Journal of Korean Institute of Intelligent Systems, Vol. 22, No. 5, October 2012, pp. 624-630
http://dx.doi.org/10.5391/JKIIS.2012.22.5.624

# 효율적인 피어리뷰 학습을 위한 회귀 모델 기반 학습성과 예측 방법

# A Prediction Method of Learning Outcomes based on Regression Model for Effective Peer Review Learning

신효정* · 정혜욱** · 조광수*** · 이지형**†

Hyojoung Shin, Hye-Wuk Jung, Kwangsu Cho, and Jee-Hyoung Lee+

*삼성전자

**성균관대학교 컴퓨터공학과

***성균관대학교 인터렉션사이언스학과

**요 약**

피어리뷰(peer review)를 통한 학습은 학습자간 피드백을 주고받으며 다양한 정보를 관찰, 분석하는 과정을 통해 학습성과를 향상시키는 방법이다. 피어리뷰 시스템의 중요한 문제 중 하나는, 학습자의 여러 특징을 고려하여 학습자의 학습성과를 향상시키는데 적합한 평가자를 찾는 것이다. 그러나 기존 피어리뷰 시스템에서는 학습자들이 가지는 다양한 특징을 고려하지 않고 단순히 피어리뷰 평가자를 임의로 할당하거나 제한적인 학습 전략에 따라 피어리뷰 평가자를 편성하였다. 본 논문에서는 학습자와 평가자의 다양한 특징을 고려하여, 특정 학습자와 평가자의 조합으로 피어리뷰 학습이 이루어졌을 때 학습자에게 어느 정도의 학습성과 향상이 있을지 예측하는 방법을 제안한다. 제안하는 방법은 학습자와 평가자의 프로파일 정보로부터 대표 속성을 추출하고 다양한 회귀 모델을 적용하였다. 또한 학습자들의 다양한 특징으로 인하여 나타날 수 있는 이상치(outlier)가 학습성과 예측에 미치는 영향을 알아보기 위해, 회귀 모델에 다양한 이상치 제거 방법을 적용하여 학습성과 예측 성능을 비교하였다. 실험 결과 이상치를 제거 하지 않은 SVR 모델이 평균 0.47%의 에러율을 보이며 가장 우수한 학습성과 예측결과를 보였다.

**키워드** : 피어리뷰, 회귀모델, SVR, 이상치, 이러닝

**Abstract**

The peer review learning is a method which improves learning outcome of students through feedback between students and the observation and analysis of other students. One of the important problems in a peer review system is to find proper evaluators to each learner considering characteristics of students for improving learning outcomes. Some of peer review systems randomly assign peer review evaluators to learners, or chose evaluators based on limited strategies. However, these systems have a problem that they do not consider various characteristics of learners and evaluators who participate in peer reviews. In this paper, we propose a novel prediction approach of learning outcomes to apply peer review systems considering various characteristics of learners and evaluators. The proposed approach extracts representative attributes from the profiles of students and predicts learning outcomes using various regression models. In order to verify how much outliers affect on the prediction of learning outcomes, we also apply several outlier removal methods to the regression models and compare the predictive performance of learning outcomes. The experiment result says that the SVR model which does not removes outliers shows an error rate of 0.47% on average and has the best predictive performance.

**Key Words** : Peer review, Regression model, SVR, Outlier, e-learning

## 1. 서 론

피어리뷰를 통한 학습방법은 기존 교수자 중심 평가와는 달리 학습자들이 피드백을 주고받는 상호작용을 통해 학습 능력을 향상시킬 수 있는 교육방법이다[1]. 피어리뷰에 참여하는 학습자는 다양한 학습 특징을 가지고 있다. 예를들어 글쓰기 학습에 참여한 학습자들의 논리적인 작문 능력이나 동료가 작성한 글을 평가하는 능력은 학습자에 따라 다르다. 또한 동료가 작성한 글에 대한 평가 의견(comment)가 비판적인 성향을 보일 수 있고 이와는 반대로 긍정적인 표현을 통해 의견을 제시하기도 한다. 이와같이 피어리뷰 과정

접수일자: 2012년 6월 28일
심사(수정)일자: 2012년 9월 20일
게재확정일자 : 2012년 9월 26일
† 교신 저자
이 논문(저서)은 2010년도 정부재원(교육과학기술부 인문사회연구역량강화사업비)으로 한국연구재단의 지원을 받아 연구되었음(NRF-2010-32A-H00011)

에서 관찰되는 학습자의 특징은 학습성과에 중요한 역할을 한다. 따라서 어떠한 학습 특징을 가진 동료들끼리 피어리뷰를 수행하였는가에 따라 학습성과가 달라 질 수 있기 때문에 최적의 학습성과를 낼 수 있는 학습자들을 구성해주는 과정이 필요하다.

Liu et al.[2]는 컴퓨터공학 전공 대학생을 대상으로 전공과목의 학습을 위한 웹기반 피어리뷰 방법을 제안하였다. 이 방법은 한 학생당 6명의 평가자를 임의로 할당하여 단계별로 피어리뷰를 실행한 후 평가척도에 따라 학습성과를 측정하는 방법을 제시하였지만, 피어리뷰 구성원을 편성하는 과정에서 학습자의 특성, 학습자들 사이의 상호작용 사이에 발생하는 관계에 대해 고려하지 않았다.

Crespo et al.[3]는 학습자 프로파일 정보를 이용하여 피어리뷰 학습자의 프로토타입을 설계하고 이를 기준으로 피어리뷰 구성원을 찾았다. 그러나 제한적인 프로토타입 기준을 적용하기 때문에 보다 다양한 학습 특징을 가지는 학습자들에게 적용하기 어렵다.

Gehringer[4]는 피어리뷰 환경의 학습자 특징에 따라 정의된 리뷰어 할당 전략 세우고 이를 이용하여 피어리뷰 학습자들을 구성하였다. 그러나 피어리뷰에 참여한 학습자들의 학습 특징을 고려하지 않는 문제가 있다.

본 논문에서는 보다 효율적인 피어리뷰 시스템 구축을 위해 피어리뷰 학습자들의 학습성과를 회귀 모델을 통해 예측해 보았다. 기존 글쓰기 학습의 피어리뷰 정보를 이용하여 학습자 프로파일 생성 및 대표 속성을 추출한 후 다양한 회귀 모델을 이용하여 학습성과 예측 성능을 평가 하였다. 또한 학습자들의 다양한 특징에서 나타날 수 있는 극단 값이 학습성과에 미치는 영향을 알아보기 위해, 가장 낮은 평균 에러율을 보인 SVR(Support Vector Regression) 모델에 다양한 이상치 제거방법을 적용하여 실험한 결과 이상치를 제거하지 않은 SVR이 가장 높은 예측력을 보였다.

본 논문의 구성은 2장에서 피어리뷰 과정을 통해 수집된 원본 데이터를 가공하여 학습자 프로파일을 생성하고 특징 선택(feature selection) 방법을 이용하여 대표 특징을 추출하는 과정을 설명 하였다. 3장에서는 앞서 추출한 대표속성을 다양한 회귀 모델을 적용하여 예측 성능 결과를 비교 분석 하였다. 4장에서는 가장 좋은 예측 성능을 보이는 SVR 모델에 이상치 제거방법을 적용한 결과를 분석한 후 5장에서는 결론 및 향후 연구 과제를 제시하였다.

## 2. 피어리뷰 학습자의 프로파일 생성 및 대표속성 추출

본 연구는 글쓰기 학습의 피어리뷰 과정에서 수집된 원본 데이터를 가공하여 학습자의 프로파일 정보를 생성하고, 생성된 프로파일에 특징 선택 방법을 이용하여 대표속성을 추출 하였다.

### 2.1 피어리뷰 학습자의 프로파일 생성

학습자의 프로파일 정보는 글쓰기 과제를 수행하는 과정에서 수집된 원본 데이터를 사용하여 생성 하였다. 과제에 참가한 학습자는 작성자(writer)와 평가자(reviewer) 역할을 번갈아 하게 된다.(작성자:1번, 평가자: 3~4번) 과제 문서 평가 수행을 위한 평가자는 한 명의 작성자에게 3~4명을 랜덤으로 할당하는 방식으로 진행된다[5]. 그림 1은 글쓰기 학습의 피어리뷰 과정으로 작성자가 (1차)과제 문서를 작성하여 평가자에게 전송하면 평가자는 (1차)과제 문서를 평가하여 작성자에게 전송한다. 작성자는 평가자가 제시한 평가의견에 따라 (2차)과제 문서를 작성(수정작업 포함)하고, 평가자를 평가 한 후 평가자에게 전송한다. 마지막으로 평가자는 (2차)과제 문서를 평가 하면서 피어리뷰 과정을 마친다. 이때, 작성자의 작문능력(WS), 평가자의 평가능력(RS), 평가자의 평가점수(ES)는 7점 척도로 측정하였고, 평가자의 평가의견(WC)는 각 의견이 언급하는 내용에 따라 35개의 개념(idea)으로 분류하여 정리하였다. 예를 들어 평가의견이 문서의 구조와 관련되어 있다면 평가의견의 개념은 “organization”이 되고, 평가의견이 “문서가 잘 구성되었다”이면 평가의견은 최종적으로 “good_organization"으로, 그렇지 않으면 "poor_organization"으로 분류된다. 이러한 개념에는 "organization", "surface", "detection", "diagnosis" 등이 있다.

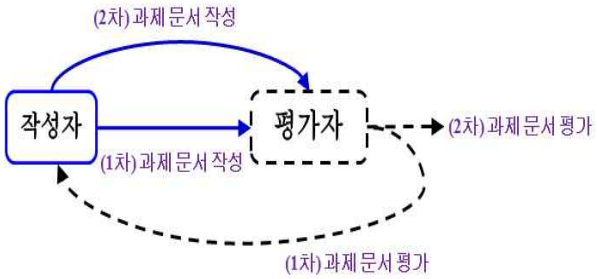


그림 1. 글쓰기 학습의 피어리뷰 과정
Fig. 1. The peer review process of writing learning

표 1. 동료 평가 학습자의 프로파일
Table 1. A profile of peer review learner

| 속성 종류 | 설 명 |
|---|---|
| WS | 작성자의 과제 문서를 평가자가 평가한 점수 |
| RS | 평가자가 작성한 평가의견을 작성자가 평가한 점수 |
| LO | 글쓰기 능력 점수의 변화량(평가 전/후) |
| ES | 과제 문서에 대한 평가자의 평가점수 |
| WC | 과제 문서에 대하여 평가자가 작성한 평가의견 |

이러한 학습자의 피어리뷰 속성을 통계적인 기법을 이용하여 프로파일 정보로 가공한 항목은 표 1과 같다. 선행 연구를 통해 수집된 원본 데이터는 크게 3가지(작성자의 과제 문서를 평가자가 평가한 점수, 평가자가 작성한 평가의견을 작성자가 평가한 점수, 전문가가 측정한 평가자의 평가의견에 대한 속성)로 나누어진다. 평가의견에 대해서는 평가의견의 길이와 속성으로 나누고, 1차와 2차 과제 문서 평가 점수의 변화량을 학습성과 향상 점수로 정의하여 학습자 프로파일 속성으로 사용 하였다.

이와 같은 과정은 원본 데이터로 부터 프로파일 정보를 생성한 것으로 학생 개인의 특성을 판단하는 지표가 될 수 있다.

### 2.2 대표 특징 추출

피어리뷰 학습자의 학습성과를 향상 시킬 수 있는 평가자 예측을 위해서는 앞서 생성한 학습자 프로파일 정보인 75개 속성 중 학습성과에 영향력을 미치는 즉, 높은 상관관계를 가지는 대표 속성 추출이 필요하다. 피어리뷰 과정에서 나타나는 다양한 속성 중 모든 학생에게 발생하는 빈도가 높거나 낮은 경우, 특정 학습자에게 나타나지만 학습성과에 연관성이 낮은 잡음 속성이 존재하기도 한다. 이러한 요소들은 학습자의 특성을 파악하는데 유용하지 않다. 또한 피어리뷰 학습자의 학습성과 예측을 위한 모델 생성에 적절하게 사용 될 수 있는지 확신하기 어렵다. 따라서 본 연구에서는 WEKA[6]를 사용하여 다양한 특징 선택 방법으로 대표 속성을 추출 하였다.

각 학습자의 속성은 작성자의 작문능력(WS), 평가자의 평가능력(RS), 평가자의 평가점수(ES), 평가자의 평가의견(WC)으로 구성되어 있다.

대표 특징을 추출에는 피어리뷰를 수행한 54명 학습자의 모든 속성을 이용하였고 특징선택(Feature Selection) 방법은 Relief와 CFS(Correlation based Feature Selection)을 사용 하였다. Relief를 이용한 속성 선택 방법은 인스턴스 기반의 학습을 이용하여 속성의 연관성을 평가하는 방법이다. 이 방법은 임의의 샘플 속성과 같은 클래스 그리고 서로 다른 클래스에서 가장 유사도가 높은 후보 속성을 탐색하여 후보 속성의 부분 집합을 생성한다. 이렇게 생성된 후보 속성의 부분집합들 중 가장 우수한 속성을 추출하기 위해 각 속성에 가중치를 부여하여 적합도를 계산한다. 예를 들어, 사전에 임의의 가중치를 각 속성에 부여하여 분류기를 학습 시키고 분류 성능에 따라 가중치를 증가 또는 감소시켜 조정한다. 이와같이 Relief는 분류하고자 하는 클래스에 대한 속성 값들의 차이를 구하고 가중치를 변경해 가며 연관성이 높은 속성을 찾는 방법이다[7].

Correlation based Feature Selection(CFS)는 사용 가능한 $n$개의 특징 집합에 대한 가능한 모든 조합을 고려하여 최적의 조합을 찾아내는 방법이다. 즉, 임의의 클래스에서 $k$개의 특징들에 대한 조합을 위해 공분산 및 상관계수를 이용하여 특징 상호간의 상관계수를 구한다. 이렇게 산출된 특징은 한 개의 클래스와는 연관성이 높고, 다른 클래스와는 연관성이 낮게 나타난다. 이 방법은 속성의 부분 집합 선택과정에서 모든 하위 집합의 상관관계를 계산하는데 많은 시간이 소요되기 때문에 $n^2$의 조합으로 특징 집합을 생성 하여 탐색시간을 줄일 수 있는 다양한 검색 방법이 사용된다[8]. 그림 2는 입력된 데이터에 대해 특징추출 방법을 이용하여 각 속성의 부분집합을 평가 및 탐색하여 최적의 대표 특징을 선택하는 특징 선택 처리 과정을 나타낸다.

본 논문에서는 CFS에 bestfirst, genetic search, greedy stepwise을 이용한 경험 기반 탐색 방법을 적용하였다. bestfirst는 유망한 노드를 탐색하는 방법으로 greedy(Hill Climbing)와 제한적인 backtracking을 통해 속성의 부분집합 공간을 검색하고 비어있는 부분 집합으로부터 시작하여 하나씩 속성을 추가하며 평가한다. Genetic search는 유전 알고리즘을 기반으로 랜덤하게 속성 집합을 선택 후 연관성을 평가하고 최적의 속성을 찾아낸다.

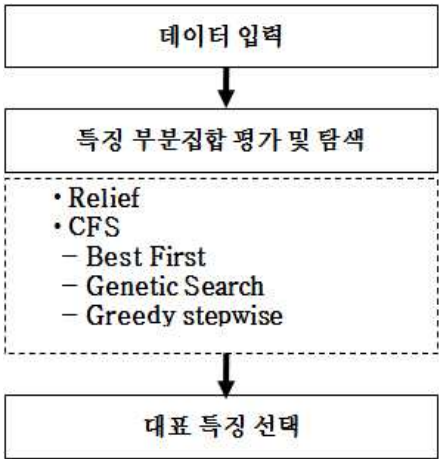


그림 2. 특징 선택 처리 과정
Fig. 2. A feature selection process

표 2. 특징 선택 결과(작성자)
Table 2. The result of feature selection(writer)

| 방 법 | 작성자의 특징 |
|---|---|
| Relief | **Diagnosis**, Organization_Macro, **Good_Surface**, Good_MechanicsExpression, Good_Organization, rs score |
| CFS (BestFirst) | **Diagnosis**, Others, **Good_Surface**, Good_MechanicsExpression |
| CFS(Genetic search) | **Diagnosis**, Expression, Organization_Micro, Development_Micro, Development_Macro, **Good_Surface**, Good_Macro, Poor_Surface, Good_MechanicsExpression, Good_Organization, Good_Supporting(Set3) |
| CFS(Greedy stepwise) | **Diagnosis**, **Good_Surface** |

Greedy stepwise는 연관성이 높은 속성을 단계적으로 선택하는 방식이다. BestFirst에서와 같이 속성을 하나씩 빼거나 추가하는 방법으로 검색하지만 백트래킹(backtracking) 과정이 생략된 방법으로 속성의 중요성을 판단하여 중요한 속성만 남겨 놓고 중요하지 않는 속성은 제거하는 방법이다. 이와같은 방법을 통해 추출된 속성 중 작성자의 정보로부터 추출된 결과는 표 2와 같다. Relief 방법으로 추출된 rs score(작성자의 리뷰능력 점수)를 제외한 모든 속성이 작성자가 피어리뷰 수행시 작성한 평가의견에 존재하는 항목이었다. 이중 각 방법에서 공통적으로 추출된 속성은 “Diagnosis”와 “Good_surface”였고 CFS(Greedy stepwise)의 경우 최종적으로 연관성이 높은 속성 두 개만 선택되었다.

표 3. 특징 선택 결과(평가자)
Table 3. The result of feature selection(reviewer)

| 방 법 | 평가자의 특징 |
|---|---|
| Relief | Validity_Macro, Supportive_Materials, **Good_Surface**, Good_Micro, Poor_Supporting(Set3), Good_MechanicsExpression, **rs score**, **Good_Organization**, Poor_Supporting(Set4), |
| CFS (BestFirst) | Expression, Organization_Micro, Focus_Micro, **Good_Surface**, Good_Micro, ws score, Good_MechanicsExpression, **Good_Organization**, **rs score** |
| CFS(Genetic search) | Diagnosis, Focus_Macro, Organization_Micro, Validity_Macro, Others, **Good_Surface**, Good_Micro, Good_Macro, Poor_Macro, Poor_Supporting(Set3), **Good_Organization**, Good_FocusDevelopmentValidity, Good_Supporting(Set3), Poor_Organization, **rs score**, Poor_Supporting(Set4), ws score |
| CFS(Greedy stepwise) | Organization_Micro, Focus_Micro, **Good_Surface**, **Good_Organization**, **rs score** |

표 3은 특징 선택 결과 중 평가자 정보로부터 추출된 대표 속성이다. 이것은 총 40개로 작성자 정보로부터 추출된 23개의 속성보다 많은 분포를 보였다. 또한 4가지 방법으로부터 “Good_Surface”, “Good_Organization”, “rs score” 속성이 공통적으로 추출되었다.

## 3. 회귀 모델을 이용한 피어 매칭 예측의 실험적 비교

회귀 모델은 독립 변수들의 연관관계를 통해 종속변수를 예측하기 위해 일반적으로 사용되는 방법으로 이러닝(e-Learning) 환경에서 교수와 학습자간 상호작용 분석, 학습자들의 학습성과 분석 등에 다양하게 적용된다. 본 논문에서는 평가자에 따른 학습자들의 학습성과를 분석하기 위해 선형 회귀, 다중 회귀, 회귀나무, 서포트 벡터 회귀 모델을 대상으로 예측 성능을 비교해 본다.

### 3.1 회귀 모델

선형 회귀(Linear Regression, LR)는 $X$ 변수를 이용해 $Y$ 변수를 예측하기 위한 최적의 선형 함수를 찾는 것을 목표로 한다. 즉, 종속 변수 $X$는 $Y$ 변수를 추론하기 위한 정보로 사용된다. 선형 회귀는 식(1)과 같이 종속 변수들을 선형 결합으로 나타낼 수 있다[9].

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_iX_i+\epsilon,\quad i=1,2,3,\cdots,n \tag{1}$$

다중 회귀(Polynomial Regression, PR)는 선형 회귀와 비슷한 개념이지만, $Y$ 변수와 $X$ 변수들의 선형 관계가 아니다. 또한 종속 변수들은 어떤 차수로도 표현이 가능하다. 다중 회귀 모델은 식 (2)와 같이 표현된다[10].

$$Y=\beta_0+\beta_1X_1+\beta_2X_1^2+\beta_3X_1^3+\cdots+\beta_{3i-2}X_i+\beta_{3i-1}X_i^2+\beta_{3i}X_i^3+\epsilon,\quad i=1,2,3,\cdots,n \tag{2}$$

회귀 나무(Regression Tree, RT)는 모든 학습샘플을 포함하고 있는 최상위 노드에서 시작하여 학습 데이터 변위의 평균값을 기준으로 부모노드에서 자식노드로 이진 분할하는 반복 수행을 거쳐 최종 노드에 이르게 된다[11].

서포트 벡터 회귀(Support Vector Regression, SVR)는 두 부류의 클래스를 갖는 데이터를 분류하는 SVM(Support Vector Machine) 이론을 선형 회귀 문제에 확장하여 적용한 방법이다. $x$는 입력벡터, $y$는 출력 벡터인 $n$개 학습 데이터가 존재 할 때, 입력과 출력에 대한 에러가 가장 작은 식(3)과 같은 함수 $f(x)$를 찾는 원리로 동작한다. 여기서 $w$는 가중치 벡터, $b$는 바이어스 이다. 또한 비선형 회귀를 하기 위해 커널 함수를 사용한다[12].

$$f(x)=w\cdot x+b \tag{3}$$

### 3.2 회귀 모델의 성능

본 절에서는 학습성과 예측 위한 다양한 회귀 모델의 성능을 평가 및 분석 한다. 회귀 모델 LR, PR, RT는 SPSS사의 PASW Statistics 18[13]을 이용하여 생성하였고, SVR 모델은 LibSVM[14]을 사용하였다.

표 4는 피어리뷰 학습자 54명에 대해 4가지 특징 선택 방법으로 추출된 속성들을 4개의 회귀 모델에 적용하여 학습성과 에러율의 평균(mean) 및 표준편차(stdev)를 산출한 결과이다. 모든 특징은 특징 선택 수행 전 모든 속성을 포함하여 실험한 결과로 4가지 특징 선택 방법의 결과와 비교하기 위한 기준치로 사용하였다. 각 모델의 총 평균과 표준편차(Total(Avg))를 산출한 결과, SVR이 0.47%의 가장 낮은 에러율과 0.41%의 표준편차를 보였다. 각 특징 선택의 Total(Avg)의 경우, CFS(Greedy stepwise) 방법이 0.57%의 평균 에러율로 가장 좋은 성능을 보이고 있으나, RT의 경우 CFS(Genetic search) 방법이 0.56%로 평균보다 더 낮은 에러율을 보였다. 또한, SVR 모델에서는 4가지 특징 선택 방법이 모두 평균보다 낮은 에러율을 보이며 좋은 성능을 보였다.

PR 모델은 다른 모델에 비해 평균 에러율과 표준편차가 2배 이상 높은 수치를 보였다. 이와 같은 결과의 원인은 PR이 단순한 곱셈을 이용하여 독립 변수 $Xs$를 회귀하기 때문에 $Xs$의 수가 증가하고 이는 중복 변수가 종속 변수에 연관되는 것을 보장하지 못한다. 결과적으로 PR은 학습성과를 예측하는 모델로 사용하기에는 부적합하다. 결론적으로

표 4. 선택된 feature를 이용한 각 회귀 모델의 성능 비교
Table 4. Performance comparison of regression models using selected features

| Regression models / 특징 선택 | LR | | PR | | RT | | SVR | | Total(Avg) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | mean | stdev | mean | stdev | mean | stdev | mean | stdev | mean | stdev |
| 모든 특징 | 2.20 | 1.53 | 2.38 | 2.59 | 0.73 | 0.64 | 0.56 | 0.43 | 1.47 | 1.30 |
| Relief | 0.58 | 0.50 | 1.05 | 1.10 | 0.61 | 0.50 | 0.48 | 0.43 | 0.68 | 0.63 |
| CFS(BestFirst) | 0.51 | 0.43 | 0.88 | 0.82 | 0.63 | 0.58 | 0.43 | 0.38 | 0.61 | 0.55 |
| CFS(Genetic search) | 0.74 | 0.59 | 3.10 | 5.19 | 0.56 | 0.59 | 0.52 | 0.43 | 1.23 | 1.70 |
| CFS(Greedy stepwise) | 0.44 | 0.42 | 0.79 | 0.93 | 0.58 | 0.59 | 0.45 | 0.40 | 0.57 | 0.59 |
| Total(Avg) | 0.57 | 0.49 | 1.46 | 2.01 | 0.60 | 0.57 | 0.47 | 0.41 | - | - |

표 5. 이상치 제거 방법을 적용한 SVR 모델의 성능 비교
Table 5. Performance comparison of SVR models applying an outlier removal method

| SVR & Outlier Detection / 특징 선택 | SVR | | SVR & Leave One Out[16] | | SVR & SVDD[17] | | Bounded Influence SVR[18] | |
|---|---|---|---|---|---|---|---|---|
| | mean | stdev | mean | stdev | mean | stdev | mean | stdev |
| 모든 특징 | 0.55 | 0.43 | 0.60 | 0.51 | 0.59 | 0.46 | 0.56 | 0.46 |
| Relief | 0.48 | 0.43 | 0.49 | 0.45 | 0.50 | 0.47 | 0.48 | 0.44 |
| CFS(BestFirst) | 0.43 | 0.38 | 0.47 | 0.39 | 0.46 | 0.41 | 0.49 | 0.42 |
| CFS(Genetic search) | 0.52 | 0.43 | 0.57 | 0.49 | 0.51 | 0.45 | 0.55 | 0.47 |
| CFS(Greedy stepwise) | 0.45 | 0.40 | 0.44 | 0.46 | 0.48 | 0.44 | 0.49 | 0.42 |
| Total(Avg) | 0.49 | 0.41 | 0.51 | 0.46 | 0.51 | 0.45 | 0.52 | 0.44 |

비록 모델은 간단하지만, LR과 RT가 충분히 좋은 결과를 보였고, SVR 모델은 다른 회귀 모델에 비해 가장 낮은 예측 에러율을 보였다. 따라서 피어리뷰 학습자의 학습성과를 예측하는데 SVR이 가장 좋은 성능을 보였다.

## 4. 이상치 제거 방법을 적용한 실험 분석

피어리뷰 학습 환경에서 발생하는 데이터는 학습자 개인의 성향에 따라 다양한 변화량을 보이기 때문에 학습자 프로파일 정보에는 극단값이 존재한다[15]. 예를들어, 피어리뷰 환경의 writter 한 명에게 평가자 3명을 매치한 경우 각 평가자의 review 평가 결과는 개인의 성향에 따라 다르므로 작성자를 평가한 결과에는 이상치가 존재 할 수 있다. 이러한 학습자들의 모든 피어리뷰 정보를 프로파일 정보로 사용하게 되면 일부 학습자 학습성과에 편중된 영향을 미치거나 에러율을 높일 수 있기 때문에 학습자 프로파일 정보 중 극단값에 해당하는 이상치를 제거하는 방법을 적용하여 보았다.

본 연구에서는 SVR 모델에 사용한 이상치 제거 방법은 Leave One Out[16], SVDD[17], Bounded Influence SVR[18] 세 가지 이다. 교차 검증 방법 중 하나인 Leave One Out은 예측 모델이 현실에서 얼마나 정확하게 동작할지를 추정하는데 사용된다. 따라서 커널 함수를 통해 특징이 되는 데이터 세트를 포함하는 최적의 구를 찾은 후 범위 밖을 벗어난 예외적인 데이터를 이상 데이터로 판별한다.

SVDD(Support Vector Data Description)는 이상 데이터를 검출 할 수 있는 단일 클래스 분류기법 중의 하나로 데이터 세트에서 특징을 나타내지 않는 객체를 검출 한다. 즉, 특징 데이터에 비해 아주 작거나 큰 특징 값을 나타내는 경우 해당 데이터를 하나씩 제거 하면서 예측 모델을 생성한다. 이 경우 검증 데이터로 모든 데이터를 다 사용한 경우와 비교해 예측 성능이 저하되면 이상치로 분류 한다. Bounded Influence SVR의 경우 회귀 모델에 이상치 제거 방법을 적용한 예로 가중치를 이용해 이상치의 영향을 감소시켰다. 이 방법은 큰 회귀잔차를 위한 강인한 적응적 스케일 추정 법칙과 레버리지 포인트 제거를 위해 커널 함수를 이용하고 hat matrix의 통계에 기반한 적응적 가중치 적용 방법이다.

본 논문에서는 앞서 SVR 모델이 학습성과를 향상 시킬 수 있는 매치 페어 예측에 적합하다는 것을 확인했다. 따라서 SVR 모델에 이상치 제거방법을 적용하여 학습성과 예측 성능을 비교해 보았다.

표 5는 총 54명의 글쓰기 학습의 피어리뷰 학습자 데이터를 SVR 모델을 이용하여 학습성과를 예측한 결과와 SVR 모델에 관련연구[16][17][18]의 이상치 제거 방법을 적용하여 실험한 결과이다. Leave One Out과 SVDD를 적용한 SVR의 경우 학습 데이터에서 이분법적으로 이상치 데이터를 제거하기 때문에 각 특징 선택 방법 중 CFS(Greedy stepwise), CFS(Genetic search) 데이터에서는 에러율이 감소되었지만 평균 에러율은 0.51%로 SVR에 비해 높은 결과를 나타내었다. Bounded Influence SVR을 실험결과 relief 방법으로 추출된 속성을 이용한 경우, SVR과 같은 에러율을 보였고 전체 평균 에러율은 0.52%로 가장 높게 나타났다. 따라서 피어리뷰 학습자 프로파일 정보를 이용한 예측모델은 이상치를 포함한 SVR을 적용했을때 가장 낮은

0.47%의 에러율을 보였으므로 학습자들의 매치 페어를 예측력이 가장 높음을 의미한다. 이러한 결과는 일부 사람들의 다양한 특징에 따라 만들어낸 교육 데이터가 예외적인 극단 값의 성격을 가지고 있더라도, 교육학적 측면에서는 긍정적인 영향을 보일 수 있기 때문에 이상치를 제거하지 않는 것이 바람직하다고 볼 수 있다.

## 5. 결론 및 향후 연구

본 연구에서는 피어리뷰 학습자들의 학습성과 예측에 가장 적합한 예측 모델을 찾기 위해 다양한 회귀 모델을 통해 실험 및 분석을 수행하고 이상치 제거 방법을 적용하여 보았다. 피어리뷰 학습자 정보에 특징 선택을 수행하여 대표 속성을 추출하고 다양한 회귀 모델로 학습성과 예측 성능을 평가, 분석 하였다. 실험 결과, CFS(Greedy stepwise) 방법으로 추출된 속성들은 선형 회귀 모델에서 가장 좋은 예측 결과를 보였다. 또한 가장 낮은 예측 에러율을 보이는 SVR에 이상치 제거방법을 적용하여 실험한 결과 이상치를 포함한 SVR이 가장 좋은 성능을 보였다.

향후 연구에서는 SVR을 이용한 학습성과 예측 모델의 성능을 더욱 최적화 하여 학습자의 학습성과를 향상 시킬 수 있는 평가자를 찾는데 적용 할 계획이다.

## References


[1] Cho, K., Chung, T. R., King, W. R., and Schunn, C. “Peer-based computer-supported knowledge refinement: an empirical investigation,” *Commun. ACM*, vol 51, no. 3, pp. 83 - 88. 2008.

[2] Eric Zhi-Feng Liu, Sunny S. J. Lin, Chi-Huang Chiu, and Shyan-Ming Yuan, “Web-Based Peer Review: The Learner as both Adapter and Reviewer,” *IEEE Transactions on education*, vol. 44, no. 3, pp. 246-251, 2001.

[3] R.M. Crespo, A. Pardo, J.P. Somolinos and C. Delgado-Kloos, "An algorithm for peer review matching using students profiles based on fuzzy classification and genetic algorithms," *Lecture Notes in Computer Science*, vol. 3533, pp. 685-69, 2005.

[4] Gehringer, E. F, “Assignment and quality control of peer reviewers,” *Proceedings, ASEE Annual Conference*, Session 3230, 2001.

[5] Hye-Wuk Jung, Kwangsu Cho and Jee-Hyong Lee, "The Analysis of student pattern using peer review information of writing instruction" *Proceedings of the Korean Institute of Intelligent Systems*, vol. 21, no. 1, pp. 75-77, 2011.

[6] Machine Learning Group at University of Waikato, "Weka 3: Data Mining Software in Java," Available: http://www.cs.waikato.ac.nz/ml/weka/, 2008. [Accessed: May 1, 2012]

[7] Kenji Kira, Larry A. Rendell, "A Practical Approach to Feature Selection," *Proceedings of the ninth international workshop on Machine learning*, pp. 249-256, 1992.

[8] Hall, Mark A., Smith, Lloyd A., "Feature Subset Selection: A Correlation Based Filter Approach," *International Conference on Neural Information Processing and Intelligent Information Systems*, pp. 855 - 858, 1997.

[9] Michael H. Kutner, John Neter, Chris J. Nachtsheim: Applied Linear Statistical Models. © Richard D. Irwin, Inc. 1990.

[10] Elias Masry, "Multivariate regression estimation: Local polynomial fitting for time series", *Stochastic Processes and Their Applics*, vol. 65, pp. 81-101, Dec. 1996.

[11] Breiman, L., Friendman, J. F., Olshen, R. A., Stone C. J.: Classifcation and Regression Trees, © Wadsworth International Group, 1984.

[12] Vladimir Vapnik , Steven E. Golowich , Alex Smola, "Support vector method for function approximation, regression estimation, and signal processing," *Advances in Neural Information Processing Systems*, vol.9, pp. 281 - 287, 1996.

[13] SPSS Inc: SPSS for Windows: http://www.spss.co.kr/trial/trial_main.asp#

[14] R.-E. Fan, P.-H. Chen, and C.-J. Lin. "Working set selection using second order information for training SVM," *Journal of Machine Learning Research 6*, pp. 1889-1918, 2005.

[15] Hyojoung Shin, Hye-Wuk Jung, Kwangsu Cho and Jee-Hyong Lee, "Inference Model for Learning outcomes based on Support Vector Regression with Outlier Detection" *Proceedings of the Korean Institute of Intelligent Systems*, vol. 21, no. 2, pp. 224-225, 2011.

[16] Stone M., "An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion" *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 44-47, 1977.

[17] David M. J. Tax, Robert P. W. Duin, "Support Vector Data Description," *Machine Learning Journal*, vol. 54 Issue 1, pp. 45-66, January 2004.

[18] Franck Dufrenois, Johan Colliez, and Denis Hamad. "Bounded Influence Support Vector Regression for Robust Single-Model Estimation," *IEEE Transactions On Neural Network*, vol. 20, no. 11, pp. 1689-1705, 2009.

## 저 자 소 개

**신효정(Hyojoung Shin)**
2010년 : 충북대학교 정보통신공학과 학사
2012년 : 성균관대 임베디드 소프트웨어 학과 석사
2012년～현재 : 삼성전자

관심분야 : 임베디드소프트웨어, 지능 시스템, Flash translation layer
E-mail : shinhj0728@nate.com

**정혜욱(Hye-Wuk Jung)**
1999년 : 한성대학교 정보공학 공학사
2005년 : 성균관대학교 정보보호 석사
2005년～현재 : 성균관대학교 대학원 컴퓨터공학과 박사과정

관심분야 : 패턴인식, 생체인식, 지능시스템, 정보보호, 사용자 모델링.
E-mail : wukj@skku.edu

**조광수(Kwangsu Cho)**
2004년 : University. of Pittsburgh 인지과학 박사
2010년～현재 : 성균관대 인터랙션사이언스학과 부교수

관심분야 : Collaborative Learning Systems, Human-Computer Interactiong
E-mail : kwangsu.cho@gmail.com

**이지형(Jee-Hyong Lee)**
1993년 : 한국과학기술원 전산학과 학사
1995년 : 한국과학기술원 전산학과 석사
1999년 : 한국과학기술원 전산학과 박사
2002년～현재 : 성균관대 정보통신공학부 부교수

관심분야 : 지능시스템, 기계학습, 사용자 모델링
E-mail : jhlee@ece.skku.ac.kr