기계학습 알고리즘을 이용한 기능적 요구사항과 비기능적 요구사항으로 요구사항 분류에 관한 연구

조병선 [1] 이석원 [2]
아주대학교 컴퓨터공학과[1] 아주대학교 소프트웨어학과[2]
chobs1028@ajou.ac.kr[1], leesw@ajou.ac.kr[2]

A Comparative Study of Classification into Functional and Non-functional Requirements using Machine Learning Algorithms

Byung-Sun Cho[1] Seok-Won Lee[2]
Ajou University Dept. of Computer Engineering[1] Ajou University Dept. of Software[2]

요 약

본 논문은 소프트웨어 공학의 요구사항을 기능적 요구사항과 비기능적 요구사항으로 분류하는데 있어서 기계학습 알고리즘을 적용한 연구로, 자연어로 작성된 소프트웨어 개발 요구사항에 대하여 자연어 처리 기법을 이용한 전처리 과정, 대표적인 예측 분류기인 나이브 베이지안, 다층 퍼셉트론, 의사 결정 나무 그리고 서포트 벡터 머신을 적용하여 각각 정확도, 재현율, 카파통계치 결과를 비교 및 분석한다.

1. 서 론

1.1 연구 배경

소프트웨어 공학에서 요구분석은 소프트웨어가 만족시켜야 할 요구사항을 도출, 분석, 명세화, 그리고 검증하는 과정으로 소프트웨어 또는 시스템 개발에 초기에 요구되는 중요한 프로세스이다. 그러기에 이해관계자의 요구사항의 분석은 개발 및 프로젝트의 성공에 결정적이고 또한 소프트웨어 또는 시스템 개발주기에서 많은 비중을 차지하는 프로세스이다[1].

이러한 소프트웨어 시스템 요구사항은 기능적 요구사항(Functional Requirement)과 비기능적 요구사항(Non-Functional Requirement)으로 분류할 수 있다. 기능적 요구사항은 시스템이 제공하는 기능 또는 서비스에 대해 상세하게 기술된 요구사항으로 특정 상황에 대하여 시스템이 어떻게 반응 및 동작을 해야 하고 하지 말아야할 것에 대하여 기술되어 있다. 그에 반해 비기능적 요구사항은 시스템 속성이나 시스템에 의해 제공되는 서비스나 기능에 대한 제약사항에 대하여 기술되어 있다. 요구사항 분석 과정에서 요구사항을 기능적 요구사항 또는 비기능적 요구사항으로 정확하게 분류하는 일은 오늘날 소프트웨어 개발을 위한 요구분석에 있어서 많은 도움이 된다.

오늘날에는 소프트웨어 요구사항이 다양해지고 그 중요성이 커지면서 소프트웨어 요구 명세서(Software Requirements Specification)에서 잘 정의 되어 있는 반면에 인간이 아닌 시스템을 이용한 요구사항의 분류에는 아직 많은 어려움이 있다. 가장 큰 이유 중 하나는 요구사항이 자연어로 작성되어 있기 때문이다. 이는 이해관계자 간의 서로 다른 용어의 사용과 같은 요구사항의 다른 묘사 및 설명 방식을 초래할 수 있고 이에 따라서 요구사항의 분석은 정확히 되지 않는 경우가 많다. 이는 시스템을 이용한 요구사항의 분류에 큰 어려움이자 소프트웨어 개발에 큰 걸림돌이 될 수 있다.

1.2 연구 목표

본 연구의 목표는 요구사항을 기능적 요구사항과 비기능적 요구사항으로 정확하게 분석 및 분류하여 시스템 개발 시 요구사항 분석에 도움을 주는 것이다. 이는 기계학습을 이용하여 요구사항을 기능적 요구사항과 비기능적 요구사항으로 분류한 결과를 비교 및 분석하여 그 목표를 달성할 수 있다.

이에 따라 본 논문에서는 인공지능의 한 분야인 데이터 마이닝을 알고리즘과 함께 이용 및 적용하여 요구사항을 기능적 요구사항과 비기능적 요구사항으로 분류하는데 있어서 더 정확한 방법을 제시한다.

2. 본 론

2.1 데이터

본 논문에서는 사용되는 데이터는 The Quality Attributes (NFR) 데이터와 Global Personal Marketplace (GPM) 내의 요구사항 데이터를 사용한다.

The Quality Attributes (NFR)는 OpenScience tera-PROMISE repository에서 제공하는 공개 데이터세트로 255개의 기능적 요구사항과 370개의 비기능적 요구사항으로 총 625개의 요구사항으로 구성되어 있다.

Global Personal Marketplace (GPM)는 Global Personal Marketing에서 공개한 Global Personal Marketplace(GPM) 시스템 요구 명세서에서 요구사항만

추출한 데이터세트로 119개의 기능적 요구사항과 84개의 비기능적 요구사항 총 203개의 요구사항으로 구성되어 있다.

<표 1>은 본 연구에서 사용하는 데이터 샘플을 보여주고 있다. 데이터 샘플은 자연어로 작성되어 있고 클래스에서 F는 기능적 요구사항을 U는 가용성, SE는 보안성으로 비기능적 요구사항으로 분류된다.

<표 1> 데이터 샘플

데이터세트	요구사항	클래스
NFR	The system shall display Events or Activities.	F
GSM	The system shall allow modification of the display.	F
NFR	The product shall be easy to learn	U
GSM	The product shall free of computer viruses.	SE

<표 2>는 각각의 데이터의 기능적 요구사항과 비기능적 요구사항의 클래스의 분포를 보여주고 있다.

<표 2> 요구사항 클래스 분포

데이터 세트	클래스 분포	
	기능적 요구사항	비기능적 요구사항
NFR	255 (40.8%)	370 (59.2%)
GPM	119 (58.62%)	84 (41.38%)
Total	374 (45.17%)	454 (54.83%)

2.2 텍스트 전처리

본 연구에서 사용되는 데이터 세트 내 데이터 타입은 텍스트이기 때문에 정보검색(Information retrieval) 기법의 전처리(preprocessing) 과정을 통해 기계학습 알고리즘에 적용할 수 있게 데이터를 정제가 필요하다.

본 연구의 전처리과정으로는

1) 오타 및 단어의 문자 코드 오류 수정

이 과정에서 실제 NFR과 GPM의 데이터에 오타 및 단어의 문자 코드 오류가 존재 및 수정하였다.

2) 단어빈도-역문서빈도 변형(tf-idf Transform)

단어빈도-역분서빈도는 정보 검색과 텍스트 마이닝에서 주로 이용되며 핵심어를 추출하거나, 특정 문서 내에서 단어의 중요도를 통계적 수치로 표현한다[2].

3) 토큰화 (tokenization)

이 과정에서는 텍스트 분할(text segmentation)로 문자열의 토큰화를 처리 하였다[3]. 이를 통하여 문자열을 특수 문자를 제외하고 문자열을 문자로 변형하여 데이터 전처리 작업을 수행하였고 이를 통해 문자열 데이터 타입을 알고리즘에 맞는 데이터 형식으로 변환하였다.

<표 3>에서는 <표 2>에서 보여준 예시를 데이터를 전처리 과정을 수행한 결과를 보여준다.

2.3 분류 알고리즘

2.3.1 나이브 베이지안

나이브 베이지안 (naïve bayes) 알고리즘은 베이즈 정리 (bayesian theorem)에 기반한 알고리즘으로 각 예측에 대한 속성 값이 서로 조건부 독립이라는 것과 두번째로 드러나지 않거나 잠재적인 속성이 예측 과정에 영향을 미치지 않는다는 것이 특징이다.

2.3.2 다층 퍼셉트론

다층 퍼셉트론 (Multi-layer Perceptron)은 피드포워드(feed forward) 인공 신경망으로 최소한 세 개의 노드 계층(layer)으로 구성되어 선형과 비선형 분류에 모두 활용된다. 이 때, 세 개 이상의 노드 계층은 입력 계층 (input layer), 은닉 계층 (hidden layer), 출력 계층 (output layer)을 포함하고 있고 입력 계층을 제외한 각 노드는 비선형 활성화 함수를 사용한다.

2.3.3 의사 결정 트리

의사 결정 트리 (Decision Tree)는 데이터 마이닝, 기계 학습에서 사용하는 예측 모델링 방법 중 하나로 어떤 항목에 대한 관측값과 목표값을 연결시켜주는 예측 모델로 트리 구조를 이용하여 데이터 관계를 시각화하여 그 결과를 보여준다.

2.3.4 서포트 벡터 머신

서포트 벡터 머신은 기계 학습 분야 중 패턴 인식과 자료 분석을 위한 지도 학습 모델로 주로 분류와 회귀 분석을 위해 사용된다. 분류 결과인 클래스에 속한 데이터의 집합이 주어질 때, 서포트 벡터 머신 알고리즘은 주어진 데이터 집합을 바탕으로 새로운 데이터가 어느 집합에 속하는가를 판단하는 비확률적 이진 선형 분류 모델을 생성한다[4].

2.4 실험

2.4.1 실험 방법

기본적으로 NFR과 GPM 두 데이터세트는 각각의 요구사항과 그에 따른 클래스 레이블(class label)만 존재하기에 본 연구에서 각각의 데이터세트가 가지고 있는

<표 3> 예시 데이터 전처리 결과

Class	Attribute																	
	activities	allow	display	event	modification	of	or	shall	system	the	be	computer	easy	free	learn	product	to	viruses
F	0.96	0	0.48	0.96	0	0	0.96	0	0.48	0	0	0	0	0	0	0	0	0
F	0	0.96	0.48	0	0.96	0.48	0	0	0.48	0	0	0	0	0	0	0	0	0
NF	0	0	0	0	0	0	0	0	0	0	0.96	0	0.96	0	0.96	0.48	0.96	0
NF	0	0	0	0	0	0.48	0	0	0	0	0	0.96	0	0.96	0	0.48	0	0.96

데이터가 많지 않기에 두 데이터세트를 합쳐서 새로 만든 데이터세트도 기능적 요구사항과 비기능적 요구사항 분류하는데 포함하여 결과를 비교 분석하였다.

기계학습의 알고리즘은 자바 기반의 오픈소스 데이터 마이닝 툴인 WEKA를 사용하여 적용하였고 검증 방법으로는 10겹 교차 검증 방법(10 fold cross validation)을 이용하였다.

10겹 교차 검증 방법은 기계학습에서 데이터의 양이 충분치 않을 때 분류기 성능측정의 통계적 신뢰도를 높이기 위해서 쓰는 방법이다. 본 연구에서 기계학습 알고리즘을 적용했을 때 과최적합(overfitting) 현상을 피하고 데이터의 양이 충분하지 않는 이 연구에서 실험 결과의 신뢰도를 높이기 위해 사용하였다[5].

2.4.2 실험 결과 및 분석

정확도 (Precision), 재현율 (Recall) 그리고 카파 (kappa) 통계치를 통해 정확하고 신뢰도 높은 결과를 도출하기 위해 <표 4> 에서 데이터 세트에 대해 나이브 베이지안, 다층 퍼셉트론, 의사 결정 트리 알고리즘, 그리고 서포트 벡터 머신을 적용한 결과를 정확도, 재현율 그리고 카파 통계치를 보여주고 있다.

정확도, 재현율, 그리고 카파 통계치의 결과를 보면 서포트 벡터 머신이 가장 높은 수치를 나타내고 있고 그 다음으로 다층 퍼셉트론, 나이브 베이지안 마지막으로 의사 결정 트리순으로 높은 수치를 보여 주고 있다. <표 4> 를 보면 다층 퍼셉트론과 서포트 벡터 머신에서 비기능적 요구사항의 정확도와 기능적 요구사항의 재현율 부분에서 같은 값을 보여주면서 동시에 100%라는 결과값을 보여주고 있다. 하지만 다른 부분에서는 다층 퍼셉트론이 서포트 벡터 머신보다 낮은 결과를 보여 주고 있고 이는 카파 통계치를 보면 많은 차이를 확인할 수 있었다. 본 연구에서 서포트 벡터 머신이 좋은 결과를 보이는 가장 큰 이유는 서포트 벡터 머신이 이진 선형 분류 모델에 최적화된 모델이기에 기능적 요구사항과 비기능적 요구사항의 이진분류(Binary Classification)로 표현되는 실험 데이터에서 더 좋은 결과를 보여 주는 것으로 분석된다.

3. 결 론

본 연구에서는 요구 사항을 기능적 요구사항과 비기능적 요구사항으로 분류를 하는데 있어서 기계학습 알고리즘을 이용하였다. 여러 알고리즘을 적용한 후 결과의 정확도, 재현율, 그리고 카파 통계치를 비교 및 분석을 통해서 보다 더 정확하고 신뢰도 높은 결과를 도출할 수 있었다.

분류를 하는데 있어서 기능적 요구사항과 비기능적 요구사항 두 클래스만으로 분류하기 때문에 이진 분류라 할 수 있었고 본 연구에서 여러 알고리즘을 적용하였고 그 중 서포트 벡터 머신이 정확도, 재현율, 카파 통계치 부분에서 다른 알고리즘에 비해 가장 좋은 결과 값이 도출 되어서 종합적으로 요구사항을 기능적 요구사항과 비기능적 요구사항으로 분류하는 본 연구를 통해 이 예측 방법이 좋다고 결론을 지을 수 있다.

4. Acknowledgement

이 논문은 2017년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2017R1D1A1B03034279)

참 고 문 헌

[1] Ian Sommerville, Software Engineering, 10th Edition; Pearson, Harlow, England, 2015

[2] https://en.wikipedia.org/wiki/Tf%E2%80%93idf

[3] https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html

[4] Black Nathanael, T.; Wolfgang, E. Introduction to Artificial Intelligence; Springer: Berlin, Germany, 2011

[5] F. Provost and T. Fawcett, “Data Science for Business – What You Need to Know About Data Mining and Data-Analytic Thinking”, OREILLY, pp. 26 – 34 (2013)

<표 4> 분류 알고리즘 정확도, 재현율, 카파 통계치 결과

분류 알고리즘	데이터 세트	정확도(Precision)		재현율(Recall)		카파 통계치
		F	NF	F	NF	
나이브 베이즈	NFR	0.757	0.861	0.808	0.822	0.623
	GPM	0.948	0.897	0.924	0.929	0.8485
	Total	0.759	0.840	0.818	0.786	0.6004
다층 퍼셉트론	NFR	0.811	0.899	0.859	0.862	0.7145
	GPM	0.915	**1**	**1**	0.869	0.8861
	Total	0.844	0.924	0.914	0.861	0.77
의사 결정 트리	NFR	0.736	0.817	0.733	0.819	0.5526
	GPM	0.884	0.932	0.958	0.821	0.7933
	Total	0.749	0.803	0.765	0.789	0.5522
서포트 벡터 머신	NFR	**0.904**	**0.876**	**0.916**	**0.859**	**0.7774**
	GPM	**0.967**	**1**	**1**	**0.952**	**0.9591**
	Total	**0.886**	**0.928**	**0.914**	**0.903**	**0.8152**