K-NN과 최대 우도 추정법을 결합한 소프트웨어 프로젝트 수치 데이터용 결측값 대치법

(A Missing Data Imputation by Combining K Nearest Neighbor with Maximum Likelihood Estimation for Numerical Software Project Data)

이 동 호 † 윤 경 아 † 배 두 환 ††
(Dong-Ho Lee) (Kyung-A Yoon) (Doo-Hwan Bae)

요 약 소프트웨어 프로젝트 데이터를 이용한 각종 분석·예측 모델 생성시 직면하는 문제 중 하나는 데이터에 포함된 결측값이며 이에 대한 효과적인 방안은 결측값 대치법이다. 대표적인 결측값 대치법인 K 최근접 이웃 대치법은 대치과정에서 결측값을 포함하는 인스턴스의 관측정보를 활용하지 못한다는 단점이 있다. 본 연구에서는 이러한 단점을 극복하기 위해 K 최근접 이웃 대치법과 최대 우도 추정법을 결합한 새로운 소프트웨어 프로젝트 수치 데이터용 결측값 대치법을 제안한다. 또한 결측값 대치법의 정확도를 비교하기 위한 새로운 측도를 함께 제안한다.

키워드 : 결측값 대치법, K 최근접 이웃 대치법, 최대 우도 추정법, 소프트웨어 프로젝트 데이터

Abstract Missing data is one of the common problems in building analysis or prediction models using software project data. Missing imputation methods are known to be more effective missing data handling method than deleting methods in small software project data. While K nearest neighbor imputation is a proper missing imputation method in the software project data, it cannot use non-missing information of incomplete project instances. In this paper, we propose an approach to missing data imputation for numerical software project data by combining K nearest neighbor and maximum likelihood estimation; we also extend the average absolute error measure by normalization for accurate evaluation. Our approach overcomes the limitation of K nearest neighbor imputation and outperforms on our real data sets.

Key words : missing data imputation, K-NN, maximum likelihood estimation, software project data

· 본 연구는 지식경제부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었고(IITA-2009-(C1090-0902-0032)), 또한 방위사업청과 국방과학연구소의 지원으로 수행되었습니다.
· 이 논문은 제35회 추계학술대회에서 'K-NN 최대 우도 추정법을 결합한 소프트웨어 프로젝트 수치 데이터용 결측값 대치법에 관한 연구'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : KAIST 전산학과
dhlee@se.kaist.ac.kr
kayoon@se.kaist.ac.kr
†† 종신회원 : KAIST 전산학과 교수
bae@se.kaist.ac.kr

논문접수 : 2008년 12월 18일
심사완료 : 2009년 3월 3일

정보과학회논문지: 소프트웨어 및 응용 제36권 제4호(2009.4)

1. 서 론

소프트웨어 프로젝트 데이터를 이용한 각종 분석·예측모델 생성시 직면하는 문제 중 하나는 데이터에 포함된 결측값(missing data)이다. 결측값이 포함된 데이터는 일반적으로 활용되지 않고 버려지는데, 이는 정보 손실을 유발하여 특정 데이터에 편향된(biased) 모델을 생성할 수 있다. 특히 소프트웨어 프로젝트 데이터는 일반적으로 소규모이기 때문에 상기 방법은 데이터 규모를 더욱 작게 하여 심각한 오류를 초래할 가능성을 더욱 증가시킨다. 이에 대한 대안으로 결측값을 관측값에 기반한 적절한 추정값으로 대치하는 결측값 대치법(missing data imputation)이 연구되었다. 결측값 대치법은 소프트웨어 프로젝트 데이터의 특성을 고려할 때 보다

효과적이라는 것이 많은 연구결과로 입증되고 있다[1-3].

이러한 연구결과 중 소프트웨어 프로젝트 데이터를 위한 결측값 대치법으로 K 최근접 이웃 대치법(K nearest neighbor imputation, 이하 K-NN)과 최대 우도 추정법(maximum likelihood estimation, 이하 MLE)을 추천하는 두 연구가 있었다[1,2]. K-NN은 데이터마이닝 기법으로, 통계적 기법과 달리 데이터 분포에 대한 가정이 불필요한 장점이 있는 반면, 결측값을 포함하는 데이터의 나머지 관측값을 활용할 수 없는 단점이 있다. MLE는 통계적 기법으로, 데이터 규모가 커질수록 정확한 추정이 가능한 반면, 특정 데이터 분포와 결측값이 발생하는 특정 방식에 대한 가정이 필요하다.

본 연구에서는 K-NN과 MLE를 결합한 새로운 소프트웨어 프로젝트 수치 데이터용 결측값 대치법과 이 기법의 정확도를 평가하기 위한 새로운 측도(measure)를 함께 제안한다. 이 결측값 대치법은 모든 결측값을 MLE 결과값으로 대치하고 K-NN을 적용함으로서 결측값을 포함하는 데이터의 모든 관측정보를 활용할 수 있어 보다 정확한 결측값 대치를 가능하게 한다.

본 논문은 다음과 같이 구성된다. 2장에서는 배경 지식으로 결측 메커니즘(missingness mechanism)을 소개하고, 3장에서는 대표적인 결측값 대치법들의 기본 개념과 이들을 적용한 기존 관련연구들에 대해 살펴본다. 4장에서 본 연구에서 제안하는 새로운 결측값 대치법 및 측도를 제안한다. 5장에서는 사례 연구를 통해 새로운 대치 기법과 기존 기법들의 정확도를 제안된 측도를 이용하여 비교한다. 마지막으로, 6장에서 결론 및 향후 연구로 본 논문을 맺도록 한다.

2. 배경 지식: 결측 메커니즘(missingness mechanism)

데이터에 포함된 결측값에 대한 적절한 결측값 기법을 결정하기 위해서는 해당 기법이 가정하는 결측 메커니즘을 이해하는 것이 중요하다. 결측 메커니즘은 결측값과 데이터 변수들과의 연관관계를 의미한다. Little[4]은 변수들에 대한 결측의 의존 여부에 따라 결측 메커니즘을 Missing Completely At Random(MCAR), Missing At Random(MAR), Non Ignorable(NI) 세 가지로 분류하였다.

MCAR은 결측이 변수의 결측값 포함 여부와 상관없이 어떠한 변수들과도 무관한 경우를 의미하는 것으로, 결측이 랜덤하게 분포한 경우를 뜻한다. MAR은 결측이 오직 결측값을 포함하지 않는 변수들과 연관이 있는 경우이며, NI는 결측이 오직 결측값을 포함하는 변수들과 관련이 있는 경우다. 표 1은 두 개의 변수(변수1, 2)로 구성된 6개의 프로젝트 데이터 인스턴스(이하 인스턴스)들을 통해 세 가지 결측 메커니즘을 설명한 예이다. MCAR은 변수 2의 결측이 결측값을 포함하지 않는 변수 1과 결측값을 포함하는 변수 2에 무관함을 알 수 있다. 반면, MAR은 변수 2의 실제값에 관계없이 변수 1이 B일때만 결측이 발생하여 오직 결측값을 포함하지 않는 변수 1의 값에 의존함을 알 수 있다. NI는 변수 1에 관계없이 변수 2의 값이 105 이상인 경우에 결측이 발생하여 결측값이 포함된 변수 2와 관련이 있음을 알 수 있다.

표 1 결측 메커니즘의 예

	변수 1	변수 2			
		실제값	MCAR	MAR	NI
인스턴스1	A	85	?	85	85
인스턴스2	A	105	105	105	?
인스턴스3	B	111	?	?	?
인스턴스4	B	80	80	?	80
인스턴스5	C	130	130	130	?
인스턴스6	C	97	?	97	97

? : 결측값

3. 관련 연구

이 장에서는 대표적인 결측값 대치법들의 기본 개념에 대해 설명하고, 소프트웨어 프로젝트 데이터의 결측값 대치와 관련된 기존 연구들을 살펴본다.

3.1 대표적인 결측값 대치법

다양한 결측값 대치법들 중에서 주로 사용되는 네 가지 대표적인 기법은 다음에 설명되는 평균 대치법, K-NN, MLE 그리고 다중 대치법이다.

- **평균 대치법(mean imputation, 이하 MEI).** MEI은 결측값을 포함하지 않는 완전한 인스턴스들의 산술평균을 결측값에 대치하는 기법으로, 매우 간단하지만 결측값 대치 후 변수의 분산이 편향되는 단점을 가진다.
- **K-NN.** K-NN은 결측값을 포함한 인스턴스와 가장 가까운 거리(예: 유클리디언 거리)를 가지면서 결측값이 없는 K개의 인스턴스들을 이용하여 결측값을 대치하는 방법이다. 결측값 대치시 결측값을 포함하지 않는 완전한 인스턴스들만을 대상으로 하기 때문에 결측값을 포함하는 인스턴스의 관측값들을 활용하지 못한다.
- **MLE.** MLE는 잘 알려진 통계적 추정법으로, 우도 함수(likelihood function)를 최대화시키는 모수(parameter)를 찾는 것을 기본 원칙으로 한다. MLE는 결측 메커니즘 중 MAR을 전제로 하며 다변량 정규 분포(multivariate normal distribution)가 아닌 데이터에 대해서도 강건하지만(robust) 여타 기법에 비해 상대적으로 큰 규모의 데이터가 필요하다는 단점이 있다.
- **다중 대치법(multiple imputation, 이하 MI).** MI

는 하나의 결측값에 대해 하나 이상의 추정값으로 대치하는 방법으로써 Rubin[5]에 의해 제안되었다. MI는 단일 대치법(single imputation)에서 일반적으로 나타나는 분산 추정값의 편향(biased estimates of variance) 현상을 보완하여 결측값을 가진 변수의 분산이 편향되지 않도록 하는 장점이 있다. 그러나, 알고리즘이 통계관련 전문지식을 요구하는 많은 입력 파라메터를 가지고 있어 그 사용이 어렵다는 단점이 있다.

3.2 소프트웨어 프로젝트 데이터의 결측값 대치법에 대한 기존 연구

Strike[1]는 소프트웨어 비용 추정을 위한 데이터에 포함된 결측값을 해결하기 위해 결측값 처리 기법들을 비교하였다. 비교대상으로는 결측값을 포함한 데이터를 삭제하는 기법(listwise deletion, 이하 LD)과 결측값 대치법인 MEI, K-NN이다. 연구결과는 소프트웨어 비용 추정 모델 생성을 위해 결측값 대치법이 LD보다 효과적이며, 비용 추정 모델 생성을 위한 결측값 대치법으로 K-NN을 추천하였다.

Myrtveit[2]는 결측값이 포함된 실제 소프트웨어 프로젝트 데이터를 대상으로 네 가지 결측값 처리 기법들을 비교하였다. 비교대상 기법들은 LD, MEI, 유사 응답 유형 대치 기법(similar response pattern imputation), MLE이었고, 결과로는 데이터 규모가 클 경우 MLE의 사용이 적합하다는 것을 실험결과로 도출하였다.

Cartwright[3]는 소프트웨어 프로젝트 데이터에 대한 결측값 대치법의 유용성을 MEI와 K-NN에 대하여 평가하였다. 연구결과는 결측값 대치법이 유용하며 K-NN이 MEI보다 효율적임을 보였다.

Twala[6]는 소프트웨어 공학 데이터베이스에 포함된 결측값에 대해 다양한 결측값 기법들을 비교하였다. 연구결과는 결측값이 많이 포함된 데이터의 경우 MI를 적용하는 것이 효과적이라고 제안하였다.

Song[7]은 소규모 소프트웨어 프로젝트 데이터의 결측값 대치를 위해 K-NN에 기반한 클래스 평균 대치법(class mean imputation)이라는 새로운 결측값 대치법을 제안하였다. 이 기법은 수치 및 범주형 결측값에 모두 적용할 수 있지만 소프트웨어 프로젝트 데이터가 소규모(100개 이하)인 경우에만 적용이 가능한 제약이 있다.

4. K-NN과 MLE 결합에 의한 결측값 대치

일반적으로 소프트웨어 개발조직에서 수집된 소프트웨어 프로젝트 데이터 규모는 타분야(예: 생명공학 등)의 데이터에 비해 그 규모가 매우 작다[7]. 따라서 이러한 특성에 적합하면서 데이터의 활용도를 높일 수 있는 결측값 대치법이 요구된다. 여러 기법들 중 K-NN은 구현이 간단하고 세 가지 결측 메커니즘에 강건하며 대치 정확도가 우수하다고 알려져 있다. 특히 작은 규모의 소프트웨어 데이터에서 그 성능의 유효함이 여러 문헌을 통해 입증되었다[1,3,7]. 그러나 K-NN은 결측값 대치시 결측값을 포함하는 인스턴스의 유효한 관측정보를 활용하지 않는다. 그림 1은 6개의 변수(a_1~a_6)를 갖는 6개의 인스턴스(c_1~c_6)를 이용하여 결측값 대치와 관련된 용어를 설명하고 있다. c_1~c_3은 결측값이 없는 완전한 인스턴스들이고, c_4~c_6은 결측값을 포함하는 인스턴스로 'X'로 표시한 부분이 결측값을 나타내며, 색이 칠해져 있는 변수들은 결측값을 포함하는 인스턴스에서 유효한 관측정보를 나타낸다. 예로 c_4의 a_3에 대한 결측값을 대치할 경우, K-NN은 c_1~c_3만을 대상으로 c_4와 가장 유사한 K개의 인스턴스를 찾아 이의 a_3에 해당하는 값들의 대표값을 c_4의 a_3에 대치를 한다. 이때 결측값을 포함하는 c_5와 c_6의 관측정보는 활용하지 않는다.

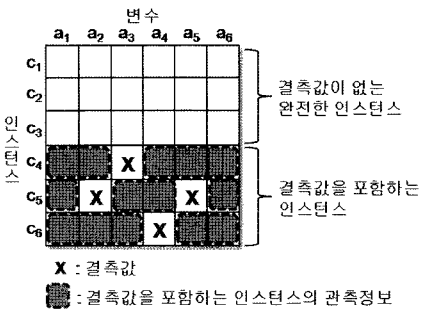

그림 1 결측값을 포함하는 데이터셋에 대한 용어

따라서 본 연구에서는 K-NN의 장점을 가지면서, 결측값을 포함하는 인스턴스의 유효한 관측정보들도 활용하여 보다 정확한 값을 대치하는 기법을 제안하고자 하는데, 이를 위해 고려한 기법이 MLE이다. MLE는 K-NN과는 달리 전체 관측값을 이용하여 결측값에 대한 추정값을 계산하며 결측값 대치의 정확도가 우수하다고 알려져 있으나 데이터 규모에 의존적임이 밝혀져 소프트웨어 프로젝트 데이터 규모에 따라 제한적으로 활용될 수 있다[2].

이 장에서는 K-NN과 MLE의 결합을 통해 이들 기법이 갖는 단점을 보완한 결측값 대치법에 대해 상세히 설명한다.

4.1 K-NN과 MLE 결합에 의한 결측값 대치법

그림 2는 본 연구에서 제안하는 소프트웨어 프로젝트 수치 데이터용 결측값 대치법의 개요를 나타내고 있다.

첫 번째 단계는 소프트웨어 프로젝트 데이터에 포함된 결측값에 대해 MLE를 적용하여 결측값을 MLE의 결과값으로 1차 대치한다. 이는 추후 K-NN 적용시 결

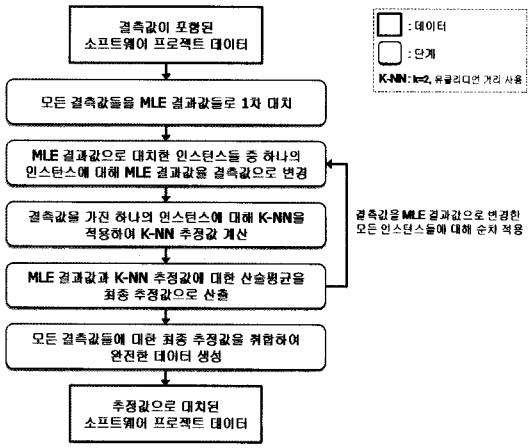

그림 2 K-NN과 MLE 결합에 의한 결측값 대치법 절차

측값을 포함하는 인스턴스들의 관측정보 활용을 가능하게 하는 것을 목적으로 한다. 그림 3은 제안하는 새로운 대치 기법의 적용 예로써 1)이 첫 번째 단계에 해당한다. 1.1 데이터셋에서 결측값 X_1, X_2, X_3가 MLE 결과값으로 변경된 것을 알 수 있다.

두 번째 단계는 MLE 결과값으로 초기화한 인스턴스들 중 하나의 인스턴스에 대해 MLE 결과값을 결측값으로 변경한다. 그림 3의 2)에 해당하며 2.1 데이터셋에서 인스턴스 c_3의 MLE 결과값이 결측값으로 변경되었다.

세 번째 단계는 두 번째 단계의 결과인 결측값을 가진 하나의 인스턴스에 대해 K-NN을 적용하여 추정값을 계산한다. 이때 첫 번째 단계에서 결측값을 MLE의 결과값으로 1차 대치한 인스턴스들도 K-NN 결과 계산에 모두 활용되어 결측값을 가진 인스턴스들의 관측정

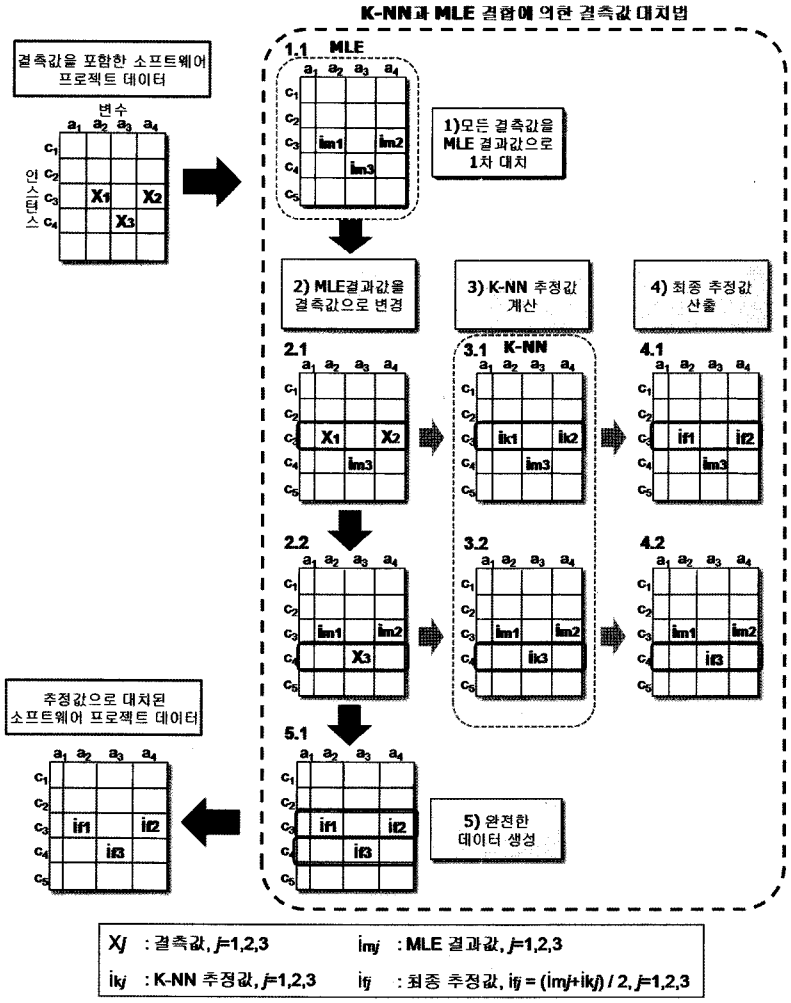

그림 3 K-NN과 MLE 결합에 의한 결측값 대치법 예

보를 모두 이용할 수 있게 된다. 기존의 K-NN에서는 결측값을 가진 c_4가 c_3의 결측값의 대치값을 찾는 과정에서 제외되지만 제안하는 기법에서는 MLE를 통해 1차 대치된 값을 c_4가 가지고 있으므로 c_4의 결측값을 제외한 나머지 관측정보를 c_3의 결측값의 대치값을 찾는 과정에 활용할 수 있다. 그림 3의 3)에 해당하며 3.1 데이터셋에서 c_3의 결측값에 대한 K-NN 추정값이 계산되었다.

네 번째 단계는 첫 번째 단계의 MLE 결과값과 세 번째 단계의 K-NN 추정값을 산술평균하여 결측값에 대한 최종 추정값을 산출하는 단계이다. 그림 3의 4)에 해당하며 4.1 데이터셋에서 c_3의 결측값 X_1, X_2의 최종 추정값으로 각각 i_{f1}, i_{f2}가 계산되었다. 이러한 일련의 과정을, 결측값을 MLE의 결과값으로 1차 대치한 모든 인스턴스들에 적용한다. 이는 그림 3에서 2)~4) 과정을 2.1, 2.2 데이터셋에 대해 반복 적용함을 의미한다.

마지막 단계에서 최종 추정값들을 모두 취합하여 결측값이 추정값으로 대치된 완전한 소프트웨어 프로젝트 데이터를 생성한다. 그림 3의 5)에 해당하며 c_3, c_4에 포함된 결측값이 최종 추정값으로 대치된 것을 확인할 수 있다.

덧붙여, 네 번째 단계에서는 하나의 인스턴스에 대한 최종 추정값이 다른 인스턴스의 추정에 영향을 주지 않도록 최종 추정값을 데이터에 즉시 반영하지 않고 임시로 보관한 후 마지막 취합단계에 반영하여 최종 결과를 생성한다.

4.2 결측값 대치법의 정확도 비교를 위한 측도

기존 결측값 대치법의 정확도 비교 연구에 사용된 측도는 Mean Magnitude of Relative Error(MMRE)[7][8], Average Absolute Error(AAE)[9,10] 등이 있다.

$$MMRE = \frac{100}{|C_{mis}|}\sum_{i=1}^{|C_{mis}|}\frac{|x_i-\hat{x_i}|}{x_i},\qquad AAE = |C_{mis}|^{-1}\sum_{i=1}^{|C_{mis}|}|x_i-\hat{x_i}|$$

$\|C_{mis}\|$	:결측값이 포함된 인스턴스 수
x_i	:결측값의 실제값
$\hat{x}_i$	:결측값의 추정값

MMRE는 동일한 오차에 대해 실제값이 작아지면 값이 증가하기 때문에 결측값 대치법의 정확도 비교에는 부적절하고, 또한 실제값이 0인 경우 사용할 수 없다는 단점이 있다. AAE는 하나의 변수에 포함된 결측값에 대한 결측값 대치의 정확도를 비교하기 위해서는 사용할 수 있지만 다수의 변수에 포함된 결측값에 대한 결측값 대치의 정확도를 비교하는 경우에는 적용하기가 어렵다. 즉, 각각의 변수가 가진 데이터의 평균과 표준편차 등 통계적 특성과 값의 단위 등이 다른 경우 변수별로 AAE가 가지는 값의 범위가 다르기 때문에 변수별 AAE를 동일한 기준으로 비교할 수 없기 때문이다.

이를 해결하기 위해, 본 연구에서는 AAE를 여러 변수 결과로 확장해서, 개별 변수의 결측값에 대해 실제값과 추정값과의 절대오차를 표준화(standardization)하여 표준화 절대오차를 구하고, 이를 다시 평균하는 Average Normalized Absolute Error(ANAE)를 제안한다. 여기서 절대오차는 하나의 결측값에 대한 실제값과 추정값과의 차를 의미한다.

$$ANAE = |C_{mis}|^{-1}\sum_{i=1}^{|C_{mis}|}\frac{|x_i-\hat{x_i}|-\mu}{\sigma}$$

μ	:개별 변수에 대한 모든 대치 기법의 오차 평균
σ	:개별 변수에 대한 모든 대치 기법의 오차 표준편차

ANAE는 하나의 변수에 포함된 모든 결측값들에 대해 결측값 대치법들을 적용하여 계산된 절대오차들의 평균과 표준편차를 이용하여 표준화한 표준화 절대오차를 사용함으로써, 절대오차가 결측값 대치법들의 절대오차 평균과 비교하여 어느 정도의 성능을 나타내는지 표현할 수 있다. 만약 표준화 절대오차가 양이면 해당 변수의 결측값에 적용한 모든 결측값 대치법들의 절대오차 평균보다 해당 수치만큼 부정확함을 의미하고, 음이면 절대오차 평균보다 해당 수치만큼 정확함을 뜻한다. 이렇게 표준화된 개별 변수들의 ANAE는 동일한 기준으로 비교할 수 있게 된다. 즉, 결측값 대치법 별로 변수들의 ANAE를 취합하여 전체 데이터에 대한 ANAE 집합을 생성하고 이들 기법별 ANAE 집합의 평균 또는 중위값을 비교하여 결측값 대치법간의 정확도를 비교할 수 있게 된다. 오차가 적을수록 ANAE의 값은 낮아지므로, 만약 어떤 결측값 대치법이 타 기법들에 비해 가장 낮은 ANAE 값을 갖는다면 이 기법은 가장 높은 정확도를 갖는 것으로 해석할 수 있다.

5. 사례 연구

이 장에서는 4장에서 설명한 K-NN과 MLE를 결합한 결측값 대치법을 실제 소프트웨어 프로젝트 데이터에 적용하고 이 결과와 타 기법의 적용결과를 ANAE를 통해 비교한 사례연구를 설명한다.

결측값 대치법의 정확도를 측정하기 위해서는 해당 결측값에 대한 실제값을 알아야 하지만 실제 소트웨어 프로젝트 데이터에 포함된 결측값의 실제값을 파악하는 것은 매우 어렵다. 따라서 이 사례연구에서는 결측값이 없는 완전한 데이터셋을 대상으로 결측값을 임의로 삽입하여 결측 데이터셋을 생성한 후 결측값 대치법들을 적용하여 정확도를 비교하였다.

실험은 그림 4와 같이 세 단계로 구성되고, 각 단계에 대한 설명과 결과는 다음과 같다.

5.1 실험 데이터 소개 및 데이터 전처리

실험을 위한 데이터는 국내의 모 금융회사(이하 회사 A)에서 수행한 소프트웨어 프로젝트 데이터이다.

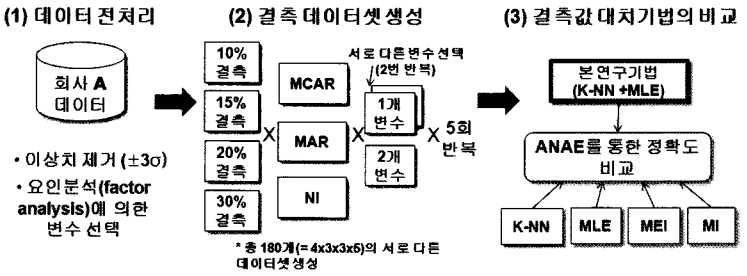

그림 4 실험 개요

표 2 실험 데이터 개요

항목	설명
출처	국내 회사 A 의 '05~'07 수행 프로젝트 데이터
프로젝트 유형	신규개발 및 유지보수
데이터 전처리	결측값 및 이상값을 포함한 인스턴스 제거
총 인스턴스 수	158개
변수	개발공수, KSLOC, 개발기간, 투입인원

표 2는 실험 데이터에 대한 세부 정보를 나타낸다.

데이터 전처리는 (1)결측값 및 이상치(outlier)가 존재하지 않는 완전한 데이터셋을 얻기 위해 결측값이나 이상치을 포함하는 인스턴스들을 제거하는 작업과 (2)각 결측값 대치기법들의 정확도를 높이기 위해 원본 데이터셋의 변수들 중에서 관련성 있는 변수들을 선택하는 작업으로 구성된다. 이상값들을 제거하기 위해 각 변수의 평균을 중심으로 ±3σ 범위에 속하지 않는 값을 포함한 인스턴스들을 제거하였고, 서로 통계적인 상관관계가 있는 변수들의 추출을 위해 요인분석(factor analysis)[11]을 사용하여 총 4개의 변수를 선택하였다.

실험 수행 환경은 이상치 제거, 관련성 있는 변수 선택, MLE 및 MI 구현을 위해 상용 통계 프로그램인 SAS를 사용하였고, K-NN, MEI 및 본 연구기법의 구현을 위해 MATLAB을 이용하였다. K=5를 사용하였고, 다른 기법들의 파라메터들은 상용 프로그램에서 제안하는 기본값(default)들로 세팅하였다.

5.2 결측 데이터셋 생성

데이터 전처리를 통해 얻은 완전한 데이터셋을 대상으로 결측값을 삽입하여 결측 데이터셋을 생성하였다. 결측 데이터셋을 생성할 때 고려한 요인은 그림 4와 같이 (1)결측률(전체 인스턴스 개수 대비 결측값을 포함한 인스턴스 개수의 비율: 10%, 15%, 20%, 30%), (2)결측 메커니즘(MCAR, MAR, NI), 그리고 (3)결측값을 포함하는 변수의 개수(1개, 2개)[1]이다. 결측 메커니즘을 고려할 때는 영향력이 가장 높은 변수인 '개발공수'를 기준으로 하여 다른 변수들에 결측값을 삽입하였다. 또한, 결측값을 포함하는 변수가 1개인 경우에는 선택된 한 특정 변수의 분포에 치우쳐진 결과를 얻는 것을 방지하기 위해 서로 다른 두 개의 변수를 선택하여 이들 각각에 대해 결측값을 삽입하였다. 실험의 정확도를 높이기 위해 위의 각 요인들의 조합을 5번 반복한 결과 총 180개의 새로운 데이터셋을 생성하였다.

5.3 결측값 대치법 간 비교결과

본 연구기법과 타 기법(K-NN, MEI, MLE, MI)들 간의 정확도 측면에서의 성능을 평가하기 위해 다음와 같은 관점에서 ANAE를 통해 비교를 수행하였다:

(1) 결측률에 따른 정확도 비교

(2) 결측 메커니즘에 따른 정확도 비교

(3) 결측값을 포함하는 변수의 개수에 따른 정확도 비교

각 비교결과에서는 결측값 대치법별로 도출된 ANAE 값들에 대한 평균과 중위값의 비교 결과가 유의미한지를 t 검정(t test)[11]과 윌콕슨 검정(Wilcoxon test)[12]을 통해 확인하였다. t 검정은 두 집단이 정규분포라는 전제하에 두 집단의 평균의 차이가 유의미한지를 검정할 때 사용하며, 윌콕슨 검정은 정규분포 가정이 불필요한 비모수검정법으로 두 집단의 중위값의 차이가 유의미한지를 검정할 때 사용한다. 여기서는 두 가지 검정시 유의수준 0.05를 사용하였다. 실험결과는 다음과 같다.

5.3.1 결측률에 따른 정확도 비교

그림 5는 결측률에 따른 결측값 대치법들의 ANAE를 결측률별로 평균한 결과를 나타내며, ANAE는 앞서 설명한 바와 같이 값이 작을수록 정확함을 의미한다.

그림 5에 따르면, K-NN과 MLE를 결합한 기법(이하 K-NN+MLE)이 모든 결측률에서 타 기법에 비해 높은 성능을 가지고 있다. MEI, MLE, MI 등 통계기반의 기법들은 결측비율이 높아질수록 성능이 전반적으로 좋아지는 현상을 보였고, K-NN은 결측률의 증가와 함께 유사한 인스턴스의 수가 줄어들어 성능이 급속도로 저하되는 특성을 보였다. K-NN+MLE는 K-NN을 적용할

1) 실험 데이터셋이 총 4개의 변수로 구성되어 있으므로 3개 이상의 변수에서 결측값이 발생되는 경우는 한 인스턴스 내에서의 결측률(75%)이 높아 실험대상에서 고려하지 않았다.

표 3 결측률에 따른 각 기법들의 ANAE값에 대한 윌콕슨 검정 및 t 검정 결과

대치기법	10%		15%		20%		30%	
	w검정 (p값)	t검정 (p값)	w검정 (p값)	t검정 (p값)	w검정 (p값)	t검정 (p값)	w검정 (p값)	t검정 (p값)
K-NN	<0.00	<0.00	<0.00	<0.00	<0.00	<0.00	<0.00	<0.00
MEI	<0.00	<0.00	<0.00	<0.00	<0.00	<0.00	<0.00	<0.00
MLE	0.16	0.99	<0.00	<0.00	<0.03	<0.02	<0.00	<0.00
MI	<0.00	<0.00	<0.00	<0.00	<0.00	<0.00	<0.00	<0.00

* w검정: 윌콕슨 검정

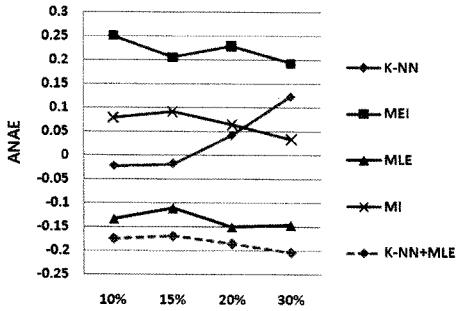

그림 5 결측률별 각 기법들의 정확도 추이

때 MLE로 대치된 값을 이용하기 때문에 K-NN의 성능 저하에 크게 영향 받지 않았다.

표 3은 결측값 대치법별 ANAE값들에 대한 평균의 비교가 유의미한지를 확인하기 위해 수행한 윌콕슨 검정과 t 검정의 결과를 나타낸다. 첫 번째 열은 K-NN+MLE와 비교된 나머지 4개 기법을 나타내고, 각 칸의 값들은 윌콕슨 검정과 t검정 결과로서 p값을 나타낸다. 결측률이 10%일 때 MLE와의 비교결과를 제외하고는 모든 p값이 유의수준인 0.05보다 작다. 이는 결측률이 10%인 경우의 MLE의 정확도는 K-NN+MLE와 차이가 없고, 다른 나머지의 경우는 K-NN+MLE의 정확도가 타 기법들에 비해 우수하다는 것을 의미한다.

5.3.2 결측 메커니즘에 따른 정확도 비교

그림 6은 결측값 대치법들의 ANAE를 결측 메커니즘별로 평균한 결과를 나타낸다.

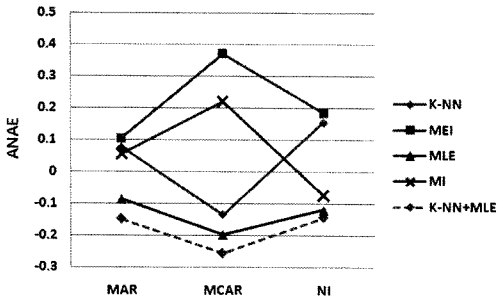

그림 6 결측 메커니즘별 각 기법들의 정확도 추이

그림 6에 따르면, K-NN+MLE가 모든 결측 메커니즘에서 타 기법에 비해 높은 정확도를 가지고 있다. K-NN, MLE, K-NN+MLE는 결측 메커니즘들 중 MCAR에서 각각 가장 좋은 성능을 보였고, MI는 NI에서 좋은 성능을 보였다. NI에서 K-NN은 지정된 변수들 내에서 특정 값 이상의 데이터가 결측이 됨에 따라 유사한 인스턴스의 수가 줄어들어 성능이 급속도로 저하되는 특성을 보였다. K-NN+MLE는 K-NN을 적용할 때 MLE로 대치된 값을 이용하고 이들 값들이 인접되어 있기 때문에 NI에서는 MLE와 유사한 성능을 보인 것으로 해석된다.

그림 7, 8, 9는 각 결측 메커니즘별 결측률에 따른 결측값 대치법들의 정확도 분석 결과를 설명한다.

이들 결과를 통해 K-NN+MLE가 결측 메커니즘과

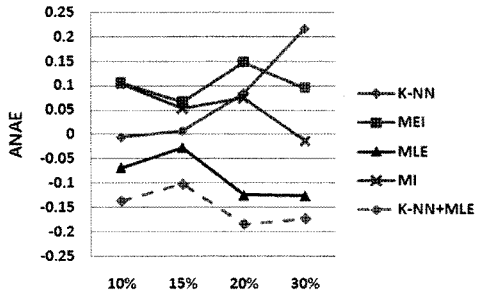

그림 7 MAR에서 결측률에 따른 각 기법들의 정확도 추이

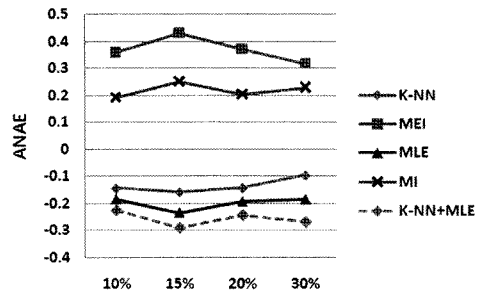

그림 8 MCAR에서 결측률에 따른 각 기법들의 정확도 추이

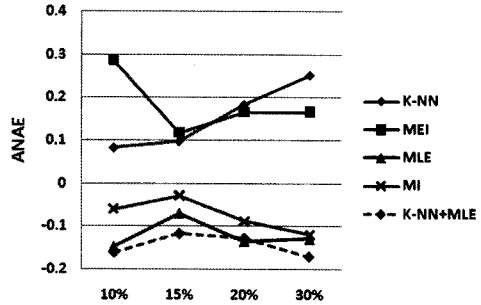

그림 9 NI에서 결측률에 따른 각 기법들의 정확도 추이

결측률에 큰 영향 없이 타 기법들에 비해 높은 정확도를 보임을 알 수 있다.

K-NN의 경우 MAR과 NI에서 결측률이 높아질수록 정확도가 급격히 떨어지는데(특히 NI), 이는 앞서 설명한 바와 같이 결측으로 인해 유사한 인스턴스의 수가 줄어들기 때문에 발생되는 현상이다. 따라서 NI에서 K-NN+MLE와 MLE는 거의 유사한 성능을 갖는 것으로 해석된다.

표 4 결측 메커니즘에 따른 각 기법들의 ANAE값에 대한 윌콕슨 검정 및 t 검정 결과

대치기법	MAR		MCAR		NI	
	w검정 (p값)	t검정 (p값)	w검정 (p값)	t검정 (p값)	w검정 (p값)	t검정 (p값)
K-NN	<0.00	<0.00	<0.00	<0.00	<0.00	<0.00
MEI	<0.00	<0.00	<0.00	<0.00	<0.00	<0.00
MLE	<0.01	<0.01	<0.01	<0.00	0.07	<0.002
MI	<0.00	<0.00	<0.00	<0.00	<0.00	<0.00

* w검정: 윌콕슨 검정

표 4는 결측 메커니즘에 따른 결측값 대치법별 ANAE 값들에 대한 평균의 비교가 유의미한지를 확인하기 위해 수행한 윌콕슨 검정과 t 검정의 결과를 나타낸다. 결측 메커니즘이 NI인 경우 윌콕슨 검정에서 MLE의 정확도와 K-NN+MLE와의 정확도 간에 차이가 없는 것으로 나타났다.[2] 이를 제외한 다른 모든 경우에서는 K-NN+MLE의 성능이 나머지 기법들에 비해 모든 결측 메커니즘에서 보다 높은 정확도를 보임을 보였다.

5.3.3 결측값을 포함하는 변수의 개수에 따른 정확도 비교

그림 10은 결측값 대치법들의 ANAE를 결측값을 포함하는 변수의 개수별로 평균한 결과를 나타낸다.

2) 동일한 조건에서 t 검정은 K-NN+MLE와 MLE간의 정확도에 차이가 있다는 결과를 보였으나, 이 경우 MLE의 ANAE 결과들이 정규분포를 따르지 않는 것으로 해석될 수도 있으므로 윌콕슨 검정의 결과를 따랐다.

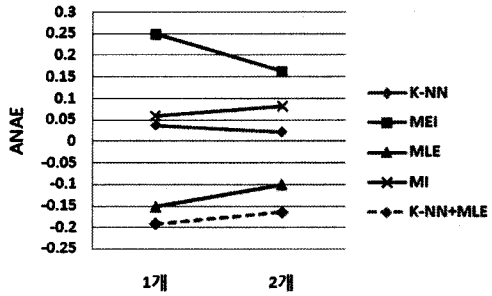

그림 10 결측값을 포함하는 변수의 개수에 따른 각 기법들의 정확도 추이(총 4개 변수에서 1개 및 2개)

그림 10에 따르면, 결측값을 포함하는 변수의 개수가 1개 또는 2개인 경우 K-NN+MLE가 타 기법에 비해 높은 정확도를 가지고 있고, 1개일 때에 비해 2개일 때 성능이 조금 떨어지는 경향을 보였다.

그림 11과 12는 결측값을 포함하는 변수의 개수에 따른 결측률의 변화에 대한 결측값 대치법들의 정확도 분석 결과를 보이고 있다. 이전의 결과와 마찬가지로 결측값을 포함하는 변수의 개수에 상관없이 모든 결측률에서 K-NN+MLE가 타 기법들에 비해 높은 정확도를 보였다.

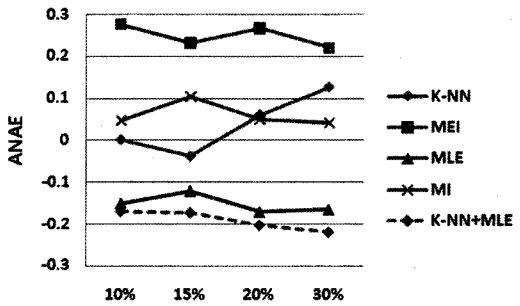

그림 11 결측값을 포함하는 변수의 개수가 1개인 경우 결측률에 따른 각 기법들의 정확도 추이

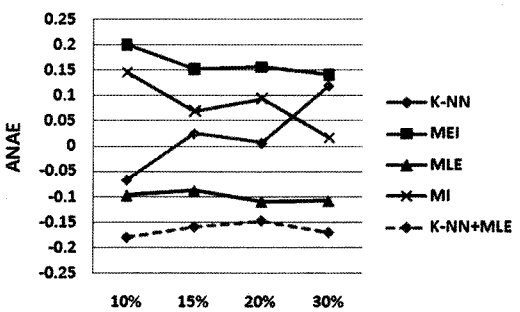

그림 12 결측값을 포함하는 변수의 개수가 2개인 경우 결측률에 따른 각 기법들의 정확도 추이

표 5 결측값을 포함하는 변수의 개수에 따른 각 기법들의 ANAE값에 대한 윌콕슨 검정 및 t 검정 결과

대치기법	변수 1개		변수 2개	
	w검정 (p값)	t검정 (p값)	w검정 (p값)	t검정 (p값)
K-NN	<0.00	<0.00	<0.00	<0.00
MEI	<0.00	<0.00	<0.00	<0.00
MLE	<0.00	<0.00	<0.00	<0.00
MI	<0.00	<0.00	<0.00	<0.00

* w검정: 윌콕슨 검정

표 5는 결측값을 포함하는 변수의 개수에 따른 결측값 대치법별 기법들의 ANAE값들에 대한 평균의 비교가 유의미한지를 확인하기 위해 수행한 윌콕슨 검정과 t 검정의 결과를 나타낸다. 두 가지 검정 결과 K-NN+MLE의 성능이 모든 경우에서 나머지 기법들에 비해 높은 정확도를 보였다.

6. 결론 및 향후 연구

본 연구에서는 K-NN과 MLE를 결합한 새로운 소프트웨어 프로젝트 수치 데이터용 결측값 대치법과 결측값 대치법의 정확도를 비교하기 위한 새로운 측도인 ANAE를 제안하였다. 그리고, 새로운 결측값 대치법의 효용성을 확인하기 위해 국내 금융회사에서 수행한 소프트웨어 프로젝트 데이터를 대상으로 네 가지 결측값 대치법들과 정확도를 비교하였다.

사례연구에서 결측률(10%, 15%, 20%, 30%), 결측메커니즘(MAR, MCAR, NI), 그리고 결측값을 포함하는 변수의 개수(1개, 2개)라는 세 가지 요인의 조합에 따라 실험을 수행한 결과, 본 연구에서 제안하는 K-NN+MLE 기법이 기존의 타 기법들에 비해 가장 높은 정확도를 보였다. 또한 이 결과들을 대상으로 통계적인 검정을 수행한 결과, 결측률이 10%인 데이터셋과 결측메커니즘이 NI인 경우 K-NN+MLE와 MLE의 정확도가 유사한 것으로 확인되었으나, 이 두 경우를 제외한 나머지 경우는 모두 통계적으로 K-NN+MLE의 성능이 가장 좋은 것으로 판명되었다.

향후 연구로는 본 기법을 범주데이터(categorical data)에 결측값이 있는 데이터셋의 경우에도 적용 가능한 기법으로 확장하고, 보다 다양한 특성을 갖는 소프트웨어 프로젝트 데이터에 적용할 예정이다.

참 고 문 헌

[1] Kevin Strike, Khaled El Emam, and Nazim Madhavji, "Software Cost Estimation with Incomplete Data," IEEE Transactions on Software Engineering, Vol.27, No.10, pp. 890-908, 2001.

[2] Ingunn Myrtveit, Erik Stensrud, and Ulf H. Olsson, "Analyzing Data Sets with Missing Data: An Empirical Evaluation of Imputation Methods and Likelihood-Based Methods," IEEE Transactions on Software Engineering, Vol.27, No.11, pp. 999-1013, 2001.

[3] M. H. Cartwright, M. J. Shepperd, and Q. Song, "Dealing with Missing Software Project Data," Proceeding of the Ninth International Software Metrics Symposium, pp. 154-165, 2003.

[4] Roderick J. A. Little, Donald B. Rubin, Statistical Analysis with Missing Data, John Wiley & Sons, 1987.

[5] Donald B. Rubin, Multiple imputation for nonresponse in surveys, John Wiley & Sons, 1987.

[6] Bhekisipho Twala, Michelle Cartwright, and Martin Shepperd, "Comparison of Various Methods for Handling Incomplete Data in Software Engineering Databases," International Symposium on Empirical Software Engineering, pp. 105-114, 2005.

[7] Qinbao Song, Martin Shepperd, "A new imputation method of small software project data sets," The Journal of Systems and Software, Vol.80, No.1, pp. 51-62, 2007.

[8] Qinbao Song, Martin Shepperd, and Michelle Cartwright, "A Short Note on Safest Default Missingness Mechanism Assumptions," Empirical Software Engineering, Vol.10, No.2, pp. 235-243, 2005.

[9] Jason Van Hulse, Taghi M. Khoshgoftaar, "A comprehensive empirical evaluation of missing value imputation in noisy software measurement data," The Journal of Systems and Software, Vol. 81, No.5, pp. 691-708, 2008.

[10] Taghi Khoshgoftaar, Andres Folleco, Jason Van Hulse, and Lofton Bullard, "Multiple Imputation of Missing Values in Software Measurement Data," International Journal of Software Measurement, Vol.1, No.1, pp. 1-12, 2007.

[11] Anthony J. Hayter, Probability and Statistics for Engineers and Scientists, 3rd Ed., Thomson Higher Education, 2007.

[12] Frank Wilcoxon, "Individual Comparisons by Ranking Methods," Biometrics Bulletin, Vol.1, No.6, pp. 80-83, 1945.

이 동 호

1992년~2000년 부산대학교 전자공학 졸업(학사). 2007년~2009년 KAIST 전산학 졸업(석사). 관심분야는 소프트웨어 데이터 품질, 정량적인 프로젝트 관리

윤 경 아

1996년 동국대학교 컴퓨터공학 졸업(학사). 1996년~2000년 삼성SDS 솔루션 사업부. 2003년 KAIST 전산학 졸업(석사). 2003년~KAIST 전산학 박사과정 관심분야는 소프트웨어 측정 및 분석, 소프트웨어 데이터 품질, 정량적인 프로젝트 및 프로세스 관리

배 두 환

1980년 서울대학교 조선공학 졸업(학사) 1987년 Univ. Of Wisconsin-Milwaukee 전산학 졸업(석사). 1992년 Univ. Of Florida 전산학 졸업(박사). 1995년~KAIST 전산학과 교수. 관심분야는 소프트웨어 프로세스, 객체지향 프로그래밍, 컴포넌트 기반 프로그래밍, 임베디드 소프트웨어 설계, 관점지향 프로그래밍