K-NN과 최대 우도 추정법을 결합한 소프트웨어 프로젝트 수치 데이터용 결측값 대치법에 관한 연구*

이동호°, 윤경아, 배두환
카이스트 정보과학기술대학 전자전산학부
{dhlee, kayoon, bae}@se.kaist.ac.kr

An approach to missing data imputation by combining K nearest neighbor with maximum likelihood estimation for numerical software project data

Dong-Ho Lee°, Kyung-A Yoon, Doo-Hwan Bae
Division of Computer Science, College of Information Science & Technology, KAIST

소프트웨어 프로젝트 데이터를 이용한 각종 분석예측모델 생성시 직면하는 문제 중 하나는 데이터에 포함된 결측값(Missing data)이다. 결측값이 포함된 데이터는 일반적으로 활용되지 않고 버려지는데이는 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 효율적이라는 것이 많은 연구결과로 입증되고 있다[1][2][3].

이러한 연구결과 중 소프트웨어 프로젝트 데이터를 위한 결측값 대치법으로K 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 반면결측값을 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 방식을 가정하는 것이 필요하다.

본 연구는 소프트웨어 프로젝트 데이터의 결측값 대치시 결측값을 포함하는 데이터의 관측정보를 활용할 경우, 보다 정확한 결측값 대치가 가능할 수 있다는 점에서 착안하게 되었다.본 연구에서는 이러한 두 가지 결측값 대치법의 특징을 고려하여 두 기법을 결합한 새로운 소프트웨어 프로젝트 수치 데이터용 결측값 대치법을 제안하고자 한다.그리고 결측값 대치법의 정확성을 비교하기 위한 새로운 측도(Measure)를 함께 제안한다. 본 연구에서 제안하는 결측값 대치법의 특징은 두 가지 결측값 대치법을 결합함으로써, K-NN 적용시 결측값을 MLE 결과값으로 대치하여 결측값을 포함하는 데이터의 관측정보를 활용할 수 있어 보다 정확한 결측값 대치를 가능하게 한다.

아래 그림은 본 연구에서 제안하는K-NN과 MLE 결합에 의한 결측값 대치법을 나타낸다.첫 번째 단계는 소프트웨어 프로젝트 데이터에 포함된 결측값에 대해MLE를 적용하여 결측값을 MLE의 결과값으로 1차 대치한다. 두 번째 단계는 MLE 결과값으로 초기화한 인스턴스들 중 하나의 인스턴스에 대해 MLE 결과값을 결측값으로 변경한다.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 관측정보를 모두 이용할 수 있게 된다.네 번째 단계는 첫 번째 단계의 MLE 결과값과 세 번째 단계의 K-NN 추정값을 산술평균하여 결측값에 대한 최종 추정값을 산출하는 단계이다.마지막 단계에서 최종 추정값들을 모두 취합하여 결측값이 추정값으로 대치된 완전한 소프트웨어 프로젝트 데이터를 생성한다.

기존 결측값 대치법의 정확성 비교 연구에 사용된 측도는 Mean Magnitude of Relative Error

*본 연구는 지식경제부 및 정보통신연구진흥원의 대학IT연구센터 지원사업(IITA-2008-(C1090-0801-0032))과, 방위사업청과 국방과학연구소의 지원으로 수행되었음

(MMRE)[4], Average Absolute Error(AAE)[5] 등이 있다. MMRE는 동일한 오차에 대해 실제값이 작아지면 값이 증가하기 때문에 결측값 대치법의 정확성 비교에는 부적절하다 또한 AAE는 다수의 변수에 포함된 결측값에 대한 결측값 대치의 정확성을 비교하는 경우에는 적용하기가 곤란하다. 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. 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.

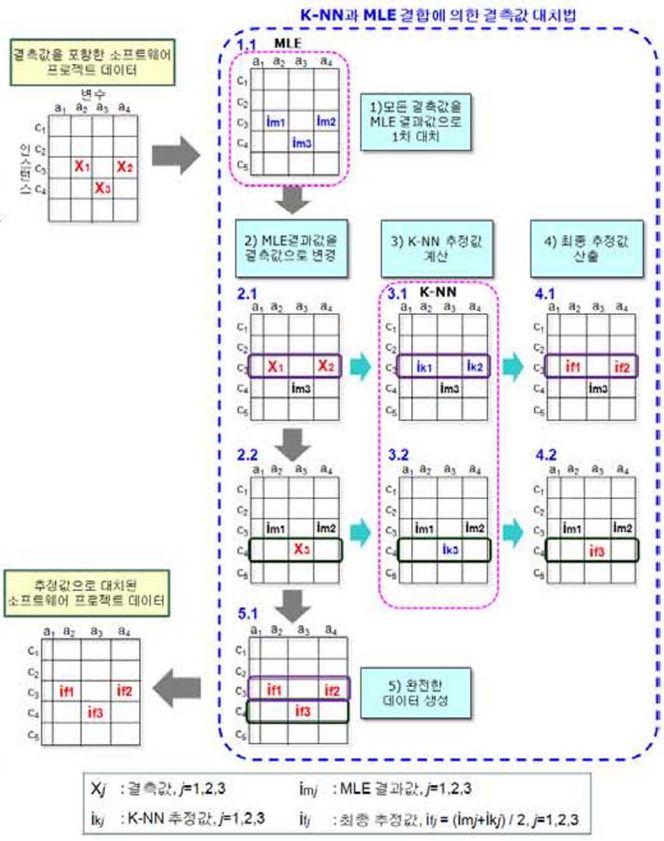

그림. K-NN과 MLE 결합에 의한 결측값 대치법

본 연구에서 제안한 결측값 대치법의 정확성 비교를 위해 평균대치법(Mean imputation), K-NN, MLE, 다중대치법(Multiple imputation) 등 모두 4개의 결측값 대치 기법들을 사용하였으며, 실험 데이터는 국내 모 금융회사의 소프트웨어 프로젝트 데이터를 사용하였다실험 방법은 결측값이 없는 완전한 데이터셋을 대상으로 결측값을 임의로 삽입하여 결측 데이터셋을 생성한 후 결측값 대치 기법들을 적용하여 정확성을 비교하였다 결측 데이터셋 생성시 고려한 요인은 결측 메커니즘 전체 인스턴스 개수 대비 결측값을 포함한 인스턴스 개수 비율결측값을 포함하는 변수 개수 등 세 가지이며 총 100개의 결측 데이터셋을 생성하였다

실험 결과는 결측값을 포함한 변수가1~2개인 경우 본 연구에서 제안하는 결측값 대치법이 가장 정확하고, 결측값을 포함한 변수가 3개인 경우엔 새로운 결측값 대치법과MLE가 동일한 정확성을 보임을 통계적 검정으로 확인하였다. 이와 같은 실험 결과를 바탕으로 소프트웨어 프로젝트 수치용 데이터에 포함된 결측값이 전체 변수의50% 이내에 분포하는 경우 본 연구에서 제안하는 대치 기법이 유용할 것으로 판단된다. 향후 연구로는 보다 다양한 형태의 결측 데이터셋에 대한 실험 및K-NN과 다중대치법 등 다양한 결측값 대치법 조합과의 비교를 수행할 예정이다

참고 문헌

[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] 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.
[5] 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.