논문 2013-2-2

데이터마이닝과 소프트웨어공학

권기태*

Data Mining and Software Engineering

Ki-Tae Kwon*

요 약

데이터마이닝은 기존의 데이터 분석 기법과 대규모 데이터 처리를 위한 기법으로 데이터 분석을 통한 유용한 정보를 제공한다. 본 논문은 소프트웨어공학 분야에서 수집되는 다양한 소프트웨어 프로젝트 데이터 분석을 위해 데이터마이닝 기법을 적용하기 위한 시도로 수행되었다. 데이터마이닝 기법을 소프트웨어공학 분야 중 소프트웨어 프로젝트의 성공과 실패를 좌우하는 결정적 요소 중의 하나인 소프트웨어 비트웨어 비용산정 분야에 적용하는 연구가 진행되었다. SVR을 이용하여 소프트웨어 비용을 산정하였고, 이 때 SVR 모수들의 최적 조합을 면역계의 동작 원리를 응용한 면역 알고리즘을 이용하여 발견하였다. 소프트웨어의 정확한 비용산정을 위해 다양한 세대수, 메모리셀수, 대립유전자수를 대상으로 IA-SVR 알고리즘을 적용하였다. 실험결과는 기존의 연구들보다 우수함이 입증되었다.

Abstract

The accurate estimation of software development cost is important to a successful project in software engineering. Until recent days, the models using regression analysis based on statistical algorithm and data mining method have been proposed. However, this paper estimates the software development cost using support vector regression, a sort of new data mining technique. Also we presents the best set of optimized parameters applying immune algorithm, changing the number of generations, memory cells, and allele. The proposed IA-SVR algorithm outperforms some existed researches published in the literatures.

한글키워드 : 데이터마이닝, 소프트웨어공학, 소프트웨어 비용산정, 면역 알고리즘, SVR

Ⅰ. 서 론

정보화 사회가 컴퓨터와 네트워크 기술의 발전과 더불어 개인과 조직 들이 매일 대량의 데이터를 만들고 이를 축적하고 있다. 대량의 데이터는 데이터의 관측수와 더불어 속성변수도 많다는 의미를 내포하고 있다. 과거에는 자료로부터 유용한 정보를 얻기 위해 수작업으로 분석하였지만 근래에는 자료의 양이 커지고 복잡해짐으로써 자동화된 분석이 필요하게 되었다. 최근 데이터베

* 강릉원주대학교 컴퓨터공학과 교수
(email: ktkwon@gwnu.ac.kr)
접수일자: 2013.12.5 수정완료: 2013.12.17

이스, 압축, 통신 등 정보산업의 발전으로 이러한 대용량 자료에 대한 수집, 저장, 분석이 가능하게 되었다[1]. 또한 1950년대에 신경망, 군집분석, 유전 알고리즘, 1960년대에 의사결정나무, 1990년대에는 서포트 벡터 기계 등의 방법론들이 개발됨으로써 대용량 자료를 분석하는 것이 어느 정도 가능하게 된 것이다. 데이터마이닝은 이러한 기법들을 적용하여 자료에 숨겨진 패턴을 찾는 과정으로 볼 수 있으며 학계, 산업, 정부 등 여러 부문에서 널리 활용되고 있다[2].

데이터마이닝은 현재로 활발하게 연구되고 있는 분야이므로 이를 명확히 정의하는 것은 쉽지 않은 일이나, 소프트웨어공학 분야에는 거의 적용되지 않고 있다. 대용량의 소프트웨어공학 자료로부터 소프트웨어 프로젝트 정보의 요약과 미래에 대한 예측을 그 목표로 하며 소프트웨어공학 자료에 존재하는 관계, 패턴, 규칙 등을 탐색하고 이를 통계적으로 모형화함으로써 이전에는 알려지지 않은 소프트웨어 프로젝트의 유용한 지식을 추출하는 일련의 과정이 필요하다[3].

본 논문은 데이터마이닝 기법을 소프트웨어공학 분야에 응용하는 연구의 일환으로 특히 기존에 거의 연구되지 않았던 소프트웨어 비용산정 분야에 적용하는 연구를 진행한다.

본 논문은 1장 서론에 이어서, 2장에는 소프트웨어 비용산정의 문제점을 분석하고, 3장에는 IA-SVR 알고리즘을 소프트웨어 비용산정에 적용하는 방안을 제안한다. 4장에서는 제안한 방법을 이용한 시험 결과를 기술하며, 마지막 5장에서는 연구의 결론과 향후 발전방향을 기술하는 방식으로 구성되었다.

2. 소프트웨어 비용산정과 데이터마이닝

소프트웨어 프로젝트 초기에서 개발 비용을 정확하게 추정하는 것은 소프트웨어 개발의 성공과 실패를 좌우하는 중대한 요인이다, 예측한 비용이 예산을 초과하여 프로젝트를 취소할 수도 있고, 개발 업체는 실제 소요될 비용보다 비용을 과소 추정하여 이윤을 남기지 못함으로써 맏개한 손실을 초래할 수도 있다. 프로젝트 초기에 정확한 개발 비용 예측을 통해 관리자들은 필요한 자원을 적절하게 배정할 수 있다.

소프트웨어 비용산정에 관한 연구는 1965년 104가지 속성을 중심으로 하는 169개 프로젝트를 대상으로 하는 SDC 연구로부터 시작되었다. 이 연구를 계기로 1960년대 말과 1970년대 초에 부분적으로는 활용성이 있었던 일부 모델들이 개발되었다. 1970년대 말에는 SLIM, Checkpoint, PRICE-S, SEER, COCOMO 등의 알고리즘 모델이 개발되었다. 이들 대부분 모델의 개발자들은 동일 시기에 비용산정 모델을 개발했지만, 모두 유사한 난관에 봉착하였다. 즉, 소프트웨어 규모가 커지고 복잡성이 증가함에 따라 개발비용을 정확하게 추정하기가 매우 어렵다는 점이다. 알고리즘 모델 자체가 가지는 문제점과 더불어 소프트웨어 개발 환경이 급속하게 변함에 따라 이를 반영하는 정확한 비용산정 모델을 개발하기가 매우 난해하다는 점이다.

1980년대에는 알고리즘 모델들이 폭넓게 이용되었으며, 이 시기의 모델들은 다양한 크기와 다양한 환경을 반영하는 데이터 집합을 이용하여 비교되었다. 이러한 연구에서 얻은 주요 결론은 비용산정 모델들은 환경이 다른 경우에는 고라지 못한 요인들이 적용된다면 비영산정의 정확도가 하락한다는 점이다. 1990년에는 Abdel-Hamid,와 Madnick 같은 연구들을 통해 소프트웨어 개발이란 복잡한 동적 프로세스로 복잡성과 다양성을 기술할 수 있는 변경과 관련한 요인들을 충분히 알아낼 수 없음을 이해하게 되었다. 그 결과 1990년대에는 기계학습 알고리즘에 기반한 비모수 모델링의 소개와 산정 기술들이 등장하였다[4].

기계학습은 환경과의 상호작용에 가반한 경험적인 데이터로부터 스스로 성능을 개선하는 시스템으로 데이터 축적을 기초로 실행모델을 자동적으로 생성하는 기술이다[5, 6].

기계학습에 의한 소프트웨어 비용산정 기법으로 가장 먼저 사용된 기법은 신경망에 의한 비용추정 기법이다. 그 다음에 사례기반추론 기법을 도입하여 소프트웨어 개발비용을 예측하였고, 또한 트리 기반 기법으로 회귀트리, 의사결정트리 등을 활용하여 소프트웨어 개발비를 예측하기 위한 연구들이 줄을 이었다. 신경망, 사례기반추론 또는 회귀모형을 이용한 소프트웨어 비용산정의 정확도를 비교하면 기존 방법들과 비교하여 볼 때 연구자와 모델에 따라 다소 정확도의 차이가 있는 것으로 밝혀지고 있다[4].

[표 1] SW 추정과 데이터마이닝

SW 추정과 예측	데이터마이닝 기법
품질	GP, NN, CBR, DT, CL, ILP
규모	NN, GP
개발 비용	DT, CBR, BL
공수	CBR, DT, NN, GA, GP,
유지보수 공수	NN, DT
자원 분석	DT
수정 비용	GP, DT, ILP
신뢰도	NN
결함	BL
재사용성	DT
릴리스 시간	NN
생산성	BL
실행 시간	GA
모듈 시험용이성	NN

[표 1]은 소프트웨어 비용산정과 관련된 분야에 적용되는 데이터마이닝 기법을 정리한 것이다[7]. 본 연구에서는 기존에 거의 연구가 진행되지 않은 면역 알고리즘과 SVR의 결합, 유전 알고리즘과 SVR의 결합 기법을 소프트웨어 비용산정에 적용한다.

3. IA-SVR의 소프트웨어공학 적용

3.1 면역 알고리즘

생물학적 측면에서 면역 시스템은 외부 병원체에 대응하여 자율 분산 시스템이 생체 방어와 유지를 수행하는 것으로 뇌의 명령을 시스템 요소들이 따르지 않고 자율적으로 각 요소에 대응하는 것이다. 면역 시스템은 일반적으로 외부 항원에 반응하는 항체를 구성하고 이 항체들이 기억세포를 형성하고 분화한다. 이러한 반응체계를 공학적으로 적용하는 시스템이 면역 알고리즘이다[8, 9, 10].

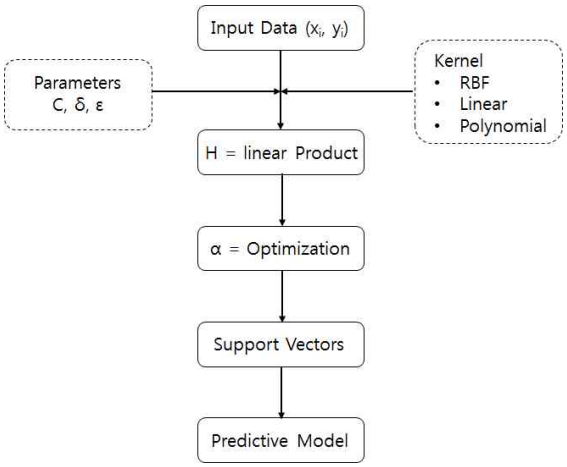
다른 비결정론적인 알고리즘과 동일하게 면역 알고리즘은 여라 개의 가능해를 동시에 최적화 진행을 수행하며, 해의 값 자체를 그대로 사용하는 것이 아니라, 코드화된 수의 배열을 사용한다. 또한 최적화 목적함수를 미분값과 그 외 다른 정보를 요청하지 않고 있는 그대로 사용한다는 장점이 있다. 그리고 이러한 비결정론적인 알고리즘의 특징 밖에도 면역 알고리즘의 가장 큰 특성은 최적해를 향한 수렴을 보장하는 메모리 셀을 가지고 최적화 과정을 진행한다[10, 11, 12].

최적화 문제에 면역 알고리즘을 대응시키면 면약 알고리즘의 항원은 최적화 문제의 제약 조건과 목적 함수로 대응되고, 항체는 최적화 문제의 해집단 후보가 된다. 또한 면역 알고리즘의 메모리 셀은 최적화 문제의 해집단이 되고, 면역 시스템에서의 항원과 항체 사이의 친화도는 최적화 문제에서의 적합도가 된다[10, 13].

3.2 서포트 벡터 회귀

서포트 벡터 머신 분류를 회귀 문제에 적용하여 훈련 데이터에 의존한 SVR 예측 모델을 만들 수 있다. SVR은 분류 최적화에 대한 일반화 능

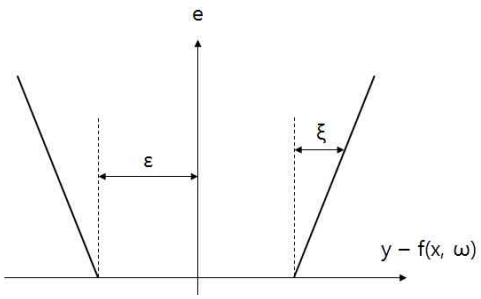
력이 뛰어나다. SVR의 절차는 [그림 1]과 같다 [14, 15].



[그림 1] SVR 처리 순서도

SVR은 학습 데이터 $D = \{(x_i, y_i) \in R^n \times R_i, i = 1,2,\cdots,l\}$ $x \in R^n, y \in R$가 있을 때 선형회귀 초평면은 식 (1)과 같다.

$$f(x, w) = W^T X + b \qquad (1)$$



[그림 2] ε-intensitivity 함수

SVR에서는 분류의 마진 대신 근사값 오류를 측정한다. 이 ε-intensitivity zone을 갖는 오류함수는 [그림 2]와 같다.

3.3 IA-SVR 알고리즘

본 논문에서는 소프트웨어 비용산정 분야에 적용하기 위한 데이터마이닝 기법으로 IA-SVR 알고리즘을 제안한다. SVR은 손실함수인 ε-intensitivity와 페널티 함수를 사용하여 이상값에 민감하지 않고 그 영향을 최소화시키는 장점을 가지고 있다. 그러나 SVR의 성능에 중요한 영향을 미치는 모수의 값은 훈련 데이터에 따라 사용자가 적절한 값을 설정해야 한다. 입력된 모수값에 따라 결과값이 크게 달라지기 때문에 훈련 데이터에 적합한 모수값을 찾기 위해 반복적인 작업을 수행해야 한다[16].

본 논문에서는 모수의 최적값을 찾고 비용산정의 성능을 높이기 위해 면역 알고리즘을 적용한다. 면역 알고리즘을 통해 훈련 데이터에 적합한 SVR 모수값을 최적화하고, 최적화된 모수값을 SVR에 설정하여 비용산정에 이용한다. 본 논문에서 제안하는 IA-SVR 알고리즘 기반의 소프트웨어 비용산정은 다음 6단계의 순서로 진행된다.

[1단계] 항원 인식

소프트웨어 비용산정을 위한 훈련 데이터, 시험 데이터, 초기값을 설정한다.

[2단계] 초기 항체군 생성

최초 과정은 유효 항체를 랜덤하게 생성한다. 각 항체는 SVR의 모수 조합이 되며, 이 항체를 이용하여 비용산정을 실시한다.

[3단계] 친화도 산출

친화도는 항체 간 또는 항원과 항체 간의 결합력으로 도출할 수 있다. 본 논문에서는 정보 엔트로피를 이용하여 생산 항체의 다양성을 측정하여 친화도를 산출한다,

[4단계] 메모리 셀 분화

이전 단계에서 산출된 친화도를 이용해서 항체 간의 친화도가 가장 큰 항체들을 삭제시켜 다양성을 향상하고, 항원과의 친화도가 큰 항체들을 메모리 셀에 저장한다.

[5단계] 항체 생성 촉진과 억제

항체 생성과 억제를 기댓값에 의존하지 않고 앞에서 도출된 친화도에 의하여 친화도가 가장 큰 1개를 메모리 셀에 저장한다. 또한 항체 간의 친화도를 도출하여 친화도가 큰 항체들을 상위부터 제거한다. 결론적으로 항원과의 친화도가 큰 항체 생산을 장려하고, 면역 시스템 전체에 항체 간 적합도가 큰 항체 생성을 억제하여 면역 시스템의 다양성 조절기구로 작용하게 한다.

[6단계] 항체 생성

친화도가 가장 큰 항체와 메모리 셀의 분화로 저장된 차세대 메모리 셀을 구성하기 위하여 현재 메모리 셀에 존재하는 항체들을 뽑아 돌연변이를 유도하여 새로운 항체를 최초 메모리 셀의 개수와 동일할 때까지 반복한다.

4. 실험 및 분석

IA-SVR 알고리즘을 이용한 소프트웨어의 비용산정은 LIBSVM Version 2.86과 Python 2.5를 사용하여 구현하였으며, 실험 데이터는 비용산정 연구분야에서 가장 널리 이용되고 있는 데이터 중의 하나인 NASA 소프트웨어 프로젝트 데이터 [17] 집합을 사용한다. 이 데이터 집합은 총 18개의 프로젝트로 구성되어 있으며 2개의 특성을 이용하여 공수(Y)를 추정하게 된다.

본 논문에서 제안한 IA-SVR 비용산정 모델을 평가하기 위한 척도로는 MMRE와 PRED(25)를 사용하였다. MMRE는 실제값과 추정값 사이의 상대적 오차의 평균 크기를 나타내며, 값이 작은 모델이 일반적으로 좋은 모델이다. PRED(25)는 추정값이 실제값의 25% 범위 내에 있는 비율이다. 좋은 모델일수록 PRED(25) 수치는 높게 나타나게 된다.

SVR의 평가방법은 교차 검증 방법인 LOOCV를 사용한다. N개의 데이터 집합을 이용할 경우, 이를 N개의 부분집합으로 나눈 후 (N-1)개의 부분집합을 훈련 데이터로 사용하고, 나머지 한 개를 테스트 데이터로 사용하여 SVR 성능 평가를 수행한다. 이를 N번 반복하여 평균 MMRE와 PRED(25)를 구한다.

훈련 데이터로 SVR 모델을 생성 후 실험 데이터로 테스트한 결과인 세대수, 메모리셀수, 대립유전자수에 따른 실험결과를 보면 MMRE는 세대수보다는 기억세포수에 영향을 많이 받으며, PRED(25)는 세대수, 메모리셀수, 대립유전자수에 뜨라 근소하지만 각 증가함에 따라 양호한 결과를 보인다.

실험결과에 의하면 세대수 20, 메모리셀수 40, 대립유전자수 10일 때 MMRE = 0.1001, PRED(25) = 89.11%로 가장 좋은 결과를 보이고, 이 때의 최적 모수 조합은 C = 29482, σ = 0.07662, ε = 0.09281이다.

훈련 시 대립유전자수와 세대수 증가에 따른 MMRE의 변화를 그래프로 살펴본 결과, 면역 알고리즘의 성능 평가를 위하여 세대수, 메모리셀수, 대립유전자수를 변경해가면서 실험하면 메모리셀수와 대립유전자수가 많은 것이 양호한 결과를 보였고, 세대수에 대한 변화는 초반인 10세대 이내에서 큰 변화가 있었고, 이후로는 변화가 미역하여 후반부에서는 큰 영향을 주지 못했다.

[표 2]는 본 논문에서 제안한 IA-SVR 알고리즘의 성능평가를 위해 동일한 NASA 소프트웨어 프로젝트 집합을 이용한 기존 연구와의 비교 결

과이다. 본 논문에서 제안한 방법이 기존 연구보다 우수함을 확인할 수 있다.

[표 2] 기존 연구와의 비교

출처	사용 모델	MMRE	PRED(25)
[18]	RBF	0.187	72.2%
[19]	SVR	0.165	88.89%
[20]	GA-SVR	0.1120	88.89%
본 연구	IA-SVR	0.1001	89.11%

5. 결론

본 논문에서는 데이터마이닝 기법을 소프트웨어공학에 적용하는 시도를 하였다. 특히 소프트웨어공학 분야에서 데이터마이닝 기법의 적용이 거의 시도되지 않았던 소프트웨어 비용산정 분야를 대상으로 데이터마이닝 기법을 적용하였다. 데이터마이닝 기법은 면역 알고리즘과 SVR을 결합하여 사용하였다. SVR은 분류 문제에 있어서 뛰어난 일반화 능력을 보이지만 데이터 집합에 따라 적합한 커널 함수와 모수값을 매번 찾아야 하는 단점이 있다. 이러한 단점을 해결하기 위해 본 논문에서는 면역 알고리즘을 이용하여 SVR에서 사용되는 모수값을 최적화시키는 알고리즘을 제안하였다. 본 논문에서 적용한 면역 알고리즘은 항체 간의 친화도가 큰 것은 항체 생성 억제 자기조절기능, 항원 간의 친화도가 큰 것은 신규 항체 생성을 통해 우수 항체 보존과 다양성을 통해 모수들을 최적화했다.

면역 알고리즘을 적용하여 발견한 최적 모수 조합으로 수행한 소프트웨어 비용산정 결과는 기존 연구보다 우수한 결과를 보임으로써 IA-SVR 알고리즘 기반의 소프트웨어 비용산정 방법이 기존의 SVR뿐만 아니라 데이터마이닝 기법의 적용의 모범 사례가 됨을 알 수 있었다.

향후 연구 과제로는 적절한 대립유전자와 항체를 기호화하는 연구와 함께 본 논문에서 제안하는 IA-SVR 알고리즘을 다양한 소프트웨어 공학 데이터에 적용하는 연구 등이 남아 있다.

참 고 문 헌

[1] 신택수, 홍태호, “비지니스 인텔리전스를 위한 데이터마이닝”, 사이텍미디어, 2009.

[2] 이정진, “R, SAS, MS-SQL을 활용한 데이터마이닝”, 자유아카데미, 2011.

[3] 김영옥, “데이터마이닝 기법을 이용한 소프트웨어공학 데이터 분석”, 강릉원주대학교 박사학위논문, 2031.

[4] 권기태, 이준길, “소프트웨어 비용산정을 위한 SVM의 파라미터 선정과 응용에 관한 연구”, 한국컴퓨터정보학회 논문지, 제14권, 제3호, pp. 209-216, 2009.

[5] 박혜영, 이관용, “패턴인식과 기계학습”, 이한출판사, 2011.

[6] 한학용, “패턴인식 개론”, 한빛미디어, 2009.

[7] Du Zhang and Jeffrey Tsai, “Machine Learning Applications in Software Engineering”, World Scientific, 2005.

[8] Dipankar Dasgupta, “Artificial Immune Systems and Their Applications”, Springer, 1999.

[9] 심귀보 외, “컴퓨터 면역기스템 개발을 위한 인공면역계의 모델링과 자기인식 알고리즘”, 한국 퍼지 및 지능시스템학회 논문지, 제11권, 제10호, pp. 910-918, 2001. 10.

[10] 권기태, 이준길, “소프트웨어 비용산정을 위한 면역 알고리즘 기반의 서포트 벡터 회귀”, 한국컴퓨터정보학회 논문지, 제14권, 제7호, pp. 17-24, 2009.

[11] 정형환 외, “면역 알고리즘을 이용한 전력계통 안정화 장치의 최적 파라미터 선정”, 전기학회 논문지, 제49A권, 제9호, pp. 433-445, 2000. 9.

[12] 박진현 외, “DC 모터 파라미터 변동에 대한 면역 알고리즘 제어기 설계”, 한국 퍼지 및 지능시스템학회 논문지, 제12권, 제4호, pp. 353-360, 2002. 4.

[13] Pang-Ning Tan et al., “Introduction to Data Mining”, Addison Wesley, 2006.
[14] Changha Hwang, “Support Vector Median Regression”, Data and Information Science, Vol. 14, No. 1, pp. 67-74, 2003.
[15] Toby Segaran, “Programming Collective Intelligence”, O’relly, 2007.
[16] 권기태, 이준길, “면역 알고리즘 기반의 서포트 벡터 회귀를 이용한 소프트웨어 신뢰도 추정”, 한국IT서비스학회지, 제8권 제4호, pp. 129-140, 2009. 12.
[17] Alaa F. Sheta, "Estimation of the COCOMO Model Parameters Using Genetic Algorithms for NASA Software Projects“, Journal of Computer Science 2, pp. 118-123, 2006.
[18] Miyoung Shin and Amrit L. Goel, “Empirical Data Modeling in Software Engineering Using Radical Basis Functions”, IEEE TSE, Vol. 26, No. 6, pp. 567-576, 2000.
[19] Adrina L. I. Oliveria, “Estimation of Software Project Effort with Support Vector Regression”, Neurocomputing, Vol. 69, pp. 1749-1753, 2006.
[20] 권기태, 박수권, “유전 알고리즘 기반의 서포트 벡터 회귀를 이용한 소프트웨어 비용 산정”, 정보처리학회논문지D, 제16-D권, 제5호, pp. 729-736, 2009.

저 자 소 개

권기태

1986년 서울대학교 계산통계학과 졸업
1988년 서울대학교 계산통계학과 석사 졸업.
1993년 서울대학교 계산통계학과 박사 졸업
1996년 미국 Univ. of Southern California, 전산학과 Post-Doc.
현재 강릉원주대학교 컴퓨터공학과 교수

<주 관심분야 : 소프트웨어공학, 데이터마이닝, 지능시스템>

※ 이 논문은 2013년도 강릉원주대학교 교수연구년연구 지원에 의하여 수행되었음.