려지지 않은 결측 값(Missing Values)들을 포함하고 있어 데이터 마이닝 시스템에 완벽한 데이터를 제공하는 것은 거의 불가능하다. 따라서 데이터 마이닝의 여러 기법에 적용하기 위한 입력 데이터에 결측 값이 존재하는 경우 사전에 데이터를 보간(imputation) 하는 전처리 방법이 필수적이다[1].

현재 활용되고 있는 전처리 방법으로는 결측 값을 무시하는 삭제 방법과 결측 값을 다른 값으로 대치하는 방법이 있다[2]. 그러나 삭제 방법의 경우 데이터의 특성을 분석하는데 중요한 속성을 제거할 수 있는 위험 부담 및 데이터 손실의 우려가 있기 때문에 일반적으로 대치 방법이 많이 사용된다. 결측 값을 대치하는 일반적인 방법은 "zero" 대치 방법과 통계적인 방법을 이용한 평균(mean), 중앙(median), 최빈(mode) 값 대치 방법 등이 있다[3][4][5]. 이 두 방법은 특정한 하나의 값으로 결측 값을 대치하기 때문에 데이터 분포에 따라 민감한 결과를 보인다. 따라서 이러한 문제를 해결하기 위해 결측 값을 현재 속성들의 연관성에 의한 평가 예측 값으로 대치하는 신경망 이용 방법이 연구되고 있다. 신경망 이용 방법 중 "estimation" 모델 처리는 결측 값이 존재하는 속성들을 출력 노드로 설정하고 그 외의 속성들을 입력 노드로 설정하여 모델을 설계하고 처리하는 방법이다[3][6][7]. 그러나 "estimation" 모델 처리 방법은 결측 값이 생성될 수 있는 모든 경우에 대한 모델을 생성해야 하는 단점을 가지고 있다.

본 논문에서는 일반적인 "zero" 방법과 평균, 최빈 대치방법 및 신경망 이용 방법인 "estimation" 모델 처리 방법을 소개한다. 또한 신경망을 이용한 새로운 결측 값 처리 모델을 제안하고, "estimation"모델 처리 방법과 비교 분석한다. 마지막으로 결측 값 처리 방법에 의해 대치된 자료의 성능 평가를 위해 앞에서 소개한 "zero", 평균, "estimation" 모델 처리 방법과 본 논문에서 제안한 방법으로 대치한 자료를 분류(classification) 문제에 적용하여 결측 값이 없는 경우와의 분류 성능을 비교한다[4][5].

2. 관련연구

2.1 일반적인 대치 방법

대치 방법이란 어떠한 변경도 없이 단순하게 선택된 값으로 결측 값을 처리하는 방법으로 일반적으로 "zero" 방법과 통계적인 방법을 이용한 평균, 최빈 방법 등이 있다.

가장 단순한 방법으로 결측 값을 대신하여 "0" 값을 선택하여 결측 값을 대치하는 "zero" 방법이 있다. "zero" 방법은 무시하는 삭제 방법과 유사한 방법으로 값이 없는 데이터라고 할 수 있다. 데이터에 대한 사전 조사를 통한 적당한 값으로 처리하는 통계적 방법으로는 평균 방법과 가장 빈도수가 높은 값으로 결측 값을 대치하는 최빈 방법이 대표적이다. 그러나 이 두 가

본 연구는 한국 과학기술부에서 지원하는 뇌신경 정보학 연구사업으로 수행되었음.

지 방법의 경우 입력 데이터의 배치가 서로 상이하다면 좋은 결과를 기대하기 어렵다.

2.2 신경망 이용 대치 방법

신경망 이용 방법이란 결측 값을 기존의 연관된 속성들을 입력으로 하여 신경망 모듈을 거쳐 평가 예측된 보다 정확한 값으로 대치하는 방법이다. 신경망 이용 방법에는 결측 값을 제외한 현재의 알려진 속성만을 가지고 결측 값을 예측 평가하는 "estimation"모델 처리 방법이 있으며, 입력, 은닉, 출력 층으로 구성된 three-layer feed forward network 구조를 가지고 back propagation 알고리즘으로 학습한다.

"estimation" 모델은 속성이 n개일 경우, 입력 노드 수는 결측 값이 없는 속성 개수이고, 출력 노드 수는 결측 값 속성 개수이다. 즉, 결측 값이 없는 속성만을 입력으로 하여 결측 값을 가진 속성을 평가 예측 하는 것이다. 따라서 속성의 조합만큼 처리 모델이 요구되므로 많은 속성을 가진 데이터 처리에는 적절하지 않다. 하지만 입력 노드가 결측 값이 없는 속성으로만 구성되어 있기 때문에 결측 값을 표현하지 않아도 되는 장점을 가지고 있다.

3. 제안한 결측 값 처리 모델

앞에서 기술한 "estimation" 모델 처리 방법의 경우 속성의 조합 개수만큼의 모델이 존재해야 하므로, 많은 속성을 가진 데이터의 결측 값 처리에는 적절하지 않다. 이러한 문제점을 해결하기 위해 본 논문에서는 신경망을 이용한 결측 값 처리 방법으로 많은 속성을 가지고 있는 대용량 데이터의 결측 값 처리를 위한 "N-to-N" 모델을 제안한다.

3.1 결측 값 표현 방법

"estimation"모델의 경우 입력노드가 결측 값이 없는 속성으로만 구성되기 때문에 결측 값을 표현하지 않아도 되는 장점이 있다. 그러나 제안한 방법은 최소한의 결측 값 처리 모델을 생성해야 하기 때문에 결측 값을 포함하는 속성이 입력노드로 구성될 수 있어 결측 값에 대한 표현 방법이 요구된다.

1) 수치데이터 결측 값 표현

제안한 방법에서는 수치 데이터의 결측 값을 "0.0"으로 표현하며 이를 위해 모든 데이터를 각각의 속성별(차원별)로 [0.1..0.9]사이로 정규화 한다.

2) 기호 데이터 결측 값 표현

기호데이터 모델 생성 시, 신경망 입력은 각 속성들을 이진형(boolean)으로 변환하여 사용해야 한다. 따라서 결측 값을 "0"으로 표현하며, "0"은 해당 속성의 속성 값이 없음을 의미한다.

3.2 "N-to-N" 모델

많은 속성을 가진 대용량 데이터에서 결측 값 처리를 위해 제안한 신경망 이용 대치 방법인 "N-to-N" 모델의 구조는 "estimation" 모델과 마찬가지로 입력, 은닉, 출력 층으로 구성된 three-layer feed forward network구조를 가지고 back propagation 알고리즘으로 학습한다. "N-to-N" 모델은 입력노드와 출력노드 개수를 동일하게 설계한다. 즉, 목표 출력 값과 입력 값이 서로 같은 모델이다.

그림1과 같이 속성으로만 구성된 모델을 수치 모델이라 하고, 기호 모델은 하나의 속성이 가질 수 있는 속성 값을 이진형으로 표현하여 그림2와 같이 구성되는 모델이다.

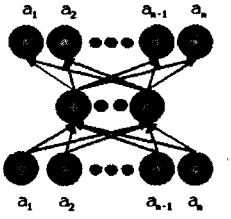

그림 1. 수치 모델

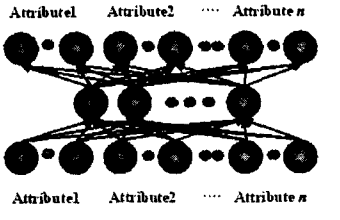

그림 2. 기호 모델

신경망 이용 방법 중 "estimation"모델이 예측하고자 하는 결측 값을 제외하고 평가 예측 하는 반면, 제안한 "N-to-N" 모델은 예측하고자 하는 결측 값 속성도 입력 노드로 선택되어 결측 값이 없는 현재 입력 속성들과 함께 결측 값을 예측하는 모델이다. 따라서 속성별로 모델이 존재하거나 속성의 조합 수만큼의 모델이 생성될 필요가 없다.

4. 실험 및 결과

4.1 실험 데이터

실험에 사용된 수치 데이터인 Iris 데이터[8]는 네 개의 속성이 각각 50개씩 3개의 클래스로 구성되어진 총 150개의 레코드로 구성된 데이터이다. 각 클래스에서 순차적으로 100개의 레코드는 학습 데이터로 50개의 레코드는 테스트 데이터로 구성하고, 결측 값을 "0.0"으로 표현하기 위하여 각각의 입력 데이터를 속성별(차원별)로 범위를 [0.1..0.9]로 정규화 하였다.

신경망 이용 방법의 결측 값 처리 평가를 위한 실험 테스트 데이터는 4개의 입력 속성 중 한 개의 속성만을 결측 값으로 설정하여 고려한 테스트 데이터와 결측 값이 없는 테스트 데이터 총 2개로 구성된다.

분류 문제를 통해 본론에서 서술했던 결측 값 처리 방법의 평가를 위한 테스트 데이터는 결측 값이 처리 대치되어 구성된 전처리 테스트 데이터와 결측 값이 없는 실제 테스트 데이터로 구성된다.

4.2 실험

실험은 신경망 시뮬레이션인 JavaNNS version 1.1을 사용한다[9]. 파라메터는 기존의 여러 테스트를 거쳐 학습률(learning rate)은 0.25로 설정하고 다른 파라메터들은 기존에 설정된 디폴트 값으로 실험한다.

1) 실제 값과 예측 값의 오차에 대한 비교 분석 실험

실제값과 예측 값과의 오차를 분석하기 위해 Iris 데이터를 사용 4개의 입력 속성 중 테스트 데이터에 결측 값이 없을 경우와 하나의 속성만이 결측 값일 경우만을 고려한다. 각 모델에 적용하여 테스트하고 테스트된 데이터의 속성별 평균 절대 오

차(Mean Absolute Error)는 식(1)을 통해 계산된다.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} T_i - R_i \qquad \text{식(1)}$$

T_i : 목표출력값, R_i : 실제출력값

표 1. 신경망 이용 방법의 예측 값과 실제 값 오차

모델	유무	예측 값과 실제 값의 오차							
		속성 1		속성 2		속성 3		속성 4	
		μ	σ	μ	σ	μ	σ	μ	σ
N-to-N (4-3-4)	있음	0.07	0.08	0.04	0.06	0.08	0.07	0.04	0.07
	없음	0.05	0.04	0.06	0.04	0.06	0.04	0.05	0.03
estimation (3-3-1)	있음	0.06	0.05	0.08	0.07	0.06	0.03	0.06	0.04
	없음	0.06	0.05	0.11	0.08	0.06	0.03	0.06	0.04

표1은 실제 값과 예측 값과의 오차 결과를 보여주고 있고, 표1의 결과를 토대로 각 속성별로 통계적 분석 방법인 가설 검정에 의하면 유의수준 5%에서 테스트 데이터에 결측 값이 없을 경우 모델의 예측 값과의 오차 차이가 거의 없음을 확인 할 수 있다.

2) 분류모델 적용 실험

본 실험은 결측 값 처리 방법의 성능 평가를 위해 입력, 은닉, 출력 층으로 구성된 three-layer feedforward network (4-3-3) 구조로 신경망 분류모델을 생성한다. 학습데이터는 결측 값이 없는 정제된 데이터를 사용하고, 앞에서 서술한 결측 값 처리 방법으로 결측 값을 처리한 데이터를 테스트 데이터로 사용한다.

표 2. 하나의 결측 값 처리 후 분류 모델 인식률(Iris)

결측 값 처리 방법	인식률(%)			
	속성1	속성2	속성3	속성4
zero	98.0	73.6	100	85.6
mean	94.0	100	100	87.6
estimation	96.2	100	100	93.6
N-to-1	96.2	100	100	93.6
N-to-N	98.4	100	100	92.4

표 3. 두개의 결측 값 처리 후 분류 모델 인식률(Iris)

결측 값 처리 방법	인식률(%)	
	속성1과 속성2	속성3과 속성4
zero	100	34.0
mean	100	74.0
estimation	99.6	82.4
N-to-1	100	94.0
N-to-N	98.0	84.8

표2와 표3의 결과는 각각의 데이터에 결측 값이 한 개와 두개가 있을 경우 앞에서 제시한 다섯 가지 방법의 결측 값 처리에 의해 예측된 값으로 결측 값을 대치한 자료를 분류모델에 적용한 인식률을 보여주고 있다. 표2에서 보이는 바와 같이 분류할 레코드에 결측 값이 존재 할 경우 다섯 가지 방법에 대한 인식률의 차이가 나타나고 있지 않다. 그러나 표3의 결측 값이 두 개 있을 경우 신경망을 이용한 결측 값 대치 방법이 "zero"나 평균 방법보다 나은 성능을 보이고 있으며, 또한 제안한 두 방법이 "estimation" 모델 처리 방법의 인식률과 유사한 성능을 나타내고 있는 결과로 보아 많은 속성을 갖는 대용량의 데이터에 대해 보다 효율적인 모델임을 알 수 있다.

5. 결론 및 향후 연구

본 논문에서는 실세계 데이터에 포함된 결측 값을 처리하는 전처리 방법을 기술하고 있으며, 신경망을 이용한 결측 값 처리 방법인 "N-to-N" 모델을 제안하고 있다. "N-to-N" 모델은 수치 데이터에 대하여 기존의 대치 방법인 "zero", 평균방법보다 분류 모델에 대해 성능 향상을 나타내고 있음을 확인할 수 있다. 또한 estimation" 모델과는 유사한 성능을 나타내고 있음을 확인 하였다. 따라서 제안하고 있는 모델은 많은 속성을 가진 대용량 데이터에 대해서 기존 "estimation"모델 처리 방법보다 최소의 처리 모델을 생성하기 때문에 보다 효율적인 모델로 사료된다.

향후 연구로는 기호 모델에 대해 벤치마크 실험을 통한 타당성 검증 및 수치적인 데이터와 기호적인 데이터를 동시에 고려하여 결측 값을 포함하고 있는 속성의 개수에 따른 성능을 비교할 것이다.

6. 참고문헌

[1] Oscar Ortega Lobo and Masayuki Numao "Ordered *estimation of missing values*" *1999*

[2] Angela L. Cool "A Review of Methods for Dealing with Missing Data" *2000*

[3] Vamlpew P, Clark D, adams A, Muench J "Techniques for Dealing with Missing Values in *Feedforward* Networks" 1992

[4] W.Z.Liu, A.P.white, S.G.Thompson and M.A.Bramer "Techniques for Dealing with Missing Values in Classification" *Symposium on Intelligence Data Analysis* 1997

[5] Colleen M. Ennett, Monique Frize, C.Robin Walker "Influence of Missing Values on Artificial Neural Network Performance" Proceeding of *Medinfo London*, pp449-453, 2001

[6] Peter Vamplew and Anthony Adams "Missing Values in a Backpropogation Neural Net" Proc. 3rd Australian Conf. on Neural Networks(ACNN92) *pp64-66*, 1992

[7] P.K. Shape & R.J.Solly "Dealing with Missing Values in Neural Network-Based *Diagnostic Systems*" 1995

[8] ftp://ftp.ics.uci.edu/pub/machine-learning-databases

[9] http://www-ra.informatik.uni-tuebingen.de/SNNS/.