소프트웨어 공수 예측의 정확성에 대한 이상치 제거의 영향 분석

(Analyzing Influence of Outlier Elimination on Accuracy of Software Effort Estimation)

서 영 석 † 윤 경 아 † 배 두 환 ††
(Yeong-Seok Seo) (Kyung-A Yoon) (Doo-Hwan Bae)

요 약 정확한 소프트웨어 공수 예측은 소프트웨어 관련 여러 커뮤니티들에서 예전부터 항상 이슈가 되어 왔다. 소프트웨어 공수 예측의 정확도를 향상시키기 위해 지금까지 많은 연구들에서는 데이타 품질이 공수 예측에 중요한 요소들 중 하나임에도 불구하고 이것에 대한 고려 없이 공수 예측 기법들에만 초점을 맞추어 왔다. 본 연구에서는 소프트웨어 공수 예측 기법과 이상치 제거 기법들 사이의 영향 관계를 공수 예측 정확도의 관점에서 실험적으로 살펴본다. 두 개의 프로젝트 데이타들(ISBSG와 국내의 한 금융 조직으로부터 수집된 데이타)에 대해 일반적으로 많이 사용되는 세 가지 공수 예측 기법(최소제곱법, 신경망 네트워크, 그리고 베이지안 네트워크)과 두 가지 이상치 제거 기법(최소절사제곱법과 K-means 클러스터링)을 적용시켜 그 결과들을 서로 비교해 보고 이상치 제거 기법을 적용하지 않은 결과와도 비교해 본다.

키워드 : 이상치 제거, 공수 예측, 데이타 품질

Abstract Accurate software effort estimation has always been a challenge for the software industrial and academic software engineering communities. Many studies have focused on effort estimation methods to improve the estimation accuracy of software effort. Although data quality is one of important factors for accurate effort estimation, most of the work has not considered it. In this paper, we investigate the influence of outlier elimination on the accuracy of software effort estimation through empirical studies applying two outlier elimination methods(Least trimmed square regression and K-means clustering) and three effort estimation methods(Least squares regression, Neural network and Bayesian network) associatively. The empirical studies are performed using two industry data sets(the ISBSG Release 9 and the Bank data set which consists of the project data collected from a bank in Korea) with or without outlier elimination.

Key words : outlier elimination, effort estimation, data quality

· 본 연구는 지식경제부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었고(IITA-2008-(C1090-0801-0032)), 또한 방위사업청과 국방과학연구소의 지원으로 수행되었습니다.

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

논문접수 : 2008년 4월 14일
심사완료 : 2008년 9월 22일

정보과학회논문지 : 소프트웨어 및 응용 제35권 제10호(2008.10)

1. 서 론

최근 소프트웨어의 활용범위가 다양한 분야로 확대되고 소프트웨어의 크기 및 복잡도 등도 점점 증가함에 따라 소프트웨어 프로젝트 관리의 중요성이 더욱 커지고 있다. 특히, 현실성 있고 효과적인 프로젝트 관리를 위해서는 계획단계에서의 정확한 소프트웨어 비용 산정이 필수적인데, 이는 전체 프로젝트의 성패에 큰 영향을 준다[1]. 예로, 프로젝트 비용에 대한 과소 산정은 프로젝트 기간이나 예산의 초과로 인한 프로젝트 실패의 원인이 되는 반면, 프로젝트 비용에 대한 과대 산정은 소프트웨어 개발 자원의 초과할당으로 인한 비용낭비의 원인이 된다.

소프트웨어 비용 산정은 주로 공수의 예측을 통해 수

행되는데[2], 공수 예측은 과거 프로젝트 데이타를 기반으로 한 예측모델의 생성을 통해 이루어진다[2-5]. 그러나 수집된 과거 프로젝트 데이타에는 프로젝트 데이타의 품질을 저하시키는 이상치가 존재하는데, 이는 보편적인 특성을 가진 데이타와는 다른 특성을 가지는 데이타를 의미한다. 이상치 발생 원인으로는 (1)비슷한 크기의 프로젝트를 하던 한 조직이 때때로 아주 대규모나 소규모의 프로젝트를 수행하거나 (2)특정 프로젝트가 불안정하게 수행될 경우, 그리고 (3)프로젝트 정보에 대한 측정 기록자나 측정시스템의 잘못된 측정 등이 있다.

'Garbage In, Garbage Out'이 의미하듯, 이상치가 많이 포함된 과거 프로젝트 데이타로 공수 예측 모델을 만든다면 현실을 잘 반영하지 못하고 왜곡된 공수 예측 결과를 얻게 된다. 예를 들어, 그림 1에서 이상치 제거 전(그림 A)과 후(그림 B)에 공수 예측 모델을 비교하면, 2개의 이상치를 제거하기 전의 기울기 변화가 현저함을 확인할 수 있다. 결과적으로 조직 내에서 주로 수행되는 프로젝트의 일반적인 특징이 잘 반영된 공수예측을 위해서는 이상치의 제거가 반드시 필요하다.

이처럼 데이타의 품질이 정확한 공수 예측을 위해 매우 중요한 요소임에도 불구하고 지금까지 수행된 공수 예측과 관련한 많은 연구들[2-5]에서는 이상치 제거에 대해 고려하지 않고 있고, 이와 관련된 연구도 이제 관심을 갖기 시작하는 추세이다[6]. 또한 다양한 이상치 제거 기법이 공수 예측 결과에 주는 효과에 대해서도 연구된 바가 거의 없다.

따라서, 본 연구에서는 소프트웨어 공수 예측 기법과 데이타 품질을 저해하는 이상치를 제거하는 기법들 사이의 영향 관계에 대해 공수 예측의 정확도 관점에서 살펴보고자 한다. 실제 업체에서 수집된 2개의 데이타셋을 대상으로 최소절사제곱법과 K-means 클러스터링 기법에 의해 각각 이상치를 제거하고, 이 데이타셋과 원본 데이타셋을 바탕으로 일반적으로 공수 예측 기법으로 많이 사용되는 최소제곱법, 신경망 네트워크, 베이지안 네트워크를 이용하여 공수 예측을 하고 이상치 제거 전후의 공수 예측 정확도를 비교하였다.

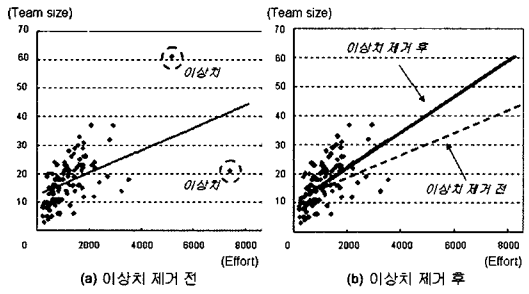


그림 1 이상치 제거에 의한 공수 예측 모델의 변화

본 논문의 구성은 다음과 같다. 2장에서는 이 논문에서 사용된 공수 예측 모델을 만들기 위한 기법과 이상치의 제거 기법에 대해 살펴본다. 3장에서는 관련연구에 대한 소개로써 기존 연구의 방식과 몇 가지 한계점을 제시한다. 4장에서는 본 연구의 전체 실험 설계와 과정에 대해 설명하고 5장에서는 본 연구의 전체 실험 결과를 보여준다. 6장에서는 결론 및 향후 연구로 본 논문을 맺도록 한다.

2. 배경 지식

본 연구에서 적용한 공수 예측 기법들과 이상치 제거 기법들은 각각 최소제곱법, 신경망 네트워크, 베이지안 네트워크와 최소절사제곱법, K-means 클러스터링으로서 각 기법들은 서로 다른 이론적 배경을 가지고 있다. 이 장에서는 이 기법들에 대한 기본 개념들에 대해 설명한다.

2.1 공수 예측 기법

최소제곱법(Least Squares). 최소제곱법은 측정된 데이타의 실제값과 그 데이타에 대한 예측값의 차이(잔차)를 제곱하여 합한 값이 최소화되는 직선을 찾는 통계 기반의 회귀 모델 생성 기법이다. 이 기법은 다른 기법들과 비교했을 때 상대적으로 간단하고, 대부분의 통계 소프트웨어 툴에서 쉽게 사용해 볼 수 있기 때문에 [7] 실제 공수 예측을 위한 많은 연구들에서 가장 일반적으로 적용되고 있는 기법이다[2-4].

신경망 네트워크(Neural Network). 신경망 네트워크는 먼저 측정된 데이타들을 네트워크를 통해 학습시킨 후 각 뉴런에 대한 적절한 가중치가 부여되면 입력 파라미터들에 해당하는 값을 입력하여 원하는 값을 예측하는 기계 학습 기반의 기법[8]이다. 신경망 네트워크에는 여러가지 형태들이 존재한다. 그 중에서 공수 예측을 위해 가장 일반적으로 많이 사용되는 형태는 역전파 알고리즘을 갖는 전방향 신경망(Feed-forward neural network with the back propagation algorithm)이다[2,7]. 이 기법은 다른 공수 예측 기법들에 비해 좀더 정확한 공수 예측 결과를 보여준다.

베이지안 네트워크(Bayesian Network). 베이지안 네트워크는 측정된 데이타들을 학습시켜 각 데이타의 변수들이 발생할 확률적 관계를 통해 예측하는 확률 기반의 기법이다. 이 기법은 각 변수들 사이의 인과관계 모델(causal model)을 생성하는 단계와 베이즈 정리[9]를 이용하여 각 변수의 확률테이블을 만들고 그것을 이용하여 전체 결합확률분포를 구하는 단계로 크게 나누어진다. 이 기법은 프로젝트 위험관리에 많이 사용되지만 공수 예측을 위해서도 연구들이 이루어지고 있다

[5,9]. 공수를 low, medium, high처럼 도메인 지식에 따라 의미있는 수준(level)으로 나누고 인과관계 모델을 만든 후, 각각의 수준별로 제시되는 결합확률분포 값을 이용하여 공수 예측이 이루어진다.

2.2 이상치 제거 기법

최소절사제곱법(Least Trimmed Squares). 2.1에서 설명한 최소제곱법은 잔차 제곱합이 최소화 되는 직선을 찾는 기법이지만, 최소절사제곱법은 절사값(h)까지의 잔차 제곱합이 최소화되는 직선을 찾는 기법이다[10,11]. 최소절사제곱법에서는 모든 데이타에 대한 잔차 제곱합의 크기를 작은 순에서 큰 순으로 정렬시켜서 $(\varepsilon_1^2 \le \varepsilon_2^2 \le ... \le \varepsilon_n^2)$ 절사값(h) 이내의 데이타만 이용하기 때문에 본 연구에서는 절사값 이후의 값들을 이상치로 식별하고 제거한다. 그림 2는 최소절사제곱법에 의해 식별된 이상치의 예를 나타낸다. 잔차가 매우 큰 값들이 이상치로 식별되고, 잔차 제곱합이 최소화되는 직선에 가까운 데이타들이 정상치로 식별된다.

K-means 클러스터링(K-means Clustering). 클러스터링은 비슷한 특성을 가지는 데이타들끼리 하나의 집단으로 묶는 데이타 마이닝 기법이다[12]. 이 기법은 자료를 집단별로 구분해주어야 하는 감독 학습(supervised learning)과는 달리 자료집단의 유사성을 바탕으로 스스로 집단을 나누어 나가는 비감독 학습(unsupervised learning)방식을 가진다. 클러스터링을 위한 알고리즘에는 여러가지가 있는데 본 연구에서는 K-means 클러스터링 기법을 적용하였다. K-means 클러스터링에서 가장 중요한 문제는 전체 데이타를 몇 개의 클러스터로 나눌지에 대한 것이므로 이를 해결하기 위해 본 연구에서는 평균 실루엣(silhouette) 값을 이용하였다. 하나의 데이타에 대한 실루엣 값의 의미는 다른 클러스터들의 데이타와 비교해 그것이 속한 클러스터 내부 데이타와의 유사성에 대한 척도이다[13]. 실루엣 값의 범위는 -1에서 1까지 이며 1에 가까울수록 그 데이타가 적합한 클러스터에 속해있다고 볼 수 있다. K-means 클러스터링을 이용하여 본 연구에서는 0보다 작은 실루엣 값을 갖는 데이타 그리고 하나 또는 두 개의 데이타를 포함하는 클러스터를 이상치로 식별하고 제거한다. 그림 3은 K-means 클러스터링에 의해 식별된 이상치의 예를 나타낸다. 유사한 집단을 잘 이루지 못하는 데이타들이 이상치로 식별되는 것을 확인할 수 있다.

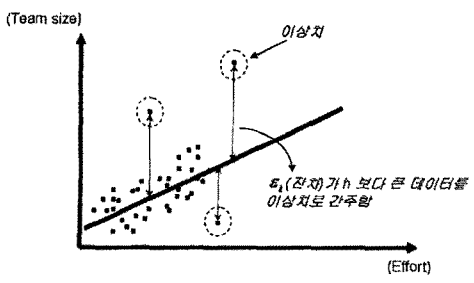


그림 2 최소절사제곱법에 의해 식별된 이상치의 예

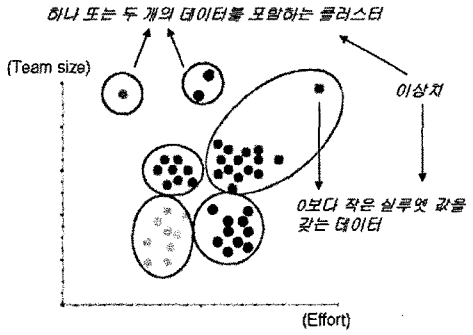


그림 3 K-mean에 의해 식별된 이상치 데이타의 예

3. 관련연구

이상치 제거 기법이 공수 예측 결과에 주는 효과에 대한 연구는 현재까지 많이 이루어지지 않은 상태이나 수집된 데이타의 품질과 이를 기반으로 한 다양한 정량적 분석모델의 성능향상에 대한 관심이 증가하면서 이에 대한 연구의 필요성이 강조되고 있다. 최근 수행된 관련연구로는 Chan의 연구[6]를 들 수 있는데, 이 연구에서는 ISBSG(release 6) 데이타[14]에 대해 최소중의 제곱법(Least Median Squares: LMS)[15]을 사용하여 이상치를 식별 및 제거한 후 공수 예측 모델을 생성했다. 이 연구는 이상치의 제거가 보다 정확한 공수 예측 모델을 생성하는데 필수적이라는 점을 보인 점에 대해서는 의미가 있으나, 연구 내용 및 실험 측면에서 다음과 같은 부족한 점을 가지고 있다. (1) 통계기반의 이상치 제거 기법만을 사용했으므로 다른 이론을 바탕으로 한 이상치 제거 기법을 사용했을 때의 공수 예측에 대한 효과를 알 수 없다. LMS등의 통계기법을 사용하기 위해서는 대상이 되는 데이타들이 특정 분포를 따라야 하거나 속성들 간 서로 독립이어야 하는 등의 조건들이 만족해야 하는데, 소프트웨어 프로젝트 데이타들은 매우 복잡한 특성을 가지고 있어서 이러한 조건을 만족하는지 여부를 알기 어렵고, 통계기법이 적합하지 않을 경우도 존재할 수 있다. 따라서 통계기반 이외의 다른 이상치 제거기법에 대한 효과에 대한 연구도 필요하다. (2) 공수 예측 정확도의 평가가 미흡하다. 일반적으로 예측 모델의 성능을 평가하기 위한 다양한 평가기준들이 존재하는데, Chan의 연구에서는 MMRE(Mean Magni-

tude of Relative Error)만을 사용하였다. 이는 각 데이타에 대한 '상대적인 에러의 크기(The Magnitude of Relative Error:MRE)'의 평균값을 의미하므로 높은 MRE값을 가지는 몇 개의 데이타에 의해 큰 영향을 받는다. 그러므로 공수 예측 모델에 대해 좀더 정확한 평가가 이루어지기 위해서는 MMRE뿐만 아니라 다른 평가기준들이 더 사용되어야 한다. (3) 데이타의 특성을 고려하지 않고 공수 예측 모델을 만들었다. 이 연구에서 사용한 ISBSG 데이타는 성격이 매우 다른 프로젝트 데이타들로 구성되어 있기 때문에 비슷한 도메인 별로 데이타를 구분하여 별개의 공수 예측 모델을 만들었다면 현실적으로 좀더 유효한 실험결과를 도출할 수 있었을 것으로 사료된다.

따라서 본 연구에서는 기존 연구의 (1), (2), (3)과 같은 부족한 점들을 보완하여, 통계기반의 기법을 포함한 여러 이상치 제거 기법들이 공수 예측 기법들의 정확성에 미치는 영향을 보다 신뢰성 있는 공수 예측 평가 기준들과 데이타 특성을 고려해 분석해 보고자 한다.

4. 실험연구

본 연구는 그림 4와 같은 4단계로 실험을 설계하였다. 연구에서 사용한 데이타와 실험환경에 대한 소개 후, 각 단계에 대한 세부적인 내용을 다음 각 절에서 설명한다.

4.1 데이타와 실험환경에 대한 소개

본 연구에서는 전세계의 다양한 조직으로부터 수집된 ISBSG 데이타와 국내의 한 금융기관으로부터 수집된 Bank 데이타를 사용하였다.

ISBSG(Release 9) 데이타[14]는 1989년부터 2004년까지 전세계 다양한 조직으로부터 수집된 프로젝트 데이타로서 공수 및 생산성과 관련된 변수들로 구성되어 있기 때문에 데이타 분포가 고르지 못함에도 불구하고 공수 예측을 위한 연구에 많이 사용된다.

Bank 데이타는 2004년부터 2006년까지 국내 한 은행에서 수집된 프로젝트 데이타이다. SW-CMM 레벨 3 단계인 회사이므로 소프트웨어 프로젝트 데이타의 분포가 ISBSG 데이타에 비해 상대적으로 고르고, 프로젝트 관리와 관련된 소프트웨어 프로젝트 생산성 및 품질과 관련된 변수들로 구성되어 있다.

실험환경으로는 최소절사제곱법, K-means클러스터링의 이상치 제거 기법들과 공수 예측 기법으로서 최소제곱법과 신경망 네트워크를 위해 MATLAB v7.3.0.267을 사용하였고, 베이지안 네트워크를 위해서는 BayesiaLAB v4.3.1을 사용하였다. 통계적 유의수준은 0.05로 설정하였다.

4.2 데이타 전처리

올바른 실험결과를 위해 입력데이타의 전처리 과정이 필요하다. 먼저 실험에 사용할 프로젝트 데이타에서 공수에 영향을 미치는 변수들을 선택하고 결측치는 보정한다. 정규분포를 가정하는 통계 기반의 이상치 제거 기법이나 공수 예측 기법을 적용할 경우 이 단계에서 정규성 검정과 독립변수 및 종속변수 사이의 상관관계 분석을 수행한다. 또한, 데이타 마이닝 기반의 이상치 제거 기법이나 공수 예측 기법을 적용할 경우에는 각 변수들의 규모(scale)의 차이를 정규화를 통해 동일하게 맞추어 이들의 영향을 동일하게 만드는 작업을 수행한다.

따라서 본 연구에서는 전체 3,024개의 ISBSG(Release 9) 데이타 중에서 데이타 품질과 금융 도메인을 고려한 기준에 근거하여 99개의 데이타와 표 1과 같은 변수들을 선택했다.

변수들 중 "프로젝트 기간", "언어"에 대해 각각 6%, 4%만큼의 결측치가 있었는데 이것을 보정하기 위하여 K-nearest neighbor(K-NN)보정 기법[16]을 사용하였다. 본 연구에서는 통계기반의 최소제곱법과 최소절사제곱법을 적용하기 때문에 데이타의 log변환 후 Shapiro-Wilks기법[17]을 적용하여 정규성 검정을 하였고, 피어슨(Pearson) 상관관계 분석과 1요인 분산분석을 통해

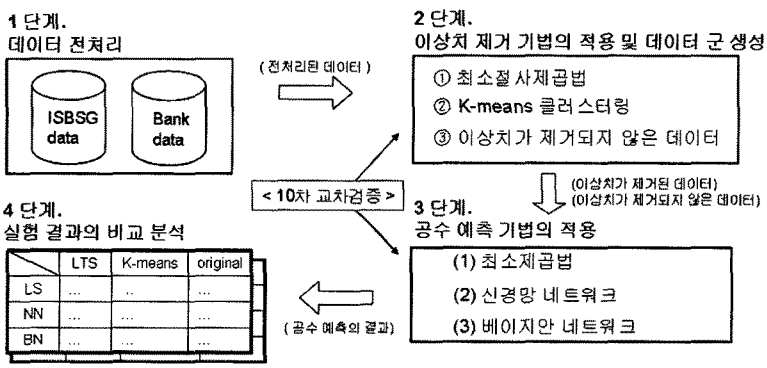


그림 4 본 연구의 실험 개요

표 1 본 연구에서 사용된 ISBSG 데이타의 변수들

변수명	척도	설명	평균 /표준편차	형태
공수	비율척도	정규화된 전체 프로젝트 공수 (person hours)	3745.85 /3859.41	종속변수
프로젝트 크기	비율척도	조정된 기능 점수	443.73 /1493.76	독립변수
프로젝트 기간	비율척도	전체 프로젝트 기간(calendar months)	7.77 /7.37	
개발 형태	명목척도	주로 사용된 개발 형태	•	
개발 플랫폼	명목척도	주로 사용된 개발 플랫폼	•	
언어	명목척도	주로 사용된 언어 종류	•	

표 2 본 연구에서 사용된 Bank 데이타의 변수들

변수명	척도	설명	평균 /표준편차	형태
공수	비율척도	전체 프로젝트 공수(person hours)	1935.57 /3769.25	종속변수
프로젝트 크기	비율척도	여러 언어로 만든 코드를 하나의 어셈블리 코드로 변환하여 계산한 LOC	340.65 /1327.61	독립변수
프로젝트 기간	비율척도	전체 프로젝트 기간(calendar days)	68.71 /57	
최대 팀 크기	비율척도	프로젝트에 투입된 최대 인원 수	20.76 /17.26	
개발 플랫폼	명목척도	주로 사용된 개발 플랫폼	•	
개발 주기 모델	명목척도	프로젝트에 사용된 개발 주기 모델	•	

독립변수와 종속변수의 올바른 상관관계를 검증해주었다. 또한, 본 연구의 K-means 클러스터링 기법을 적용하기 위하여 데이타들을 0에서 1사이의 값을 갖도록 정규화시켰다.

Bank 데이타에 대해서는 전체 127개의 데이타 중에서 측정된 공수가 0인 데이타 등 형식상 오류가 있는 데이타를 제거한 후 남겨진 120개를 사용하였다. 그리고, 전체 37개의 변수들 중 전문가의 의견을 기반으로 표 2와 같이 공수에 영향을 미치는 6개의 변수를 선택하였다.

변수들 중 "프로젝트 크기"에 대해 2.5%만큼의 결측치가 있었는데 이것을 보정하기 위하여 ISBSG와 마찬가지로 K-NN보정기법[16]을 사용하였다. 마찬가지로 비율척도를 가진 변수들을 log변환값들을 이용해 정규성을 검정하였고 변수들 사이의 상관관계를 분석 및 K-means 클러스트링 기법을 위한 데이타의 정규화를 수행하였다.

4.3 이상치 제거 기법의 적용 및 데이타군 생성

이 단계에서는 전처리된 데이타에 대해 이상치 제거 기법을 적용시켜 이상치를 제거한다. 특히, 실험결과의 신뢰도를 높이기 위해 K차 교차검증기법(K-fold cross validation)을 사용하고 이를 통해 실험에 사용될 데이타군을 생성한다. K차 교차검증기법은 그림 5와 같이 전체 데이타를 K개의 군으로 나눈 후, K-1 개의 군은 공수 예측 모델의 기반이 되는 데이타 집합, 즉 훈련군(training set)으로 사용하고 나머지 한 개의 군은 공수

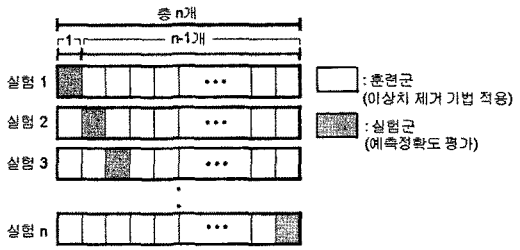


그림 5 n차 교차검증기법의 예

예측 모델의 정확도를 평가하기 위한 실험군(testing set)으로 사용하여 총K번의 실험을 거쳐 얻어지는 공수 예측 정확도를 평균하여 검증하는 기법이다. 이 때, 각 실험군들을 제외한 훈련군들에 대해 이상치 제거 기법을 적용한다.

본 연구에서는 2단계와 3단계에서 10차 교차검증기법을 사용하였다. 그림 6은 ISBSG 데이타에 10차 교차검증 기법을 이용하여 최소절사제곱법과 K-means 클러스터링을 적용한 예이다.

최소절사제곱법을 위한 절사값은 잔차 제곱합이 작은 순에서 큰 순으로 정렬했을 때 사용자가 전체 데이타 개수의 50%에서 100%내의 값을 선정하면 되지만 일반적으로 약 75%정도까지를 절사값으로 이용하기 때문에 [18] 본 연구에서도 75%이후의 절사된 값들을 이상치로 식별하고 제거하였다. K-means 클러스터링에서는 클러

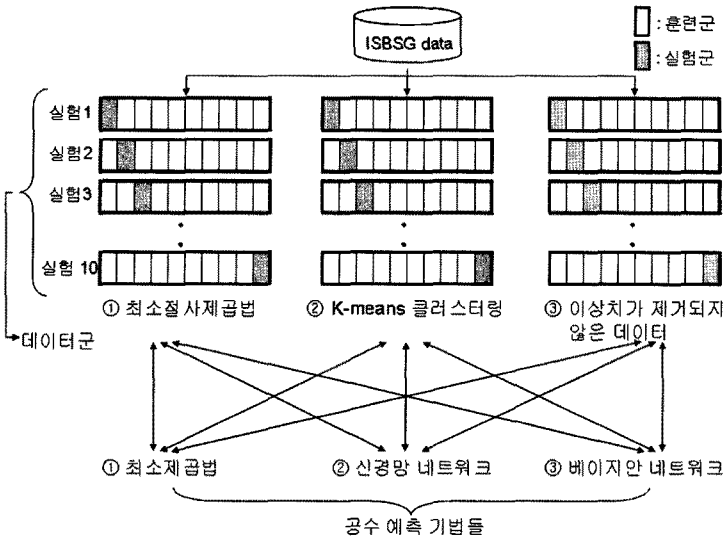


그림 6 교차검증 기법을 이용한 최소절사제곱법과 K-means 클러스터링의 적용

스터링 후 얻어지는 클러스터 내부에 데이타가 하나 또는 둘이거나 0보다 작은 실루엣 값을 가지는 데이타들을 이상치로 식별하고 제거하였다. 표 3은 ISBSG와 Bank 데이타에 대해서 최소절사제곱법과 K-means 클러스터링을 통해 식별된 이상치의 개수 분포이다. 각 기법마다 식별하는 이상치의 개수가 다르고 ISBSG 데이타에 비해 한 조직에서 수집된 Bank data의 이상치가 좀더 적음을 확인할 수 있다.

표 3 이상치 제거 기법을 통해 식별된 이상치 개수

이상치 평균 개수	13.8	5.0	11.1	5.8
이상치 평균 비율	13.94%	5.05%(7.6)	9.25%	4.83%(5.8)

(): 클러스터의 평균 개수

4.4 공수 예측 기법의 적용

그림 6에서 확인할 수 있듯이 이 단계에서는 최소절사제곱법에 의해 이상치가 제거된 10개의 훈련군, K-means 클러스터링에 의해 이상치가 제거된 10개의 훈련군, 그리고 이상치가 제거되지 않은 10개의 훈련군에 대해 공수 예측 기법을 적용시켜 예측 모델을 만들고 실험군을 이용해 각 경우 얻어지는 10개의 예측 정확도를 평균하여 평균 예측 정확도를 구한다. 본 연구에서는 세 가지 공수 예측 기법을 적용하므로 ISBSG와 Bank 데이타에 대해 각각 9개의 평균 예측 정확도가 얻어진다.

최소제곱법을 이용하여 만들어진 모델 중 가장 정확한 공수 예측 모델의 결과는 다음과 같다.

• ISBSG 데이타

$$\log(\text{공수}) = 1.7156 + 0.6186 * \log(\text{프로젝트크기}) + 0.3063 * \log(\text{프로젝트기간}) + (-0.21) * \log(\text{개발형태}) + (-0.0833) * \text{개발플랫폼\_Multi} + 0.1341 * \text{개발플랫폼\_MR} + 0.1178 * \text{언어\_3GL} + (-0.1009) * \text{언어\_4GL}$$

• Bank 데이타

$$\log(\text{공수}) = 0.9222 + 0.1725 * \log(\text{프로젝트크기}) + 0.5314 * \log(\text{프로젝트기간}) + 0.6823 * \log(\text{최대팀크기}) + 0.0398 * \text{개발플랫폼\_}Host + 0.0571 * \text{개발플랫폼\_}Unix + (-0.0093) * \text{개발주기모델\_}V + (-0.0424) * \text{개발주기모델\_}Inc$$

신경망 네트워크는 모델을 만들고 결과값을 얻어내는 과정이 모두 블랙박스 형태로 이루어지기 때문에 만들어진 모델은 알 수 없고 공수 예측 결과만 알 수 있다.

베이지안 네트워크는 그림 7과 같은 공수 예측을 위한 인과관계 모델로서 공수 예측을 수행하였다.

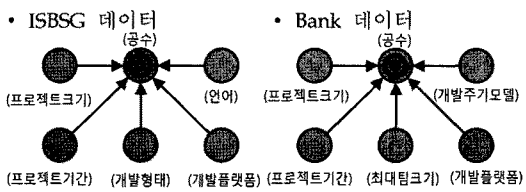


그림 7 베이지안 네트워크를 위한 인과관계 모델

4.5 실험 결과의 비교 분석

4단계에서는 얻어진 공수 예측 정확도를 서로 비교 분석한다. 예측 정확도의 평가기준은 많은 연구 및 문헌에서 주로 사용되는 MMRE[3,7], MdMRE[3], Pred(0.25)[3,19], Pred(0.5)[3,19]를 채택했다. MMRE는 가장 많이 사용되는 평가기준 중 하나이지만, 몇 개의 높은 MRE

값을 가지는 데이타들이 존재하거나 실제값이 작은 데이타를 예측할 경우 값의 변동이 매우 민감하다[20]. MMRE와 MdMRE는 그 값이 낮을수록, Pred(0.25)와 Pred(0.5)는 그 값이 높을수록 공수 예측의 정확도가 더 높다고 평가할 수 있다. 일반적으로 MMRE≤0.25, Pred(0.25)≥0.75정도가 되면 상당히 정확한 예측 결과로 간주된다[21].

5. 실험결과와 검토

5.1 ISBSG 데이타에 대한 결과

표 4는 ISBSG 데이타에 2가지 이상치 제거 기법을 적용하여 이상치를 제거한 데이타들과 그렇지 않은 데이타에 대해 3가지 공수 예측 기법을 적용시켜 얻은 전체 결과이다. 4개의 평가기준에 대한 결과값은 10차 교차기법에 의해 얻어진 10개의 값들을 평균해서 얻은 값이다.

표 상의 네모박스는 각 공수 예측 기법을 사용했을 때 가장 좋은 공수 예측 결과를 보여주는 이상치 제거 기법과의 조합을 나타낸다. ISBSG 데이타의 경우 다양한 도메인으로부터 수집되어 데이타의 분포가 고르지 못하고, 데이타들 사이의 비슷한 특성을 찾아내기가 어렵기 때문에 정확한 공수 예측을 하기는 힘들다. 전체적인 결과에서의 MMRE값 중 가장 낮은 수치도 0.6232이고 가장 높은 Pred(0.25) 값도 0.3556밖에 되지 않는다는 측면에서 공수의 예측 정확도가 전반적으로 좋지 못하다라는 사실을 파악할 수 있다. 그러나 이런 특성을 가진 데이타라 하더라도 이상치 제거 기법을 이용하여 이상치를 제거한 데이타로 공수 예측 모델을 만들었을 때 예측 정확도가 좀더 향상된 것을 확인할 수 있다. 그러므로 이상치 제거는 정확한 공수 예측을 위한 중요한 요소임을 알 수 있다.

표 4 ISBSG 데이타에 대해 이상치 제거 기법들을 적용한 후 만든 공수 예측 모델들의 예측 정확도

사용된 데이터 명 / 이상치 제거기법		ISBSG		
		최소절사 제곱법	K-means 클러스터링	이상치 제거 하지 않은 데이터
공수 예측 기법	평가기준	10차 교차검정의 평균값		
최소 제곱법	MMRE	0.7301	0.7447	0.7417
	MdMRE	0.4485	0.4328	0.4381
	Pred(0.25)	0.2822	0.2611	0.2711
	Pred(0.5)	0.5844	0.5744	0.5744
신경망 네트워크	MMRE	0.6232	0.7164	0.7181
	MdMRE	0.3644	0.3818	0.4378
	Pred(0.25)	0.3556	0.3322	0.2822
	Pred(0.5)	0.6167	0.5967	0.5878
베이지안 네트워크	MMRE	0.9941	1.0468	0.9848
	MdMRE	0.6054	0.5537	0.6044
	Pred(0.25)	0.2222	0.2844	0.2433
	Pred(0.5)	0.4133	0.4656	0.4556

신경망 네트워크는 전체 결과 중에서 4가지 평가기준 모두 각 이상치 기법에 대해 가장 정확한 공수 예측 결과를 보여주고, 다른 세 가지 공수 예측 기법들로부터 얻은 결과들과 비교해 보았을 때 이상치를 제거한 경우와 제거하지 않은 경우 모두 공수 예측 결과가 더 정확하다. 본 연구의 실험에서 신경망 네트워크와 함께 사용했을 때 가장 좋은 결과를 보이는 이상치 제거 기법은 최소절사제곱법이다.

베이지안 네트워크는 다른 공수 예측 기법과 비교하여 이상치를 제거한 경우와 제거하지 않은 경우 모두 4가지 평가기준에서 가장 정확하지 못한 공수 예측 결과를 보여준다. 이상치를 제거한 경우 공수 예측 정확도는 향상하지만 전체적으로 다른 공수 예측 기법들의 결과보다 훨씬 더 좋지 않다. 본 연구의 실험에서 베이지안 네트워크를 사용했을 때 가장 좋은 결과를 보여주는 이상치 제거 기법과의 조합은 K-means 클러스터링이다.

ISBSG에 대한 본 연구의 실험에서 가장 정확한 공수 예측 결과를 보여준 조합은 신경망 네트워크와 최소절사제곱법이다.

5.2 Bank 데이타에 대한 결과

표 5는 Bank 데이타에 대해 5.1절에서와 같은 방식으로 얻은 전체 결과들이다.

Bank 데이타의 경우 ISBSG 데이타와는 달리 하나의 은행 조직으로부터 얻은 데이타이기 때문에 상대적으로 데이타의 분포가 고르고 비슷한 특성을 가지는 프로젝트 데이타들이 더 많이 존재하므로 보다 정확한 공수 예측이 가능하다. 실제로 본 연구의 실험결과를 살펴보면 가장 낮은 MMRE값이 0.2772이고 가장 높은 Pred(0.25) 값이 0.7083이라는 측면에서 상당히 정확한 공수 예측치를 보여준다. Bank 데이타가 전반적으로 정확한 공수 예측 결과를 보여주지만 ISBSG 데이타와 마찬가지로 이상치를 제거한 이후 만든 공수 예측 모델들의 예측 정확도가 좀더 향상됨을 확인할 수 있다. Bank 데이타에서도 이상치 제거는 공수 예측 기법의 예측 정확도를 상승시키는 중요한 요소이다.

최소제곱법은 ISBSG 데이타를 이용했을 때와 비슷한 경향을 보인다. 이상치 제거 후 다른 공수 예측 기법들에 비해 상대적으로 낮은 향상치를 나타낸다. 본 연구의

표 5 Bank 데이타에 대해 이상치 제거 기법들을 적용한 후 만든 공수 예측 모델들의 예측 정확도

사용된 데이터 명 / 이상치 제거기법		Bank		
		최소절사 제곱법	K-means 클러스터링	이상치 제거 하지 않은 데이터
공수 예측 기법	평가기준	10차 교차검정의 평균값		
최소 제곱법	MMRE	0.3183	0.3161	0.3291
	MdMRE	0.1933	0.2051	0.2123
	Pred(0.25)	0.5750	0.5917	0.5500
	Pred(0.5)	0.8333	0.8583	0.8417
신경망 네트워크	MMRE	0.2962	0.2772	0.3052
	MdMRE	0.1774	0.1523	0.1840
	Pred(0.25)	0.6250	0.7083	0.6333
	Pred(0.5)	0.8417	0.8500	0.8417
베이지안 네트워크	MMRE	0.8203	0.5569	0.7584
	MdMRE	0.3811	0.3132	0.3950
	Pred(0.25)	0.3750	0.4167	0.3583
	Pred(0.5)	0.6000	0.6750	0.5667

실험에서 최소제곱법을 사용했을 때 가장 좋은 결과를 보여주는 이상치 제거 기법은 K-means 클러스터링이지만 이 조합 또한 큰 향상도는 보이지 않는다.

신경망 네트워크는 ISBSG 데이타의 결과와 마찬가지로 가장 정확한 공수 예측 결과를 보여준다. 공수 예측의 정확도가 상대적으로 정확함에도 불구하고 이상치 제거 후에는 더 상승한다. 본 연구의 실험에서 신경망 네트워크를 사용했을 때 가장 좋은 결과를 보여주는 이상치 제거 기법과의 조합은 K-means 클러스터링이다.

베이지안 네트워크도 ISBSG와 마찬가지로 4가지 모든 평가기준에서 가장 정확하지 못한 공수 예측 결과를 보여준다. 그러나 이상치를 제거한 후 다른 공수 예측 기법들에 비해 공수 예측 정확도의 상승률은 가장 높다. 본 연구의 실험에서 베이지안 네트워크를 사용했을 때 가장 좋은 결과를 보여주는 이상치 제거 기법과의 조합은 K-means 클러스터링이다.

Bank에 대한 본 연구의 실험에서 가장 정확한 공수 예측 결과를 보여준 조합은 신경망 네트워크와 K-means 클러스터링이다.

5.3 두 데이타의 결과에 대한 검토

이상치 제거기법으로서 K-means 클러스터링의 적용 결과는 ISBSG 데이타에서 그리 좋지 못하다. 그 원인은 ISBSG 데이타가 분산이 크다는 특징 때문에 정확한 클러스터링 결과와 이를 통한 이상치 검출이 용이하지 않았기 때문이다. 표 6은 ISBSG와 Bank 데이타에서 10개의 훈련군에 대해 K-means 클러스터링을 적용할 때 사용되었던 평균 실루엣 값들을 보이고 있다. ISBSG 데이타의 평균 실루엣 값은 대체적으로 Bank

표 6 ISBSG 데이타와 Bank 데이타에 사용된 평균 실루엣 값

평균 실루엣 값	
ISBSG 데이타	Bank 데이타
0.5261	0.7336
0.5294	0.7525
0.5162	0.7599
0.5609	0.7725
0.5338	0.749
0.5231	0.7424
0.5354	0.7673
0.5754	0.7581
0.5353	0.7567
0.522	0.7873

데이타에 비해 많이 낮은 편이다. 평균 실루엣 값이 낮다는 것은 데이타들이 올바른 클러스터를 형성하지 못했다는 것으로 해석할 수 있는데 이것은 결국 이상치로 식별되어야 하는 데이타들이 이상치로 식별되지 못하고 공수 예측 모델을 만들 때 사용되었다는 것을 의미한다. 결과적으로 이는 공수 예측 모델을 만드는데 악영향을 끼쳐 예측정확도를 저하시킬 수 있다.

본 연구의 실험결과를 통해 알 수 있듯이 이렇게 분산이 크고 유사한 특성을 가진 데이타가 아닌 경우에는 최소절사제곱법이 공수 예측 정확도를 높이는데 더 효과적이다. 서로 관련성이 적고 많이 흩어져 있는 데이타는 최소절사제곱법처럼 통계적으로 미리 정해진 절사값이 있고 그 값을 넘어버리면 바로 그 프로젝트 데이타는 이상치로 식별하고 제거시킴으로써 특성이 매우 다른 프로젝트 데이타들이 모여있다고 하더라도 K-means 클러스터링에 비해 비슷한 성향을 가진 데이타들이 모일 수가 있다. 이것은 곧 공수 예측 모델을 만드는데도 영향을 끼쳐 좀더 정확한 공수 예측을 가능하게 할 수 있다. 베이지안 네트워크는 공수 예측의 기법상 이런 특성에 큰 영향을 받지 않아서 최소절사제곱법보다는 K-means 클러스터링이 좀더 좋은 결과가 나온 것으로 사료된다.

반면에 Bank 데이타는 최소절사제곱법보다는 세가지 공수 예측 기법 모두 K-means 클러스터링을 사용했을 때 더 좋은 결과를 보여준다. 이 데이타는 ISBSG 데이타와는 달리 상대적으로 유사한 특성을 가지는 프로젝트 데이타들이 잘 모여있고 분산도가 적다. 그렇기 때문에 클러스터링을 하면 ISBSG 데이타에 비해 상대적으로 특성이 다른 데이타들이 다른 클러스터로 잘 구분되어서 좀더 명확히 이상치를 식별할 수 있고 이것은 공수 예측에 도움을 줄 수 있다. 표 6의 평균 실루엣 값들을 살펴보아도 좀더 나은 클러스터들이 만들어졌음을

확인할 수 있다. 이에 반해, 유사한 데이타가 잘 모여있는 데이타 분포일수록 최소절사제곱법의 이상치 식별 능력은 떨어진다.

그 이유는 만약 이상치가 없는 데이타 분포라고 하더라도 이 기법의 특성상 절사값을 넘는 데이타는 반드시 이상치로 식별하기 때문에 실제 이상치가 아닌 데이타들도 이상치로 식별하고 제거한다. 결국 이상치제거를 하더라도 오히려 공수 예측 정확도를 저하될 수 있다.

세 가지 공수 예측 기법 중 가장 예상치 못했던 결과를 보인 것은 최소제곱법이다. 이 기법은 일반적으로 이상치에 상당히 민감한 기법으로 알려져 있음에도 불구하고, 본 연구의 실험결과에서는 이상치를 제거한 데이타와 그렇지 않은 데이타를 이용한 공수 예측 모델의 예측 결과 차이가 그리 크지 않았다. 본 연구에서는 훈련군에 대해 이상치 제거를 한 후 공수 예측 모델을 만들었기 때문에 훈련군에 대한 공수 예측 정확도는 상당히 높지만, 실험군에 있는 이상치들에 대해서는 민감하게 반응하여 공수 예측 정확도가 오히려 떨어지는 이상치값들이 측정된다. 그림 8은 ISBSG 데이타에서 최소제곱법과 최소절사제곱법의 조합과 최소제곱법과 이상치 제거하지 않은 데이타와의 조합에 대해 10개 실험군에 대한 MRE값들의 상자그림을 보여준다. 그림상의 십자표시는 각 실험군에 대한 이상치값들을 나타낸다. 최소제곱법과 최소절사제곱법의 조합에 대해서 이 모든 이상치값들에 대한 MRE값의 평균값과 중앙값을 계산해 본 결과 최소제곱법과 이상치 제거하지 않은 데이타와의 조합에 대한 값들보다 더 좋지 못한 결과를 보여준다는 것을 표 7에서 확인해볼 수 있다. 결국 본 연구에서는 실험군에서 얻은 공수 예측 정확도들을 평균한 값을 이용하기 때문에 전체적으로는 이상치를 제거하지 않고 만든 공수 예측 모델과 비슷한 공수 예측 정확도를 나타낸다.

신경망 네트워크는 ISBSG와 Bank 데이타 모두에서 가장 높은 공수 예측 정확도를 보여주었으나 이상치 제거 기법의 조합의 성능에 대한 결과는 달랐다. ISBSG와 Bank 데이타에서 각각 최소절사제곱법과 K-means 클러스터링과의 조합이 가장 정확한 결과를 보여주었다. 신경망 네트워크는 프로젝트 데이타의 학습을 기반으로 하기 때문에 비슷하거나 중복된 데이타가 많이 학습될수록 더 정확한 예측 결과를 나타낸다. 그런 측면에서 ISBSG 데이타에서는 최소절사제곱법을, 그리고 Bank 데이타에서는 K-means 클러스터링이 앞에서 언급한 것처럼 가장 유사한 데이타들을 상대적으로 잘 선정했기 때문에 가장 최적인 공수 예측 결과를 보인 것이라 사료된다.

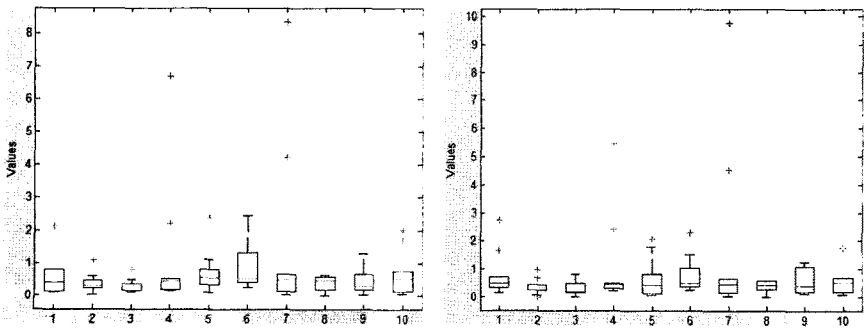
표 7 ISBSG 데이타에서 10개의 실험군에 대한 이상치 MRE값들의 평균과 중앙값

	최소제곱법과 최소절사제곱법	최소제곱법과 이상치 제거하지 않은 데이타
평균값	3.0672	2.7770
중앙값	2.163	2.08

베이지안 네트워크는 ISBSG와 Bank 데이타 모두 K-means 클러스터링 기법을 이용하는 것이 최소절사제곱법을 이용하는 것보다 더 정확한 공수 예측 결과를 보여주었다. 이것은 공수 예측 기법 특성상 K-means 클러스터링을 통해 이상치 제거를 한 데이타들이 인과관계 모델의 정확한 결합확률분포를 구하는데 도움을 주어 공수 예측 정확도를 높여 주었음을 의미한다.

6. 결론 및 향후 연구

일반적으로 수집된 프로젝트 데이타 내에는 이상치들이 존재한다. 이상치들이 많이 존재하는 데이타를 이용해 공수 예측 모델을 만든다면 앞으로 수행할 프로젝트에 대한 정확한 공수 예측은 어려울 것이다. 그러므로 본 연구에서는 두 가지 프로젝트 데이타에 대해 이상치



(최소제곱법과 최소절사제곱법) (최소제곱법과 이상치 제거하지 않은 데이타)

그림 8 ISBSG 데이타에서 10개의 실험군에 대한 MRE 값의 상자그림

제거 기법들을 적용한 후 공수 예측 모델들의 예측 정확도를 알아보았고, 다른 이상치 제거 기법과 공수 예측 기법들이 다른 결과를 보여주는 이유에 대해서도 살펴보았다. 실험을 수행해 본 결과 이상치 제거 기법을 이용해 이상치를 제거한 후 만든 공수 예측 모델들이 그렇지 않은 모델들보다 더 정확한 공수 예측 결과를 보여주었고 다른 분포를 가지는 두 가지 프로젝트 데이타에 대해 얻어진 공수 예측 결과의 차이도 확인할 수 있었다.

향후 연구로는 좀더 의미 있는 결과를 얻기 위해 다른 소프트웨어 프로젝트 데이타들을 추가하고, 이상치 제거기법으로 CART(Classificaiton And Regression Tree)를 적용할 예정이다. 또한 데이타 분포, 공수 예측 기법, 그리고 이상치 제거 기법들 사이의 관계에 대한 구체적인 분석도 수행할 예정이다.

참 고 문 헌

[1] C. van Koten and A.R. Gray, "Bayesian Statistical Effort Prediction Models for Data-centred 4GL software development," Information and Software Technology, Vol.48, No.11, pp. 1056-1067, 2006.

[2] A. Gray and S. MacDonell, "Application of Fuzzy Logic to Software Metric Models for Development Effort Estimation," Annual Meeting of the North American Fuzzy Information Processing Society, pp. 394-399, 1997.

[3] M. Jorgensen, "Experience With the Accuracy of Software Maintenance Task Effort Prediction Models," IEEE Transactions on Software Engineering, Vol.21, No.8, pp. 674-681, 1995.

[4] E. Mendes, C. Lokan, R. Harrison, and C. Triggs, "A Replicated Comparison of Cross-company and Within-company Effort Estimation Models using the ISBSG Database," 11th IEEE International Software Metrics Symposium, 2005.

[5] P.C. Pendharkar, G.H. Subramanian, and J.A. Rodger, "A Probabilistic Model for Predicting Software Development Effort," IEEE Transactions on Software Engineering, Vol.31, No.7, pp. 615-624, 2005.

[6] V.K.Y. Chan and W.E. Wong, "Outlier Elimination in Construction of Software Metric Models," Proceedings of the 22nd ACM Symposium on Applied Computing, pp. 1484-1488, 2007.

[7] A.R. Gray and S.G. MacDonell, "A Comparison of Techniques for Developing Predictive Models of Software Metrics," Information and Software Technology, Vol.39, No.6, pp. 425-437, 1997.

[8] J. Heaton, Introduction to Neural Networks with Java, Chesterfield, MO : Heaton Research, Inc, 2005.

[9] S. Chulani, B. Boehm, and B. Steece, "Bayesian Analysis of Empirical Software Engineering Cost Models," IEEE Transactions on Software Engineering, Vol.25, No.4, pp. 573-583, 1999.

[10] P.J. Rousseeuw, "Multivariate Estimation with High Breakdown Point," Mathematical Statistics and Applications, pp. 283-297, 1985.

[11] P.J. Rousseeuw and A.M. Leroy, Robust Regression and Outlier Detection, NY : John Wiley & Sons, Inc, 1987.

[12] A.K. Jain, M.N. Murty, and P.J. Flynn, "Data Clustering: A Review," ACM Computing Surveys, Vol.31, No.3, pp. 264-323, 1999.

[13] S. Lamrous and M. Taileb, "Divisive Hierarchical K-means," International Conference on Intelligent Agents, Web Technologies and Internet Commerce, 2006.

[14] International Software Benchmarking Standards Group, http://www.isbsg.org, 2005.

[15] P.J. Rousseeuw, "Least Median Squares Regression," Journal of American Statistical Association, Vol.79, No.388, pp. 871-880, 1984.

[16] Q. Song and M. Shepperd, "A new imputation method for small software project data sets," Journal of Systems and Software, Vol.80, No.1, pp. 51-62, 2007.

[17] M.Mendes and A.Pala, "Type I Error Rate and Power of Three Normality Tests," Pakistan Journal of Information and Technology, Vol.2, No.2, pp. 135-139, 2003.

[18] P.J. Rousseeuw and K. van Driessen, "Computing LTS Regression for Large Data Sets," Data Mining and Knowledge Discovery, Vol.12, No.1, pp. 29-45, 2006.

[19] B. Kitchenham, S.G. MacDonell, L. Pickard, and M.J. Shepperd, "Assessing Prediction Systems," The Information Science Discussion Paper Series, University of Otago, 1999.

[20] T. Foss, E. Stensrud, B. Kitchenham, and I. Myrtveit, "A Simulation Study of the Model Evaluation Criterion MMRE," IEEE Transactions on Software Engineering, Vol.29, No.11, pp. 985-995, 2003.

[21] S.D. Conte, H.E. Dunsmore, and V.Y. Shen, Software Engineering Metrics and Models. Benjamin/Cummings Publishing Company, 1986.

서 영 석

2006년 숭실대학교 컴퓨터학부 졸업(학사). 2006년~2008년 KAIST 전산학과 졸업(석사). 2008년~KAIST 전산학과 박사과정. 관심분야는 소프트웨어 프로세스, 소프트웨어 측정 및 분석, 소프트웨어 비용산정

윤 경 아

1996년 동국대학교 컴퓨터공학과 졸업(학사). 1996년~2000년 삼성SDS 솔루션 사업부. 2003년 KAIST 전산학과 졸업(석사). 2003년~KAIST 전산학과 박사과정. 관심분야는 소프트웨어 측정 및 분석, 소프트웨어 데이타 품질, Empirical Software Engineering

배 두 환

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