Information Systems Review
Vol. 25, No. 1 February 2023 http://dx.doi.org/10.14329/isr.2023.25.1.047

효과적인 의사결정을 위한 다중레이블 기반 속성선택 방법에 관한 연구: 감성 분석을 중심으로

Exploring the Performance of Multi-Label Feature Selection for Effective Decision-Making: Focusing on Sentiment Analysis

원 종 윤 (Jong Yoon Won) 성균관대학교 경영대학 일반대학원 박사과정
이 건 창 (Kun Chang Lee) 성균관대학교 경영대학 교수, 교신저자

요 약

본 연구는 인공지능 기법 중 다중레이블 속성선택 방법을 적용하여 복잡한 경영환경에서 의사결정의 효과성을 증대시키는 방안을 설명한다. 인공지능 기반의 의사결정 시스템은 의사결정자의 선택과 판단을 돕거나, 대신하는 중요한 역할을 한다. 더욱이 최근 인공지능을 중심으로 한 비즈니스 의사결정은 기업의 성장 동력으로 평가받는데, 이를 위해서는 효과적인 의사결정 방법이 수반되어야 한다. 이에 본 연구는 의미 있는 속성값을 선별하는 CFS-BR(이진연관성 접근 기반의 상관관계 속성선택 모델)을 제안하여, 효과적인 의사결정을 지원하는 것을 돕는다. 예시데이터와 실증데이터의 분석 결과, CFS-BR은 유의미한 속성을 최상우선선별 알고리즘 기반으로 최상의 조합을 선별하므로 효율적 의사결정을 지원할 수 있고, 기존의 다중 레이블 속성선택 방법과 비교하였을 때 정확도가 높은 것으로 보아 효과적인 의사결정을 증대시키는 데 유용하다.

키워드 : 다중 레이블, 속성 선택, 감성 분석, 데이터 마이닝, 의사결정

Ⅰ. 서 론

경영활동의 주요 업무인 의사결정은 과거에는 의사결정자가 가진 권위와 경험에서 나오는 직관에 의존했다. 반면, 이제는 데이터를 기반으로 하여 객관적인 의사결정을 선호하는 것으로 패러다임이 변화하고 있다. 특히, 이를 지원하는 다양한 방법 중 머신러닝 기법이 각광받고 있다(Phillips-Wren *et al.*, 2021). 그중 기존의 머신러닝 기법은 단일 레이블 분류 방법으로 타깃변수가 하나이지만(Huang *et al.*, 2013), 다중 레이블 분류 방법은 두 개 이상의 타깃변수에 대해 분석하는 기법으로 더 효과적으로 데이터를 분석하는 방안이라고 평가받는다(Tsoumakas and Katakis, 2007).

예를 들어, 다중 레이블을 적용한 데이터 분석 기반의 비즈니스 의사결정을 위한 정보시스템의 사례는 다음과 같을 것으로 예상한다. 정보화 시스템을 구축하기 위해 정보시스템 전문가는 기업의 요구 사항을 수렴한다. 이를 사용자 요구 사항

분석이라 한다(Meridji *et al.*, 2019). 대체로 사용자는 정보시스템 개발에 대한 지식이 전혀 없는 경우가 많다. 그래서 정보시스템 전문가는 사용자와 기업의 프로젝트 범위, 기준 등을 고려한 다양한 요구 사항에 관하여 사전에 논의한다. 사용자 요구 사항은 기능적 요구 사항, 비기능적 요구 사항으로 구분하는데, 기능적 요구 사항은 시스템이 제공해야 하는 기능을 뜻하고, 비기능적 요구 사항은 사용성, 신뢰성, 성능, 지원성, 제약사항 등을 말한다. 다중 레이블 분류 기법을 사용자 요구 사항 분석에 적용하면, 해당 기업의 다양한 요구 사항이 타깃변수로 치환된다. 그리고 기업이 보유하고 있는 역량이나 프로젝트의 범위 또는 기준이 인스턴스에 할당한다. 그 결과 기업의 여건이나 환경에 기반을 두어 시스템 구축이 가능한지 예측 및 평가할 수 있으며, 기업고객을 이해시키고, 만족하게 할 객관성을 확보할 수 있다. 또한, 온라인 리뷰에 대한 감성 분석을 예로 들면, 기존의 단일 레이블 분류 방법은 전체적인 리뷰글의 긍정 혹은 부정 등의 극성값만을 도출하지만, 다중 레이블 분류 방법을 적용한 스마트폰 온라인 리뷰를 예시로 디자인, 화질, 카메라, 속도, 무게 등 다양한 항목에 대한 각각의 감정을 추정하고 분석할 수 있다. 그뿐만 아니라 다중 레이블 분류는 여러 타깃변수를 한 번에 분석하고 타깃변수 간의 상관성을 고려할 수 있으므로 복잡한 온라인 리뷰를 분류하는데 더 효과적이라고 평가받는다(Khan *et al.*, 2016; 정풀잎 외, 2019). 이는 효과적인 의사결정을 내리는데 기존의 방법보다 더 도움이 될 것으로 여겨진다. 이렇듯 현실 세계는 동시에 다양한 의미를 내포하므로 단일 레이블 분류 방법으로 해결하지 못하는 문제가 있지만(Stamatescu *et al.*, 2019; Sun *et al.*, 2016), 다중 레이블 분류 방법은 현실 세계의 복잡성과 다양성을 동시에 반영하는 복잡한 의사결정을 지원할 수 있을 것이다.

빅데이터 시대는 더 많은 데이터를 어떻게 분석하고 처리하느냐가 기업의 경쟁력으로 평가받을 것이다. 실제로 많은 기업이 머신러닝 알고리즘을 자사의 핵심 자산으로 선언하고, 정교한 머신러닝 알고리즘을 전사 전략, 제품 개발, 위기 대응 등 다방면에 활용하고 있다(George *et al.*, 2014; Henrique *et al.*, 2019). 즉, 경쟁기업보다 더 정교한 머신러닝 모델을 만들고 이를 통해 비즈니스 의사결정에 도입하는 것이 중요한데, 이를 위해서는 더 많은 학습 데이터가 필요하다. 그러나 많은 양의 데이터를 처리할 때 몇 가지 문제점이 발생할 수 있다.

그중 수많은 데이터 가운데 불필요한 데이터로 인해 잘못된 결과가 도출하거나, 너무 많은 양의 데이터를 분석하는 데 많은 시간을 낭비하게 된다. 이러한 문제를 해결하기 위해 데이터에서 유용하고, 의미 있는 정보만을 추출하는 차원 축소 기술이 주목받고 있다(Erevelles *et al.*, 2016; Yassine *et al.*, 2019). 특히, 차원 축소 기술 중에 속성선택은 문제를 해결하기 위한 적절한 속성값을 찾아 데이터를 재구성한다(Cai *et al.*, 2018; Xu *et al.*, 2016). 이러한 속성선택 방법은 세 가지 장점이 있는데, 첫째, 데이터의 원형을 변형하지 않기 때문에 유용한 정보를 버릴 위험이 적다. 둘째, 타깃변수와 속성 간의 상관관계를 이해하는 데 도움이 된다. 마지막으로, 데이터의 양을 줄일 수 있으므로 데이터를 저장하고 관리하는 비용을 줄일 수 있으며, 분석하는 시간을 줄여주므로 효율적인 대안이다(Miao and Niu, 2016). 이러한 속성선택의 장점은 유의미한 속성만을 선택적으로 분류하고 분석할 수 있기에 빅데이터 시대에 효과적인 의사결정을 위한 핵심요소라 평가할 수 있다.

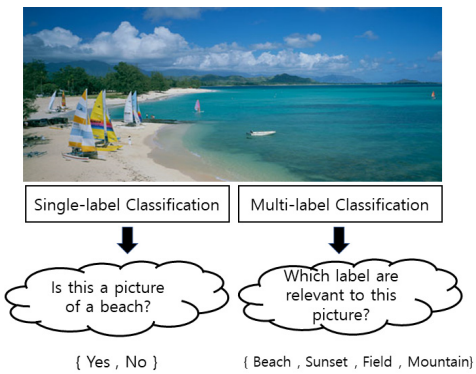
본 연구는 앞서 언급한 내용을 바탕으로 다중 레이블 기법을 적용한 효과적인 의사결정을 위해 다중 레이블 기반 속성선택 방법인 CFS-BR(Correlation-based Feature Selection based on Binary Relevance approach: 이진 연관성 접근 기반의 상관관계 속성선택 모델)을 제안하고, CFS-BR이 의사결정의 효과를 증대시키는지 감성 분석으로 실증한다.

Ⅱ. 관련연구

2.1 다중 레이블 분류

데이터를 분석할 때 목표 대상의 종류에 따라 데이터마이닝 방법이 다르게 적용된다. 데이터마이닝 방법에는 분류, 회귀, 군집, 패턴 마이닝, 시계열 예측 등이 있다(Gera and Goel, 2015). 그중 분류는 가장 대중적이고 일반적인 방법이다. 기존의 데이터마이닝에서 가장 많이 사용된 것은 단일 레이블 분류 방법이다. 이는 현실 세계의 객체가 하나의 레이블(타깃변수)로 연결된 것이라 말할 수 있다. 기계학습 패러다임에서 단일 레이블 분류 방법은 현실 세계의 객체를 하나의 인스턴스로 표현하고, 하나의 레이블과 연결된다. 이에 X는 인스턴스를 의미하고 Y는 레이블을 의미한다. 학습 데이터 세트 : $\{(x_i, y_i) \mid 1 \leq i \leq m\}$를 통해 학습 기능 f: X → Y를 수행한다(Zhang *et al.*, 2014). 이때 $x_i \in X$는 객체의 속성을 나타내는 인스턴스이고, $y_i \in Y$는 그에 대응하는 레이블이다. 그 결과 기본적인 가정으로 각 사례가 하나의 개념에만 속하며, 하나의 고유한 의미로만 해석된다.

이와 달리 다중 레이블은 하나의 사례가 여러 개념에 속할 수 있다. 다중 레이블 데이터는 인스턴스 공간인 X에 레이블 집합을 갖게 되는데, 부분집합은 $Yi \subset L$, $L=\{y_1, y_2, ..., y_q\}$로 q의 레이블 세트를 가진다(Zhang and Zhou, 2013). <그림 1>은 단일 레이블 분류와 다중 레이블 분류가 어떻게 다른지 설명하는 그림이다. '그림이 해변인가?'라는 물음은 단일 레이블 문제이며, '사진에 어떤 것들이 있는가?'라는 질문은 다중 레이블 문제라 할 수 있다. 이에 대한 답으로 단일 레이블 분류는 '예' 또는 '아니요'라고 할 수 있으며, 다중 레이블 분류는 해변, 해넘이, 산 등으로 다양한 답이 있다.



〈그림 1〉 단일레이블 분류와 다중레이블 분류의 예시

다중 레이블 분류를 단일 레이블 분류의 멀티 클래스 문제와 비교하면 다중 레이블 분류의 효과성과 효율성을 이해하기 쉽다. 데이터 구조를 기반으로 레이블 세트의 크기가 작은 멀티 클래스 분류 문제는 다중 레이블 분류 문제를 일부 해결할 수 있는 것처럼 보인다(Agrawal *et al.*, 2013). 그러나 다중 레이블 분류의 경우 레이블의 수가 증가할 때마다 변수의 개수가 크게 많아지므로 근본적인 해결책이 되지 않는다. 예를 들어, 다중 레이블의 레이블 세트가 n개일 때, 다중 레이블을 멀티 클래스로 표현하기 위해서는 타깃변수가 2n 또는 3n개 등으로 늘어나기 때문이다. 이는 높은 차원의 다중 레이블이 가지고 있는 타깃변수를 멀티 클래스가 전부 반영하는 것은 어렵다. 멀티 클래스 학습은 타깃변수 간의 상관성을 고려하지 않는다는 점에서 다중 레이블 분류 방법은 기존의 단일 레이블 분류 방법보다 개선된 분류 방법이다(Tsoumakas and Katakis, 2007). 다중 레이블 분류 연구는 크게 두 가지로 구분할 수 있다. 다중 레이블 분류를 위한 새로운 알고리즘 개발과 알고리즘을 활용한 새로운 분석모델을 제시하는 것이다. 그중 분석모델을 제시하는 초기 연구는 텍스트 분류에 관한 연구가 많다. 최근에는 텍스트뿐만 아니라 음성, 이미지와 같은 다양한 유형의 온라인 정보를 분류하는 연구가 이뤄지고 있으며, 게놈 및 생물학과 같은 분야 등 여러 응용 프로그램을 개발하는 데 다양하게 적용되고 있다(Zhang and Zhou, 2006). 다중 레이블 분류 방법에 대한 수요는 지속해서 증가하고 있으며(Zhang *et al.*, 2014),

다중 레이블 분류 방법의 최근의 연구현황은 <표 1>과 같다.

Nam *et al.*(2014)은 기존의 역전파 다중 레이블 학습 알고리즘(Back-Propagation Multi-Label Learning, BP-MLL)보다 분류성능이 우수한 신경망 학습 모델을 제시하였다. 해당 연구에서 제시하는 신경망 학습 알고리즘은 기존의 신경망 알고리즘보다 대용량의 텍스트 세트를 처리하는 데보다 빠르고 정확한 분류성능을 보인다고 보고하였다. 이는 새로운 알고리즘을 개발하는 연구로 전산학에서 주를

〈표 1〉 최근 다중 레이블 분류연구 현황

	내용	분야
Nam *et al.*(2014)	BP-MLL 알고리즘의 한계성에 대한 인지하고 오류를 최소화할 수 있는 인공신경망 학습 구조를 제시한다. 새로운 신경망 모델은 기존의 모델에 비해 대규모의 텍스트 데이터 세트를 처리하는데 우수한 결과를 보임.	컴퓨터정보
Bromuri *et al.* (2014)	만성질환 환자의 의료기록 자료를 활용하여, 다중 레이블로 만성질환 환자에 대한 분류가 보다 용이하도록 모델을 제시.	컴퓨터정보-의학
Liu and Chen (2015)	다중 레이블 분류기법을 적용하여 감성 분석모델을 제시하였다. 중국의 재해와 관련하여 사람들의 감정을 각 5가지, 10가지로 분류하고, 다양한 다중 레이블 분류기에 대해 평가, 비교함.	행정
Khan *et al.*(2016)	유튜브의 스마트폰 리뷰를 분석하는 데 단일 레이블 분류방법과 다중 레이블 분류방법의 비교를 통해 효과성을 입증함.	컴퓨터정보
Corani and Scanagatta (2016)	대기오염을 예측하는 모델을 통해 다중 레이블 분류방법이 기존의 단일 레이블 분류방법보다 더 효과적인 것을 입증함.	환경
Wehrmann and Barros(2017)	인공신경망 학습 기반의 다중 레이블 분류방법으로 영화 예고편에 대한 장르 분류 시스템을 제안함.	컴퓨터 정보
Jabreel and Moreno(2019)	이전의 감성 분석이 단일 레이블 분류방법에 초점이 맞춰져 있는 사실에 착안하여, 소셜 미디어 사용자의 감정을 딥러닝 기반의 다중 레이블 분류방법 모델을 제시함.	컴퓨터 공학
de Morais *et al.* (2019)	뉴스의 객관성과 합법성 간의 차이에 대해 고찰하고, 가짜 뉴스를 찾기 위해 다중 레이블 분류방법을 적용하였다. 그리고 다중 레이블과 멀티 클래스 문제 간의 비교를 통해 다중 레이블의 우수성을 입증하였다.	정보시스템
Liu *et al.*(2020)	다중레이블 기반의 딥러닝 알고리즘인 BrandImageNet 모델을 개발하여 소비자의 지각적 브랜드 속성을 분석하였다. 모델의 우수성을 검증하기 위해 전통적인 자기보고 방식의 결과와 일치하는 것을 확인하였으며, 해당 연구는 소비자의 이미지를 활용하고 이해하는데 기여한다.	마케팅
Le *et al.*(2021)	식사경험에 따른 온라인 리뷰에 대해 타자의 진성성, 생산자의 진정성, 자아의 진성성 측면으로 구분하여 다중레이블과 단일레이블 결과를 비교하였다. 그 결과 다중레이블 분류가 데이터를 분석하는데 정확하고 효과적이라 보고하였다.	관광
Schlegelmilch *et al.*(2022)	다중레이블 알고리즘을 적용하여, 6개국 2백만 개 이상의 코로나19 관련 트위터를 정치, 경제, 사회, 기술, 환경 및 법률로 분류하였다. 그리고 이를 기반으로 소비자의 감정을 재분류하였다. 해당 연구는 국제 마케팅에서 다국어 온라인 감성분석을 하는데 기여한다.	마케팅
Fujii *et al.*(2022)	증권 보고서에서 리스크를 추출하고 분류하였다. 다양한 알고리즘을 사용하였는데, 그중 BERT 기반의 다중레이블 분류기법으로 기업별, 산업별에 대한 위험을 분류하는데 기여하였다.	재무

이루는 연구 분야이다.

이러한 새로운 알고리즘을 개발하는 연구와는 달리 기존의 다양한 알고리즘을 적용하여 새로운 분석모델을 제시하는 논문이 있다. 기존의 감성 분석 연구는 머신러닝 기법을 많이 적용하였는데, 대체로 단일 레이블 분류 방법에 초점이 맞춰져 있다. 이에 다중 레이블 분류 방법을 적용한 감성 분석 연구는 다음과 같다. Liu and Chen(2015)은 소셜 미디어 사용자들이 올리는 짧은 글에도 하나의 감정이 아닌 두 개 이상의 감정이 내포되어 있다는 점에 착안하여, 걱정, 화남, 두려움, 놀람 등의 감정을 분류하였다. 데이터에 따라 각각 5가지, 10가지의 감정을 구분하여 분석하였다. 그 결과 중국 자연재해 관련 온라인 리뷰의 감정을 다중 레이블 분류를 적용하여 분석할 때, 가장 적합한 감정 사전과 다중 레이블 분류 알고리즘을 선정하여 제시하였다. Khan *et al.*(2016)은 소비자 온라인 리뷰를 다중 레이블 방법으로 분류하였다. 온라인 소셜 미디어의 발달로 텍스트뿐만 아니라 이미지, 영상 등의 다양한 콘텐츠가 소비자들로부터 생성되고 있다. 많은 소비자는 기존의 구매자가 남긴 온라인 리뷰에 의존하고, 실수를 줄이려는 행동을 취한다. 이에 온라인 공간에서 주고받는 개개인의 의견은 기업의 소비 매출에 크게 영향을 주었다. 이러한 관점에서 해당 연구는 유튜브의 스마트폰 리뷰를 다중 레이블 방법으로 분류하였다. 이를 위해 텍스트 데이터의 전처리를 품사 단위로 구분하고 연구자가 부여한 특정 키워드로 분류하였다. 그 결과 단일 레이블 분류 방법과 다중 레이블 분류 방법이 스마트폰의 온라인 리뷰 분석에 효과적이라는 것을 입증하였다. Jabreel and Moreno(2019)는 감성 분석을 위한 딥러닝 기법을 적용한 다중 레이블 분류 알고리즘을 제시하였다. 딥러닝 기법의 우수성을 검증하기 위해 트위터와 같은 소셜 미디어의 데이터를 활용하여 여러 감정을 분류하였다. 그 결과 딥러닝 기반의 다중 레이블 분류 알고리즘은 감성을 분석하는 데 효과적임을 입증하였다. 일반적인 감성 분석뿐만 아니라 가짜 뉴스를 탐지하는 것에도 다중 레이블 분류 방법의 효과성이 입증되었다. de Morais *et al.*(2019)은 뉴스의 합법성과 객관성에 대한 관계를 기반으로 가짜 뉴스를 탐지하기 위해 다중 레이블 학습 방법을 적용하였다. 연구 결과 단일 분류기인 멀티 클래스 학습 방법보다 다중 레이블 분류 모델이 가짜 뉴스를 탐지하는 것에 효과적임을 실증하였다. 최근 텍스트뿐만 아니라 영상과 이미지를 분류할 때 다중 레이블 방법은 더욱 효과적이라는 연구 결과가 있다. Wehrmann and Barros(2017)는 영화 예고편을 통해 영화의 장르를 분류할 수 있는 자동화 시스템 모델을 제시하였다. 온라인의 많은 콘텐츠는 사진이나 글보다 영상의 수가 급격히 증가하였다. 이러한 영상 콘텐츠는 어느 한 장르에만 국한되지 않고, 두 개 이상의 장르로 분류할 수 있다. 지금까지는 이러한 분류작업이 시간과 비용이 많이 드는 사람의 수작업으로 이뤄졌다. 이러한 현실에 착안하여 해당 연구는 영상물에 대한 자동 분류 시스템으로, 다중 레이블 영화 장르 분류를 위한 합성곱 신경망 알고리즘(Convolution-Through-Time for Multi-label Movie genre Classification, CTT-MMC)을 제안하였다. 연구결과 학습된 데이터에 따라 영상의 장르를 분류하는 데 효과적임을 입증하였다.

또한 다중 레이블 분류 방법은 여러 분야에서 실용적으로 사용되고, 다양하게 응용되고 있다. Cerri *et al.*(2016)은 단백질의 하위 기능을 분류하는 데 있어, 다중 레이블 분류방법을 적용하였다. 신체의 기능이나 단백질의 종류에 따라 다양한 상호작용이 일어나는데, 기존의 단일 레이블로 분류는 단백질의 복잡한 기능을 설명하는 데 한계가 있었다. 이에 해당 연구는 다중 레이블 분류 방법을 적용하여 기존의 다른 선행 연구보다 단백질의 하위 기능을 더욱 명확히 밝혀내었다고 보고하였다. Corani and Scanagatta(2016)은 다중 레이블 알고리즘을 적용한 대기오염을 예측하는 모델을 제시하였다. 연구모델의 유효성을 입증하기 위해 상하이, 베를린, 부르가스의 대기오염 예측 결과를 다중 레이블과 단일 레이블 분류방

법을 따로 도출하여 비교하였다. 초미세 먼지(Particulate Matter Less than 2.5μm, PM2.5)의 대기질 지수를 예측하기 위해 이산화질소 대기 질 지수, 온도, 압력, 습도, 바람, 날씨 변수 등을 고려하여 모델을 도출하였다. 연구결과 다중 레이블 분류방법을 적용한 대기오염을 예측하는 모델이 단일 레이블 분류방법보다 정확도가 더 높은 것으로 나타났다. Bromuri *et al.*(2014)은 병원 환자의 의료기록을 활용하여 환자들의 질병을 예측하였다. 병원 환자들 중 만성질환 환자들의 의료기록 텍스트를 인스턴스 데이터로 활용하였다. 그리고 다중 레이블 분류방법으로 만성질환 환자의 질병들을 분류하였다. 그 결과 환자들이 만성질환 중에 어떠한 질병에 걸렸는지, 해당 질병과 다른 질병 간의 상관성에 대해 기존 단일 레이블 분류방법보다 명확하게 예측함에 따라 환자의 질병 예측 모델을 제시하였다. Bogaert *et al.*(2019)은 금융 서비스 부문에서 다중 레이블 알고리즘을 적용한 추천 서비스 모델을 제안하였다. 금융업의 경우 금융회사들이 자체 개발한 상품뿐만 아니라 다른 금융회사의 상품까지 판매하는 방식으로 교차판매가 존재한다. 교차판매는 신규 고객을 확보하는 것보다 매출을 늘리기 쉬우므로 금융 산업에서 유용하다. 해당 연구는 소비자들의 교차 상품 판매에 대한 소비 정보를 기반으로 특정 고객이 다음에는 어떤 상품에 관심이 있으며, 구매 여부에 대해 다중 레이블 분류 방법을 적용한 상품 추천 시스템을 제안하였다.

기존의 다중 레이블 분류 관련 연구 중 다수의 연구가 새로운 알고리즘을 개발한 컴퓨터공학 계열의 연구이다. 또는 의학 계열에서 다중 레이블 분류 방법을 적용한 연구가 많이 이뤄졌다. 예를 들면 불면증, 정신 분열증, 치매 등 질병을 예측하기 위해 다중 레이블 분류 방법을 사용하였거나(Folorunso *et al.*, 2020), 단백질과 같은 세포를 분류하는데 많이 적용되었다(Cerri *et al.*, 2016).

경영학 관련 연구의 경우, Liu *et al.*(2020)은 다중레이블 기반의 딥러닝 알고리즘을 개발하여 소비자가 지각하는 브랜드 속성을 분석하였다. 그리고 해당 알고리즘의 결과를 증명하기 위해 전통적 자기보고 방식의 결과와 비교하였는데, 알고리즘과 실제 사람이 인식하는 수준은 차이가 없는 것으로 나타났다. 이러한 연구는 소비자가 업로드하는 이미지를 이해하고 기업이 활용하는 데 도와줄 것으로 기대한다. 또한 Schlegelmilch *et al.*(2022)은 트위터의 글들을 기반으로 정치, 경제, 사회, 기술, 환경 및 법률에 대해 분류하고, 이에 대한 감정을 분석하여 국제 마케팅 감성 분석에 대해 검증하였다. 이외에도 경영학적 관점에서 다양하게 연구되고 있다.

최근 다중 레이블 분류를 적용한 국내 연구는 <표 2>와 같다. 민동영, 조성준(2017)은 다중 레이블 분류방법을 적용하여 우리나라 유가증권 시장의 기업들의 산업 분류방법을 제안하였다. 기존의 산업분류 방식이 주관적이고 한 종목에만 속하지 않기 때문에 기존의 방식을 탈피해야 한다고 주장하였다. 이에 유가증권 시장에 존재하는 기업들의 사업보고서와 신문기사를 기반으로 등장하는 단어를 정리하여 산업군 사전을 구축하였고, 이를 기반으로 다중 레이블 분류방법을 적용하여 기존의 방식과는 다르게 주관적이지 않고 객관적인 산업군 분류 모델을 제시하였다.

임소라, 권용진(2017)은 기존의 특허문서를 분류하는 작업은 사람에 의해 수작업으로 이뤄지는 점에서 착안하여, 자동화 분류 모델을 제시하였다. 국제 특허 분류 기준을 기반으로 기술 분야 및 배경기술 필드 속성을 활용하여 630개의 범주를 구성하고 분류하였다. 해당 연구는 특허문서의 경우 동시에 여러 개의 코드를 갖는다는 국제 특허 기준에 따라 단일 레이블 분류방법과 달리 문제를 한 번에 처리할 수 있는 모델을 제시하였다.

이민성(2017)은 사용자가 미래에 방문하고자 하는 장소에 관한 장소 예측연구를 하였다. 이를 위해 다중 레이블 분류방법을 적용하였다. 장소예측은 마케팅, 도시계획, 교통예측 등 광범위하게 연구되는 주제이다. 다중 레이블을 적용한 장소예

〈표 2〉 국내 다중 레이블 분류 연구현황

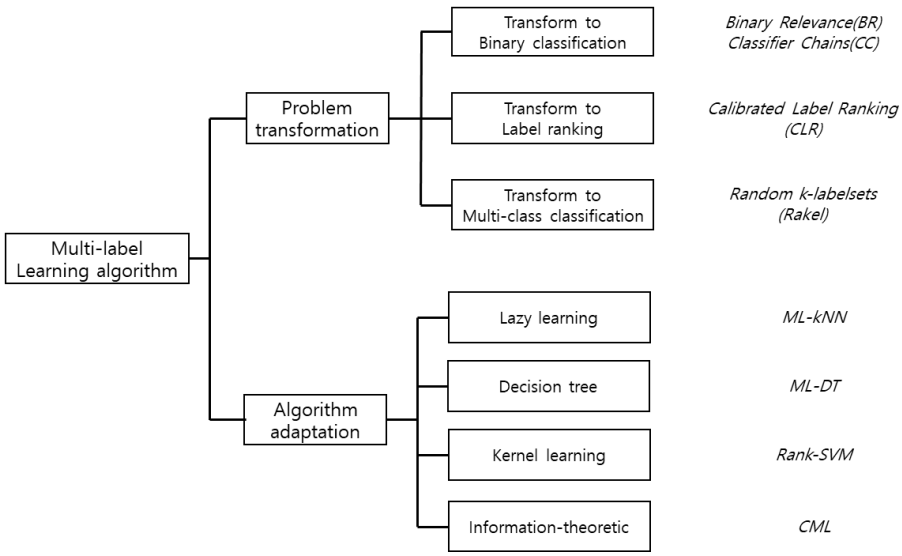
	내용	분야
임소라, 권용진 (2017)	특허문서의 경우 사람의 손으로 분류되는 배경을 바탕으로 머신러닝 기법을 적용한 자동분류 모델을 제시함.	컴퓨터정보
이민성(2017)	다중 레이블 분류기법을 적용하여 사용자의 장소예측(Location prediction)이 가능한 모델을 제시하고 분류기 간의 성능을 비교함.	컴퓨터정보
미아오쉬, 이재성 (2018)	음악에 내재 되어있는 감성을 추출하고 분류하는데 다중 레이블 분류 알고리즘을 적용하였음. 특히, 속성선택 방법을 적용하여 음악 감성 인식 모델을 제시하였음.	컴퓨터공학-인문
이상현 외(2019)	정형외과 의사는 골절이 발생하면 컴퓨터 단층촬영(CT)을 활용해 골절 범위를 식별하고 치료방법을 결정하는데, 골절이 발생하면 다발성 골절인 경우가 많음. 해당 연구는 다발성 골절 부위를 판단하기 위해 다중 레이블 분류방법을 적용함.	컴퓨터공학-의학
장수진 외(2019)	영화의 장르는 하나의 장르에 국한되지 않기 때문에 다중 레이블 연구에 적합한 데이터라 할 수 있음. 해당 연구는 영화 포스터의 이미지를 근거로 영화의 장르를 다중 레이블 분류 알고리즘을 사용하여 예측하는 모델을 제시함.	컴퓨터공학
임채현 외(2021)	다중레이블 분류 방법을 적용하여 여드름, 블랙헤드, 주근깨 등 복합적인 피부질환일 인식하는 모델을 제시함.	컴퓨터공학

측 모델을 제시하기 위해 다양한 다중 레이블 분류기를 벤치마킹하여 장소예측에 가장 적합한 분류기를 선별하였다. 이렇듯 다중 레이블 분류방법을 적용한 국내 연구의 경우 유가증권 시장의 산업군 분류, 특허문서 분류, 장소예측 등 실생활의 데이터를 활용한 사례가 증가하고 있다. 특히 앞서 설명한 바와 같이 현실 세계의 데이터는 다중 레이블 데이터 세트로 자연스럽게 분석할 수 있다는 장점이 있다.

2.2 다중 레이블 분류 알고리즘

앞서 논의한 바와 같이 다중 레이블 분류학습을 위한 분류 알고리즘은 단일 레이블 분류 알고리즘보다 어렵고 복잡하다. 따라서 학습에 따른 주요 접근법을 두 가지로 구분할 수 있는데, 이는 문제변환 방법과 알고리즘 적응 방법이다(Tsoumakas and Katakis, 2007). 전자는 다중 레이블 학습을 단일 레이블 학습 시나리오로 변환하는 것이 요점으로 여러 개의 다중 레이블 타깃변수를 하나의 레이블로 인식 또는 변환하여 분류한다. 문제변환방식의 장점은 단순하며, 단일 레이블 분류방법의 모든 알고리즘을 사용할 수 있어 많은 학습 알고리즘을 사용할 수 있다(Read *et al.*, 2011). 후자의 경우 잘 알려져 있고, 많이 사용되는 학습 알고리즘 기술을 다중 레이블 분류방법에 직접 적용하여 분석한다. 알고리즘 적응방식의 장점은 목표변수인 레이블 세트를 변형하지 않고 인스턴스와 레이블을 직접 학습하여 더 자연스럽게 해석할 수 있다. 다만, 문제변환 방식보다 복잡하고 사용할 수 있는 알고리즘이 제한된다는 단점이 있다. 요약하면, 문제변환 방법은 알고리즘에 데이터를 맞추는 것이라 할 수 있고, 알고리즘 적응 방법은 알고리즘을 데이터에 맞추는 것이라 할 수 있다(Wu *et al.*, 2020).

본 논문은 Zhang *et al.*(2014) 연구의 기준에 따라 대표성을 갖는 알고리즘을 소개하고자 한다. 대표성의 기준은 첫째, 알고리즘이 얼마나 많은 연구에서 사용되는지, 둘째, 후속연구가 많이 진행되는지, 마지막으로 알고리즘에서 고유한 특징을 갖는지 이다. 그에 따른 결과는 <그림 2>와 같다.



〈그림 2〉 다중 레이블 분류 알고리즘 구분

2.2.1 문제변환 방법

이진 연관성은 기존의 데이터 세트에 존재하는 각각의 레이블들을 하나의 기준으로 두 그룹으로 분류하는 변환 기법으로, 설정한 분류 규칙에 따라 목표와의 연관성을 0과 1의 결과로 나타난다(Montañes *et al.*, 2014). 즉, 인스턴스에 따라 각각의 레이블을 독립적으로 평가하는 것이다. 그 결과는 각각의 레이블에 대해 예측값이 계산된다. 이진 연관성은 많은 알고리즘과 접목 가능하며, 다른 모델과의 결합이 간단하다. 그렇지만 레이블 간의 상관성을 고려하지 않는다는 단점이 있다. 이로 인해 실제로 존재하지 않는 레이블 집합을 추측하는 상황이 발생하기도 한다(Zhang *et al.*, 2018).

이를 해결하기 위해 추측 정확도를 높일 수 있는 분류기 체인 등의 기법과 함께 사용한다. 분류기 체인은 레이블들의 측정 결과를 입력 데이터로 추가하여 레이블 예측값을 추론하여 각각의 레이블이 결과에 영향을 줄 수 있는 방법이다(Liu and Tsang, 2015). 첫 번째 분류기는 원래 입력 특성만 사용하여 학습하고, 출력된 레이블이 새로운 입력 속성으로 추가된 뒤 새로운 입력 공간에서 두 번째 분류기를 학습하는 것에 사용한다. 반복되는 과정의 효율성을 높이기 위해 베이지안 네트워크를 활용하기도 한다. 레이블 간의 의존성을 고려하여 분류기를 연결하고 이진 연관성이 레이블 간의 상관관계를 고려하지 못하는 것을 개선하는 방법이다(Wang *et al.*, 2019). 이진 연관성과 분류기 체인은 <그림 3>과 같은 데이터 세트 예시를 통해 도식화할 수 있다.

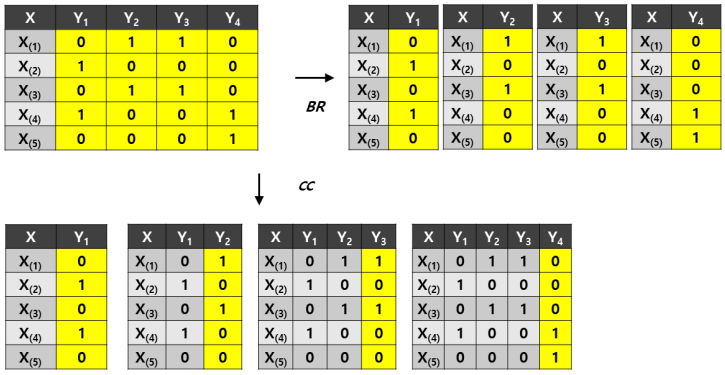
레이블 멱 집합(Label Power-set, LP)은 다중 레이블 학습 데이터 세트에 존재하는 레이블의 하위 집합들을 만들고 이를 처리할 수 있는 새로운 데이터를 만든다. 생성된 데이터들은 원래 데이터 세트의 멱 집합들과 일치하기 때문에 변환된 데이터 세트를 분류할 수 있으며, 새로 생성된 레이블 집합 중 하나를 새로운 인스턴스에 할당할 수 있다. 다중 레이블 분류학습에서 레이블 멱 집합은 레이블 수가 적은 데이터 세트에 주로 사용되며, 레이블 수가 많은 데이터 세트에 사용되면 멱 집합을 만들기가 어렵다. 레이블 멱 집합은 레이블 간의 상관성을 고려할 수 있지만, 각 레이블당 학습 데이터의 수가 감소해 예측 정확도가 감소하기도 한다. 레이블 멱 집합의 추측 정확도를 높이기

위해 라켈(Random k-label-sets, RAkeL)와 함께 사용한다(Gupta *et al.*, 2019).

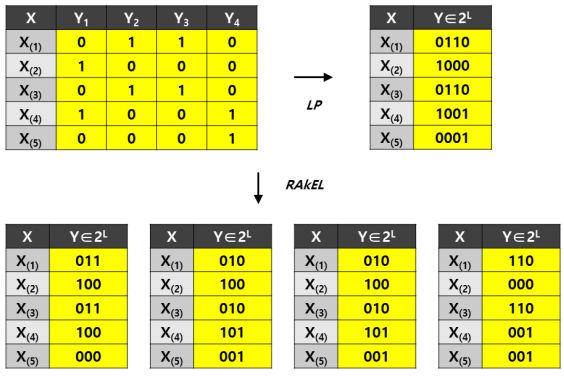
라켈은 분류기 체인과 비슷하게 무작위로 k 크기만큼의 레이블 세트를 k개 선택한 뒤 k번 레이블 멱 집합을 실행한다. 각각의 레이블 세트에 맞게 모델들이 이진 그룹으로 나누어진 예측치를 결과로 나타내고, 결과들은 다중레이블 예측 모델과 결합하여 예측 결과값이 가장 높은 정확도를 가졌는지를 판단하게 된다. 라켈은 레이블 멱 집합의 작업을 단순하게 만들어 분석의 효율성을 높이고, 하위 집합 전체를 대상으로 하지 않고 무작위로 k개를 추출하여 분석하기에 시간 절약에 좋다. 또한, k개를 추출해서 진행하지만, 라켈은 더 균형있는 하위 집합들의 분포를 포함한다. 그리고 라켈은 원래 훈련 데이터 세트에 나타나지 않던 레이블 세트를 예측할 수 있다는 장점이 있다(Rokach *et al.*, 2014). 레이블 멱 집합과 라켈은 <그림 4>와 같은 데이터 세트 예시를 통해 비교할 수 있다.

2.2.2 알고리즘 적응 방법

다중 레이블 k-최근접 이웃 알고리즘(ML-kNN)은 새로운 테스트 데이터의 특징과 모든 테스트 데이터의 특징 사이의 거리를 계산하여 가장 인접한 k개의 이웃을 찾는다. 새로운 샘플에 속해있는 인스턴스 중 가까운 것을 찾으면 해당 인스턴스와 근접한 레이블을 사용하여 새로운 샘플의 레이블을 예측한다. k-최근접 이웃 알고리즘은 모델을



〈그림 3〉 문제변환방식 예시(이진 연관성, 분류기 체인)



〈그림 4〉 문제변환방식 예시(레이블 멱 집합, 라켈)

만들지 않으며, 새 샘플이 반환된 경우에만 분류 작업을 수행하기 때문에 다중 레이블 k-최근접 이웃알고리즘으로 불린다(Wang and Zucker, 2000). 다중 레이블 k-최근접 이웃 알고리즘은 기계학습 알고리즘인 k-최근접 이웃 알고리즘을 다중레이블 시나리오에 맞게 적용한 것이다. 작동 방식은 k-최근접 이웃 알고리즘과 비슷하나 각 레이블이 가지고 있는 선험적 확률에 의해 조건부 확률을 계산하고, 이를 근거로 근접한 이웃의 최대 확률을 계산한다(Zhang and Zhou, 2005). 최대 확률의 계산이 끝나면 새로운 샘플의 레이블을 예측하는 것이 완료된다. 이때 등장하는 모든 확률은 각 레이블에 대한 신뢰수준으로 레이블의 순위를 생성하여 나열할 수 있다(Zhang and Zhou, 2007).

다중 레이블 의사결정 나무(ML-DT)는 새로운 샘플이 어떤 클래스에 속하는지 분류하는 과정을 나무 형태로 표현한다. 훈련 데이터를 주면 트리를 자동으로 생성하고, 정보의 습득이 최대가 되도록 자식 노드를 생성한다. 알고리즘 중 유일하게 학습 과정의 지식을 도출할 수 있기도 하다. 다중 레이블 의사결정 나무는 의사결정 나무의 최종 결과로 나오는 이파리 노드에 레이블 셋이 나타난다. 각 노드에 특정 속성의 값이 아닌 레이블의 값들을 기준으로 각 레이블 셋의 하위 레이블 집합들을 분류한다(Vens *et al.*, 2008).

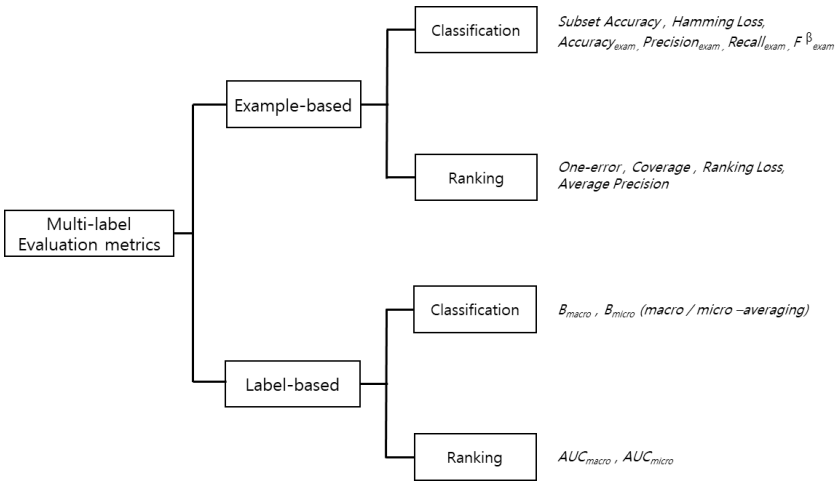
랭크 서포트 벡터 머신(Rank-SVM)은 대체로 이진 분류 문제를 해결하기 위해 많이 사용되었다. 따라서 레이블 간의 상관관계가 존재할 때 사용하는 것은 적합하지 않다. 이러한 문제를 해결하기 위해 서포트 벡터 머신 원칙을 기반으로 한 직접 접근 방식인 랭크 서포트 벡터 머신은 다중 레이블 분류의 특성에 맞게 적용된 알고리즘으로 효율성이 높고 우수하다(Elisseeff and Weston, 2002). 레이블 간의 상관관계를 고려하여 레이블의 순위 값을 산출하고, 특정 기능이 전체 순위의 하위 집합으로서 예측된 레이블 세트를 추출한 후에 이를 임계값으로 조정하여 사용된다. 다차원 공간에서 서로 다른 클래스를 분류하는 평면을 훈련 데이터로부터 결정하기 때문에, 고차원 데이터를 작업할 때에도 우수한 성능을 보여준다(Jiang *et al.*, 2008).

집단 멀티 라벨(Collective Multi Label, CML)은 고려할 수 있는 모든 상관관계를 고려하는 분류 알고리즘이다. 새로운 데이터 세트가 추가되면 관련된 속성들과 속성들의 근처에 있는 속성들 그리고 관찰되지 않은 레이블 간의 상관성까지 계산하여 결과값을 보여준다. 집단 멀티 라벨은 각각의 레이블에 독립적으로 시행되는 기존의 머신러닝의 분류기법들과는 달리 적합한 기법으로 고려되지 않는다. 다중 레이블과 현실 세계의 데이터와 같이 타깃변수가 두 개 이상일 경우에는 집단 멀티 라벨이 더욱 적합하다고 평가받는다(Ghamrawi and McCallum, 2005).

위와 같은 다중 레이블 분류 알고리즘은 비즈니스에 다양하게 활용될 수 있다. 예를 들면, 스마트 팩토리에서 수집되는 정보들을 저장하고 분석하는데 다중 레이블 분류 알고리즘은 생산 운영과정의 효율성을 가져올 것이다. 최근 의료경영이 주목받는데, 다중 레이블을 적용한 유전자 분석은 개인맞춤형 의료서비스를 현실화시킬 수 있어 의료소비자에게 큰 혜택을 제공하고, 새로운 비즈니스 모델을 형성해 부가가치를 창출할 수 있다.

2.3 다중 레이블 분류 평가지표

단일 레이블 분류는 데이터나 알고리즘을 평가하는 데 있어 '옳다', '그르다'와 같이 두 가지로 구분하기 때문에 다중 레이블에 비해 간단하지만 단편적이다. 단일 레이블에서 대표적으로 사용되는 평가지표는 정밀도, 재현율, 정확도, F-척도 등이 있다. 이와 달리 다중 레이블 분류는 단일 레이블 분류보다 레이블이 여러 개로 구성된 데이터 세트이므로 계산이 복잡하고 다양해 단일 레이블에서 대표적으로 참고하는 평가지표와는 다르다. 따라서 다중 레이블의 데이터 및 분류기의 성능을 평가하기 위해서는 단일 레이블보다 몇 가지의 지



〈그림 5〉 다중 레이블 분류 평가지표 구분

표가 더 존재한다. 이 중에서 해밍로스가(Hamming Loss) 다중 레이블을 평가할 때 가장 중요한 평가지표로 사용된다고 알려져 있다(Pereira *et al.*, 2018).

2.3.1 예제 기반 평가지표

<그림 5>는 다중 레이블 연구의 평가지표를 도식화하여 분류한 것이다. 해밍로스는 다중 레이블 평가지표 중 가장 많이 사용된다. 잘못 예측한 오류와 예측에 실패한 오류를 계산하기 위해 전체 인스턴스와 전체 레이블의 개수를 기반으로 정규화 과정을 거친다. 식 (1)과 같이 계산되며 오류들이 분자에 들어간다. 이로 인해 우수한 모델의 경우 오류가 적기 때문에 해밍로스의 값은 작게 나타난다.

$$\text{해밍로스} = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \triangle Z_i|}{|L|} \tag{1}$$

2.3.2 예제 기반 순위지표

예제 기반 순위지표는 분류된 레이블들의 순위를 평가하기 위한 지표이다. 원에러(One-error)는 인스턴스의 레이블과 관련되어 있지 않은 최상위 레이블의 빈도를 계산한다. 관련성이 없는 레이블이 분류될 경우 1로 반환한다. 원에러가 높을수록 관련되어 있지 않은 레이블의 수가 많이 분류되었다는 의미이다. 이는 모델을 다시 만들어 원에러 값이 낮게 나오도록 분류작업을 진행해야 한다.

$$\text{원에러} = \frac{1}{N}\sum_{i=1}^{N}\delta(\arg min r_i(\lambda)), \tag{2}$$
$$\lambda \in L, \delta(\lambda) = \begin{cases} 1 \text{ if } \lambda \notin Y_i, \\ 0 \; otherwise. \end{cases}$$

랭킹로스(Ranking Loss)는 분류의 결과에서 목표와 관련 없는 레이블이 관련된 레이블보다 높은 순위에 존재하는 횟수를 나타낸다. 랭킹로스의 값이 클수록 분류작업에서 관련 없는 많은 레이블이 높은 순위에 존재했다는 뜻이다.

$$\text{랭킹로스} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|Y_i||\overline{Y_i}|}|\{(\lambda_a, \lambda_b) \tag{3}$$
$$: r_i(\lambda_a) > r_i(\lambda_b), (\lambda_a, \lambda_b) \in Y_i \times \overline{Y_i}\}|$$

앞서 설명한 바와 같이 다중 레이블은 타깃변수가 두 개 이상이기 때문에 단일 레이블 분류보다 알고리즘이 복잡하다. 또한, 이에 따른 결과를 해석하는 평가지표도 복잡할 뿐만 아니라 더 많은

평가지표가 있다. 기존 연구에서 어떤 평가지표를 주 지표로 사용할지에 대해 많이 논의되었으며, 최근에도 해밍로스가 직관적이고 해석이 쉽다는 측면에서 주로 사용되고 있다. Pereira et el.(2018)은 다양한 지표들이 있는 가운데, 평가방식은 다르지만 하나의 데이터에 관한 결과인 평가지표 간의 상관성에 관해 연구하였다. 그리고 평가지표의 상관성을 고려하여 해밍로스가 종합적으로 가장 주요한 지표임을 재차 강조하였으며, 이를 기반으로 함께 참고하면 좋은 지표들의 조합을 제시하였다. Pereira *et al.*(2018)의 다중 레이블 지표 간의 상관관계 분석결과를 0.8 이상 그리고 0.9 보다 큰 값으로 구분하여 평가지표 간의 관련성을 살펴보았다. 그 결과 다중 레이블 분류에 대한 평가지표 조합을 제시하였는데, 첫째, 해밍로스, 커버리지, 랭킹로스와 정확도를 함께 사용한다면 레이블의 하위 집합들을 평가하는 지표들을 독립적으로 평가할 수 있다. 이는 해밍로스, 커버리지, 랭킹로스는 다른 지표들에 영향을 받지 않기 때문이다. 둘째, 해밍로스, 커버리지, 랭킹로스, 원에러, 평균 정밀도의 조합이다. 정확도를 대신하여 원에러와 평균 정밀도를 사용하는 이유는 두 지표가 서로 상관관계를 가지고 있고, 연구자들이 다중레이블 분류에서 우선순위를 평가할 때 효율적이기 때문이다. 셋째, 해밍로스, 마이크로 평균 정밀도, 마이크로 평균 재현율, 마이크로 평균 F-척도, 정확도, 랭킹로스를 함께 사용하는 조합이다. 이 조합은 F-척도가 정밀도와 재현율의 조화평균이라는 점에서 함께 사용하고, 다른 지표들은 다중 레이블 분류 결과를 F-척도, 정밀도, 재현율과는 다른 관점에서 평가하기 위해 사용한다. 본 연구는 위의 평가지표 조합과 기존 문헌을 바탕으로 해밍로스, 원에러, 랭킹로스를 평가지표로 채택하였다.

2.4 다중 레이블 속성선택

높은 차원과 차원의 저주라는 말은 입력 공간과 관련이 있다. 입력 공간의 차원이 클수록 데이터를 분석하는 것에 많은 시간이 필요할 뿐만 아니라 노이즈가 발생할 수 있기 때문이다(Dash and Liu, 1997). 이러한 문제는 단일 레이블 분류 학습과 다중 레이블 분류 학습 모두 영향을 준다(Liu, and Motoda, 2012). 그러므로 입력 공간의 차원 연구는 머신러닝 분야에서 중요하며, 입력 공간의 차원을 축소하는 것을 목적으로 둔다. 속성선택은 입력 공간의 데이터에서 중복되거나 관련성이 없는 속성을 제거하고 중요한 속성만을 선택하여 데이터 세트를 재구성하는 방법이다. 이에 속성선택 방법은 속성값을 훼손하지 않고 데이터의 보존이 가능하다는 장점이 있다. 그리고 고차원 데이터를 학습하고 분석할 때, 차원의 저주 문제를 해결할 수 있기에 데이터의 양이 많아질수록 더욱 각광받는 연구분야이다. 특히, 기존 연구에 따르면 속성선택은 분석정확도를 저해하는 현상 없이 속성을 제거할 수 있다고 증명되었다(Spolaôr *et al.*, 2012; Zhao *et al.*, 2011). 따라서 분류 알고리즘이 데이터를 학습하고 분석하는데 걸리는 시간을 줄이고, 노이즈를 제거함에 따라 예측 정확도를 높이는 등 효율성과 효과적인 측면에서 유용하다고 보고된다(Guyon and Elisseeff, 2003). 이러한 속성선택 방법은 데이터 전처리 단계에서 많이 사용되는데, 속성값을 줄이는 작업뿐만 아니라 속성선택 전후의 데이터 세트 비교를 통해 타깃변수에 영향을 주는 주요한 요인을 탐색하는데 사용된다. 속성선택의 초기 연구의 잠재 디리클레 할당(Latent Dirichlet allocation, LDA), 정준상관분석(canonical-correlation analysis, CCA) 방법은 하나의 속성만을 관찰하도록 설계가 되었다. 이는 타깃변수를 구별하고 특정 속성을 제거할 수 있는 속성 순위를 얻으며, 각 입력 속성이 얼마나 유용한 정보를 전달하는지 알려준다(Priyadarsini *et al.*, 2015). 이와 달리 최근 방법은 정보 이득(Information Gain, IG)이나 릴리프에프(ReliefF, RF)와 같은 필터 방법이 있다. 필터 방법은 포괄적으로 모델을 평가하여 사용하거나 래퍼 방식과 같이 학습 알고리즘에 의해 속성값을 평가한다(Azhagusundari and Thanamani, 2013). 그

러나, 다중 레이블 분야의 속성선택은 타깃변수가 두 개 이상이기 때문에 단일 레이블 속성선택보다 어렵고 복잡하다. 다중 레이블 속성선택 기법은 데이터 모델 관점에서 두 가지로 구분할 수 있다. 첫째, 단일 레이블 연구에서 사용되는 속성선택 기법을 적용하는 문제변환방식 기법과 둘째, 직접 다중 레이블 속성선택 알고리즘을 개발한 알고리즘 적용방식이다(Spolaôr *et al.*, 2013).

문제변환방식 기법을 적용한 연구는 이진연관성, 레이블 멱 집합, PPT(Pruned Problem Transformation) 등 다중 레이블 데이터를 단일 레이블 데이터 세트로 변환하여 최신 단일 레이블 속성선택 기법을 적용한다. Doquire and Verleysen(2011)은 PPT 방법을 기반으로 다차원 상호 정보 기반의 속성선택 방법의 성능을 평가하였다. 이후 Reyes *et al.*(2015)은 PPT 방법을 기반으로 RF 알고리즘을 사용하여 각 속성의 가중치를 부여하는 방법을 제시하였다. Spolaôr *et al.*(2013)은 LP와 BR을 사용하여 문제를 변환한 다음 IG와 RF 알고리즘을 적용하여 네 가지 방법의 성능을 비교하였다.

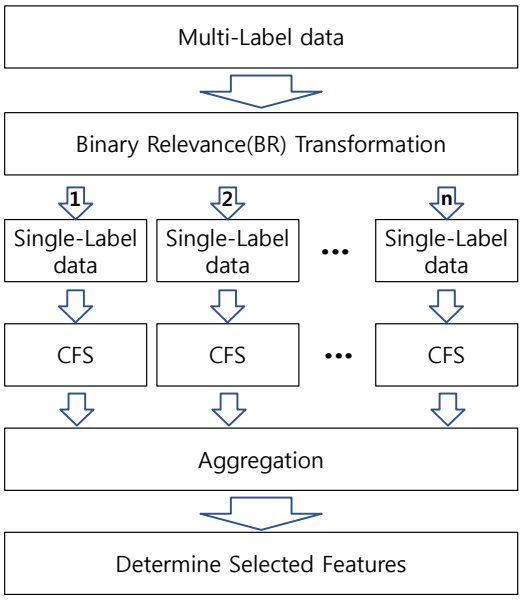
이와 달리 기존의 많은 연구는 다중레이블 분류 속성선택을 위한 알고리즘을 직접 개발하는 알고리즘 적용방식을 많이 사용하였다. Kong and Yu(2010)은 독성을 분석하고, 독성에 맞는 약물을 분류 및 예측하기 위해 그래프를 기반의 분류 알고리즘과 다중 레이블 속성선택 방법을 사용하는 모델을 제시하였다. 약물은 환자에 따라 해롭고 부작용이 있어, 환자에게 유익한지 또는 유익하지 않은지에 대해 동시에 결정하고 예측되어야 한다. 이를 위해 약물이 가지고 있는 기능을 찾는 것을 목표하여 속성선택 알고리즘을 적용하였다. 연구결과 제안하는 모델의 속성선택 알고리즘은 힐베르트 슈미트(Hilbert-Schmidt) 독립 점수의 값을 최적화하도록 하여 우수한 결과를 보여주었다. Lee and Kim(2013)은 정보의 상호작용 알고리즘에 따라 속성을 추출하는 방법을 제시했다. 적용된 속성선택은 선택된 기능의 하위 집합과 레이블 집합 간의 상호작용 정보를 최대화하여, 상호작용 정보에 대한 값을 최대치로 달성하기 위해 두 세트 간의 정보를 일련의 다변수 상호 정보 용어로 분해하였다. 연구의 주된 목적은 데이터 세트에 대해 함수의 자연적 파생을 찾는 것이다. 기존의 속성선택 방법은 주로 단일 레이블 문제로 변형한 다음 속성선택 프로세스를 적용하였지만, 연구에서 제안하는 방법은 직접 알고리즘을 적용하여 별도의 데이터 전처리를 하지 않아도 된다는 장점이 있다. 다중 레이블 분류를 위한 속성선택 방법으로 다중 레이블 속성선택과 미미틱 알고리즘(Memetic Search Algorithm)을 적용한 기법이 있다(Lee *et al.*, 2019; Zhu *et al.*, 2010). 이러한 혼합 유전알고리즘은 유전자 정보를 바탕으로 진보된 알고리즘 형태라 할 수 있다. 이는 하나의 모집단에 적합한 함수 그리고 개체 사이의 돌연변이 및 유전과 같이 진화론에 입각한 원리를 기반으로 최적의 값을 찾을 수 있는 기능이며 성능이 우수하다. 다만 처음부터 최적값에 수렴하거나, 계산하는 시간이 오래 걸린다는 단점을 가지고 있다. Lee and Kim(2015)은 기존의 유전알고리즘 모델의 한계를 극복한 속성선택 알고리즘을 제안하였다. 기존 알고리즘보다 효율성이 높고, 효용성에 따라 관련성이 높고 낮은 특징을 추출하는 속성선택 모델로 우수한 성능을 가지고 있다. 기존의 연구에서는 각 레이블에 대한 속성 의존성의 반복 계산을 포함하였지만, Lee and Kim(2017)의 SCLS(Scalable Criterion for Large label Set : 대규모 레이블 집합 확장 가능 기준)는 의존성 계산의 반복성을 배제함으로써, 속성을 선택하는 효율성을 극대화하였다. 연구결과 SCLS는 특정한 조건에서 관련성을 정확하게 평가하는 관련성 평가 프로세스로, 기존 연구의 속성선택 방법보다 학습 정확도와 분석시간을 줄여, 효율성을 향상한 모델이라 보고하였다. 본 논문에서 적용한 CFS 알고리즘의 경우 Jungjit *et al.*(2013)에 의해 문제변환 방식없이 알고리즘 적용방식의 ML-CFS(Multi-Label Correlation Based Feature Selection Method) 이 제안되었으며, Braytee *et al.*(2017) 또한 CFS를 적용하여, NMF

(Non-negative Matrix Factorization)를 기반으로 한 CMFS(Correlated-and Multi-label Feature Selection method)를 제안하였다.

위와 같이 다중 레이블 속성선택 연구는 문제변환 방법을 적용한 연구보다 알고리즘 적용방식의 연구가 많다. 그러한 이유는 다중 레이블 관련 연구가 전산학을 기반으로 하기에 직접 알고리즘을 개발하거나, 기존의 단일 레이블 알고리즘을 다중 레이블 알고리즘에 직접 적용 가능하도록 알고리즘을 개선하는 데 의의를 둔다. 이와 달리, 본 연구는 문제변환방식을 적용하여 효율적으로 의사결정을 지원하는 속성선택 방법을 제안하고자 한다. 특히, 기존 연구와 달리 부분집합 탐색 방법을 최우선 탐색 방법(Best-First Search)을 적용하여 의사결정자가 속성값의 순위 혹은 별도의 파라미터를 지정하지 않아도 최적의 결과를 도출하는지 기존 알고리즘과 비교한다.

Ⅲ. 개념적 모델 제시

본 연구는 비즈니스 환경에 적합한 다중 레이블 기반의 의사결정을 위한 속성선택 방법을 제안한다. 다중 레이블 데이터를 단일 레이블로 변환하는 문제변환 방식 중 하나인 이진 연관성 이론을 기반에 단일 레이블 연구에서 사용되는 상관관계 속성선택 방법을 적용하여 새로운 방법을 제안한다. 이는 이진 연관성 접근 기반의 상관관계 속성선택 모델을 CFS-BR이라 명명한다. 특히, 기존의 다중레이블 속성선택 알고리즘은 필요한 변수의 개수를 직접 정하거나, 연구자가 지정하는 파라미터 값에 따라 결과가 상이하므로 예측 성능의 편차가 크다. 즉, 기존의 다중레이블 속성선택 알고리즘의 최적값을 찾기 위해서는 파라미터 값을 여러차례 조정해야할 뿐만 아니라, 데이터 세트의 특징에 따라 값이 달리 적용되므로 빠르고 정확한 의사결정을 지원하는 데 한계가 있다. 반면, 우리가 제안하는 CFS-BR은 최적의 변수를 최우선 탐색 방법으로 변수를 선별하므로 의사결정에 필요한 전처리 과정을 간소화하며, 효율적인 의사결정 지원을 도울 것으로 기대한다.



〈그림 6〉 이진관련성 접근 기반의 상관관계 속성선택 모델

우리는 이를 검증하기 위해 기존의 속성선택 모델과 비교한다. Spolaôr *et al.*(2013)은 다중 레이블의 문제변환방법을 적용한 속성선택 모델을 제안하였다. 이진 연관성(BR)과 레이블 멱 집합(LP) 방법을 기반으로 다중 레이블 세트를 단일 레이블 세트로 변환하고, 기존의 단일 레이블 속성선택 알고리즘인 IG와 RF를 적용한 4가지 모델을 제시하였다. 4가지 모델은 IG-BR, RF-BR, IG-LP, RF-LP로 <표 3>과 같다. 해당 연구의 결과 IG-BR, IG-LP 모델보다 RF-BR, RF-LP 모델이 우수한 결과를 나타냈다. 따라서 본 연구는 RF-BR, RF-LP 모델 알고리즘을 벤치마킹하였다. 또한, 레이블 세트를 문제변환방식으로 변형하지 않고 알고리즘을 직접 적용하는 방식인 SCLS를 벤치마킹하였다(Lee and Kim, 2017). SCLS 알고리즘을 적용하기 위해서는 데이터의 전처리가 필요한데, 해당 알고리즘은 속성값이 이진 분류인 경우만 적용할

〈표 3〉 벤치마킹 속성선택 알고리즘

알고리즘 이름	저자	방식	약자
RF based on BR approach : 이진 연관성 접근 기반의 릴리프 에프 속성선택 모델	Spolaôr *et al.*(2013)	문제변환방식	RF-BR
RF based on LP approach : 멱 집합 레이블 접근 기반의 릴리프 에프 속성선택 모델			RF-LP
IG based on BR approach : 이진 연관성 접근 기반의 정보 이득 속성선택 모델			IG-BR
IG based on LP approach : 레이블 멱 집합 접근 기반의 정보 이득 속성선택 모델			IG-LP
Scalable Criterion for Large label Set : 대규모 레이블 집합 확장 가능 기준	Lee and Kim(2017)	알고리즘 적용방식	SCLS

수 있기 때문이다. 이에 본 연구는 SCLS 벤치마킹 하기 위해 데이터 세트를 이진 분류의 값으로 변환하고, 출력 개수를 다섯 구간으로 나누어 그 중의 최적의 값을 추출하여 비교 값으로 선택하였다.

마지막으로 해당 논문에서 제시하는 CFS-BR과 RF-BR, RF-LP, SCLS의 속성선택 성능을 비교하기 위해 다중레이블 k-최근접 이웃 알고리즘을 기반으로 비교하였다. 이는 다중레이블 분류연구에서 일반적으로 많이 사용되는 분류기이다(Zhang and Zhou, 2007).

3.1 이진 연관성 문제

다중 레이블 학습과제를 L개라고 하면 각 레이블에 대해 하나씩 q개로 구분하는 독립된 이진 분류문제로 분해한다. 그러면 다중 레이블 데이터 세트 D는 q개의 독립된 이진 분류로 구성된다. 구성된 q개의 데이터 세트는 D_{yj}(j=1 ..., q) 로 나뉘는데, 각 이진 분류문제에 해당 레이블과 관련된 예제는 양수로 간주하고 다른 예제는 음수로 간주한다(Zhang *et al.*, 2014). 마지막 과정으로 새로운 다중 레이블 인스턴스를 분류하기 위해, 이진 연관성은 q의 독립 분류를 통해 옳게 예측된 레이블의 집합을 출력한다. 이진 연관성은 레이블 집합 L의 크기 q와 선형으로 비례하며, q가 크지 않은 것이 보다 적합하다(Zhang *et al.*, 2018).

3.2 상관관계 속성선택

상관관계 기반의 속성선택 기법은 엔트로피와 카이스퀘어를 이용하여 다른 속성선택 방법보다 성능이 우수한 것으로 알려졌다. 특히, 상관관계 기반의 속성선택은 속성의 하위 집합을 평가하는 것이 특징이다(Liu *et al.*, 2002). 상관관계 속성선택 방법은 변수 집합을 평가하는 과정과 탐색하는 과정으로 구분할 수 있다. 우선 변수 집합 평가 방법으로 Merits 값을 사용한다.

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}}$$

$Merit_s$는 k개의 변수를 가진 집합 S에 대해 값을 가진다. $\overline{r_{cf}}$ 는 속성과 레이블 간의 상관관계를 나타내는 값들의 평균이고, $\overline{r_{ff}}$는 속성값들 사이의 상관관계를 나타내는 평균값이다. 대체로 분류 방법에 있어 레이블값은 명목변수이므로 상호정보를 고려하는데, 상호정보는 최댓값의 범위가 넓어 표준화시킨 대칭적 불확실성을 이용한다. 대칭적 불확실성은 엔트로피를 이용하여 상호정보 값을 표준화시킨 형태이다. 엔트로피는 확률변수의 불확실성에 대한 척도라 할 수 있다. 즉, 엔트로피와 상호정보는 대칭적 성질을 가지며 더 많은 값

을 가지고 있는 방향으로 표준화 값을 구성하여 두 변수 간의 상관관계를 측정한다.

최적의 속성을 선택하기 위해서는 모근 속성의 집합을 평가할 수도 있지만 모든 가능한 조합을 탐색하는 것은 어렵다. 따라서 데이터에 적합한 탐색 방법을 모색해야 한다. 탐색 방법으로는 후진 제거, 전진선택, 지역탐색, 유전알고리즘 등을 고려한 방법이 있으며 본 연구에서 적용한 최우선 탐색방법은 지역탐색법에서 가장 효율적이라 알려졌다(Doshi, 2014). 최우선 탐색 방법은 단계적으로 변수를 선택하면서 최적의 값이 향상되지 않을 때 이전 단계에서 차선 순위의 변수를 탐색하여 최적의 값을 선택한다(Hall and Holmes, 2003).

3.3 데이터 세트

본 연구는 두 종류의 실험을 통해 CFS-BR의 우수성을 입증하고자 한다. 첫째, Mulan 데이터 세트(mulan.sourceforge.net)의 22가지의 데이터 중 기존 연구에서 많이 인용한 오디오, 이미지, 텍스트 등을 기반으로 구성된 7개의 다중 레이블 자료를 얻었다. 이 자료는 다중 레이블 커뮤니티에서 널리 쓰이고 새로운 분석방법을 검증하는 데 많이 사용되었다(Tsoumakas *et al.*, 2011).

둘째, 온라인 리뷰에 대한 감정분석을 통해 CFS-BR의 실무적 가치를 검증하고자 하였다. 감정을 분류하기 위해 뮤지컬 데이터를 수집하였다. 국내 티켓예매 사이트(인터파크)를 통해 누적 인원수가 가장 많은 뮤지컬 8개를 선정하여 모든 데이터를 수집하였다. 41,000건의 뮤지컬 리뷰 중에 (1) 예매자가 확인된 데이터를 전처리한 결과 28,000여 건으로 줄어들었다. 그리고 (2) 200자 이상의 데이터만을 대상으로 분류하여 1,500건을 분석 대상으로 선정하였다. 그리고 리뷰 데이터에 대한 감정을 할당하기 위해, 한국말을 모국어로 하는 대학생 3명에게 인간의 기본 감정이라고 여겨지는 행복감, 공포, 놀람, 슬픔, 화남, 혐오를 분류하도록 의뢰하였다. 특히, 최소 두 가지 이상의 감정이 내포된 데이터를 선별하고, 분석을 실시하였다(Giatsoglou *et al.*, 2017). <표 4>는 각 데이터에 대한 설명으로 해당 도메인, 인스턴스, 변수, 레이블(타깃변수)의 수 등을 기술하였다.

Ⅳ. 분석결과

본 논문에서 제안하는 모델인 CFS-BR과 벤치마킹한 알고리즘(RF-BR, RF-LP, SCLS) 간의 속성선택 성능을 비교한 결과는 다음과 같다. 비교를 위한 학습 알고리즘으로는 다중 레이블 k-최근접 이웃 알고리즘을 사용하여 10-겹 교차검증으로 분석하였다. SCLS의 경우 속성선택을 위해 추출하고자 하는 변수의 개수를 입력하여야 한다는 점에서 다섯 구간으로 구분하여 가장 낮은 값을 추출하였다. 종합적인 결과는 <표 5>와 같다.

〈표 4〉 데이터 세트

데이터 이름	변수				
	도메인	인스턴스	명목변수	연속형 변수	레이블(Y)
Birds	Audio	645	2	258	19
Fags	Image	194	9	10	7
Genbase	Biology	662	1186	0	27
Yeast	Image	2417	0	103	14
Image-data	Image	600	0	294	5
TMC_500	Text	15000	500	0	22
Emotion	Music	593	0	72	6
Musical	Text	1500	0	1000	6

〈표 5〉 테스트 데이터 결과

		본 연구	RF-BR	RF-LP	SCLS
해밍로스	Bird	**0.048**	0.053	0.049	0.051
	Emotion	**0.203**	0.211	0.220	0.214
	Flag	**0.228**	0.267	0.243	0.290
	Genbase	**0.015**	0.033	0.042	0.041
	Image-data	**0.158**	0.166	0.170	0.170
	TMC2007	0.059	**0.058**	0.059	0.061
	Yeast	**0.171**	**0.171**	0.175	0.173
원에러	Bird	**0.370**	0.578	0.472	0.529
	Emotion	**0.290**	**0.290**	0.300	0.293
	Flag	**0.125**	0.151	0.203	0.145
	Genbase	**0.151**	0.418	0.620	0.568
	Image-data	**0.273**	0.298	0.313	0.315
	TMC2007	0.190	**0.187**	0.189	0.200
	Yeast	**0.177**	0.182	0.185	0.182
랭킹로스	Bird	**0.092**	0.175	0.133	0.158
	Emotion	**0.132**	0.166	0.171	0.182
	Flag	**0.159**	0.167	0.178	0.204
	Genbase	**0.008**	0.051	0.098	0.086
	Image-data	**0.129**	0.137	0.143	0.155
	TMC2007	0.036	**0.034**	0.035	0.037
	Yeast	**0.123**	0.124	0.127	0.127

해밍로스 결과는 낮은 값이 우수하다고 평가하는데, 레이블의 세트의 크기에 따라 해밍로스에 영향을 미치기도 한다. 그렇지만 데이터 간의 비교가 아닌, 속성선택 알고리즘 간의 비교가 목적이기에 레이블 세트의 크기에 따른 차이는 본 연구의 실험과 무관하다 할 수 있다. 연구결과 이진 연관성 기반의 상관관계 속성선택 모델인 CFS-BR은 일곱 개의 데이터 세트 중 다섯 개 데이터 세트 (Bird, Flag, Genbase, Image-data, Yeast) 에서 우수한 결과를 보여준다. TMC 2007의 경우 RF-BR 모델이 우수한 결과를 보여주며, Yeast의 경우 CFS-BR 모델과 RF-BR 모델의 결과가 같은 것을 보여준다.

원에러는 레이블의 긍정오류의 비율을 나타내는 지표로, 작은 값의 경우에 알고리즘 모델의 성능이 우수하다고 할 수 있다. 실험 결과 다른 속성선택 알고리즘과 비교할 때, 7개의 데이터 세트 중에 5개의 데이터 세트(Bird, Emotion, Flag, Genbase, Image-data) 에 대해 우수한 결과를 보여준다. 그 외에 Emotion 데이터 세트의 경우 CFS-BR과 RF-BR이 같은 값을 나타내었으며, TMC 2007의 경우 RF-BR 모델이 조금 더 우수한 결과를 보여주었다. 이는 해밍로스의 결과와 같다고 할 수 있다.

랭킹로스는 분류기의 성능을 예측하는 데 있어, 관련이 없는 레이블이 관련이 있는 레이블보다 이전에 얼마나 나타나는지 계산하는 지표이다. 따라서 랭킹로스는 값이 적을수록 우수한 모델이라 평가할 수 있다. 랭킹로스의 결과는 해밍로스, 원에러의 결과와 비슷한 흐름을 보여준다. 랭킹로스의 경우도 앞서 다른 지표들의 결과와 마찬가지로 TMC 2007은 RF-BR 모델이 우수한 결과를 나타내지만, 나머지 데이터 세트의 경우 CFS-BR 모델이 우수한 결과를 보여준다. 앞서 언급한 바와 같이

해밍로스는 다중 레이블 연구 분야에서 모델을 평가하는 가장 기본적이고 보편적인 지표이다. 이는 직관적으로 해석할 수 있으며, 계산과정이 간단하므로 해석이 명료해 많은 연구에서 사용된다. 본 연구에서 해밍로스 외에 다른 측정항목과 함께 비교 평가한 이유는 결과에 대한 공정성을 확보하기 위함이다. 또한, 다중 레이블 학습 결과를 해석하는 데 있어 다각적이고 광범위하게 평가하고자 하였다.

테스트 데이터 세트에 대한 실험 결과를 요약하면 다음과 같다. 첫째, 본 연구에서 제안하는 CFS-BR 모델은 벤치마킹한 다른 속성선택 방법과 비교하였을 때, 일곱 개의 데이터 중에 여섯 개의 데이터에서 더 우수한 결과를 보여주었다. 둘째, 텍스트 데이터인 TMC 2007의 경우 RF-BR 모델이 비교적 우수한 결과를 보여주었다. 이는 일곱 개의 다른 데이터 세트와 비교하였을 때 유일하게 TMC 2007만이 텍스트 데이터인 점을 고려하였을 때, 텍스트에 대한 데이터의 정형화 방식에 따라 속성선택의 우수성이 다를 수 있다는 점을 고려할 수 있다. 즉, 텍스트 기반의 데이터를 숫자로 임베딩 하는 과정은 말뭉치(Bag-of-Words)의 경우 단어의 빈도수, TF-IDF 등이 있고, Word2vec, GloVe 등 다양한 워드 임베딩 기법들이 있으므로 텍스트의 전처리 과정에 따라 다른 차이가 있는 것으로 예상한다.

기존 머신러닝 연구자들에 따르면 다중 레이블 학습뿐만 아니라, 차원 축소기법의 일환인 속성선택 방법에서도 문제변환 방식보다 알고리즘 채택 방법을 선택한 방법이 레이블 간의 관계를 분석하는 데보다 자연스럽고 우수하다고 여겨진다. 그렇지만, 본 연구에서 제안하는 CFS-BR 모델은 문제변환 방식을 적용했다는 한계점에도 불구하고, 기존 연구보다 우수한 속성선택 성능을 보여준다. 기존의 벤치마킹 속성선택 방법인 RF 알고리즘 기반의 속성선택 방식은 속성값에 대한 순위 값을 나타낸다. 그러므로 연구자가 순위 값에 따라 임의로 속성값을 선택하는 기준을 정의해야 하는 번거로움이 있다. SCLS 알고리즘 또한 원하는 속성값의 수를 사용자가 직접 입력하여 선택하는 속성선택 알고리즘이다. 또한, 속성값이 이진 분류인 경우에만 계산할 수 있으므로 연속형 변수나 실수의 경우 SCLS 알고리즘에 바로 적용하는 것은 어렵고, 전처리 작업을 수행해야 한다. 이러한 전처리 과정은 인위적으로 자료의 변이가 일어날 수 있으므로 알고리즘의 한계점이라 볼 수 있다. 이렇듯 벤치마킹한 SCLS과 RF 기반의 속성선택 알고리즘은 최적의 결과값을 구하기 위해서는 사전작업이 필요하고, 실험시간이 소요된다. 반면, CFS-BR은 이진 분류 데이터뿐만 아니라 연속형 변수와 같은 모든 변수에 대해 즉각적인 계산이 가능하다. 즉, 특별한 전처리 과정 없이 자연스러운 속성선택 적용이 가능하다. 따라서 CFS-BR을 적용한 다중 레이블 분류 모델은 의사결정을 효과적으로 지원할 수 있다.

마지막으로, 기존의 다중 레이블 연구에서 적용되는 오픈 데이터가 아닌 현실 세계에서 수집한 데이터를 대상으로, CFS-BR 모델을 검증하고자 한다. 이는 실제 비즈니스 의사결정과 같은 유사한 환경에서 의사결정자의 선택을 돕거나 대신할 수 있는지에 대한 현실 세계 데이터를 기반으로 검증하기 위함이다. 앞서 CFS-BR과 RF-BR 모델이 비교적 좋은 성과를 보여줌에 따라 두 속성선택 간의 비교를 통해 CFS-BR 모델의 우수성을 검증하고자 한다. 또한, 앞서 실험에서 적용한 다중 레이블 k-최근접 이웃 알고리즘뿐만 아니라 다양한 문제변환 방식(BR, CC, RAkEL, CLR, LP) 알고리즘으로 다각적인 검토를 하며, 새로운 알고리즘 적용방식으로는 BP-MLL을 추가하여 성과를 비교하였다. 연구결과는 <표 6>과 같다.

해당 데이터는 텍스트 데이터로 비정형데이터인 뮤지컬 리뷰이다. 전처리 과정은 해당 리뷰를 말뭉치기법(bag-of words)을 통해 TF-IDF(Term Frequency - Inverse Document Frequency)로 데이터를 변환하였다. 특히, 앞선 실험에서 텍스트 데이터의 경우, RF-BR 모델이 우수한 결과를 보여준

〈표 6〉 뮤지컬 데이터 결과

	해밍로스			원에러			랭킹로스		
	Original	CFS_BR	RF_BR	Original	CFS_BR	RF_BR	Original	CFS_BR	RF_BR
BR	0.163	**0.142**	0.155	0.235	**0.123**	0.102	0.177	**0.108**	0.120
CC	0.166	**0.140**	0.155	0.178	**0.151**	0.108	0.133	**0.118**	0.118
RAkEL	0.214	**0.152**	0.164	0.147	**0.136**	0.114	0.112	**0.104**	0.117
CLR	0.279	**0.271**	0.302	0.110	**0.100**	0.101	0.099	**0.089**	0.121
LP	0.186	**0.150**	0.116	0.233	**0.176**	0.135	0.161	**0.125**	0.129
BP-MLL	0.117	**0.094**	0.155	0.055	**0.050**	0.092	0.050	**0.049**	0.118
ML-kNN	0.143	**0.128**	0.154	0.091	**0.081**	0.092	0.091	**0.086**	0.117

것을 고려할 때, 텍스트 데이터 전처리의 절차와 방법이 중요하다는 것을 인식할 수 있다. 연구결과는 다음과 같다.

해밍로스는 작은 값일수록 분류성능의 우수성을 나타내는 지표로 모든 분류기에서 CFS-BR 모델이 우수한 것을 볼 수 있다. 특히, 알고리즘 적용방식인 BP-MLL와 다중 레이블 k-최근접 이웃 알고리즘이 다른 분류기에 비해 문제를 잘 분류하는 것으로 확인하였다.

원에러의 경우 레이블의 긍정오류를 나타내는 비율로 낮은 값이 분류성능이 우수하다. 원에러의 결과 또한 해밍로스의 결과와 비슷한 양상을 보이는데, CFS-BR 모델이 가장 우수한 결과를 보여주며 알고리즘 적용방식이 다른 분류기보다 더 분류성능이 우수한 것으로 확인하였다.

마지막으로 랭킹로스의 값을 살펴보면 다음과 같다. 기존의 다른 측정항목들과 마찬가지로 CFS-BR 모델이 가장 우수함을 보여주었다. 결과적으로 기존 연구에서 사용되는 오픈 데이터뿐만 아니라 한글 온라인 리뷰를 분석하는데 CFS-BR 모델이 우수한 속성선택 방법임을 증명하였다.

Ⅴ. 결 론

본 연구는 효과적인 의사결정을 위한 다중 레이블 분류 기반의 속성선택 방법을 제안하였다. 기존 컴퓨터 공학에서 연구되는 알고리즘은 경영실무자 또는 연구자가 원하는 속성값의 개수를 입력하여 추출하는 방식이 많다. 이에 반해, CFS-BR은 최적의 값을 자동적으로 선별하여 실무자 또는 연구자에게 최적의 솔루션을 제시할 수 있다. 각 방법에 따른 장단점이 있지만, 속성값의 개수를 선택하지 않고 자동으로 산출하는 방법은 비즈니스 환경에서 더 효과적일 것이다. 더욱이 사물인터넷이나 온디바이스의 발달로 인해 기업은 더 많은 데이터를 보유하는데, 데이터 기반의 의사결정을 지원하기 위해서는 자동으로 최적의 속성값을 찾고 해답을 제시해야 하므로 CFS-BR은 의사결정자가 빠른 선택과 결정을 내리는 데 도움이 될 것이다.

본 연구의 학술적 기여는 다음과 같다. 효과적인 의사결정을 위한 다중 레이블 분류 속성선택 방법과 분류모델을 제시하였다. 이는 경영정보, 재무, 마케팅, 인사조직 등 비즈니스의 특정 문제를 해결하기 위한 의사결정 지원 프로세스를 구축하기 위한 기초자료로 다른 연구 분야에 응용되길 기대한다. 특히, 다중 레이블 속성선택은 디지털 트랜스포메이션으로 넘쳐나는 데이터를 처리, 분석하는 데 유용하기에 데이터 기반 의사결정을 지원하고, 새로운 데이터 세트를 대상으로 하는 연구에 기여할 수 있다.

더욱이 현실 세계에 존재하는 다양한 인간의 활동들을 다중 레이블 데이터 세트로 변환하여, 적용이 가능하기에 CFS-BR은 다양하게 적용될

수 있다. 예를 들어, 개인의 다양한 성향이나 특징이 속성값이라 하면, 그에 대응하는 레이블은 하루 동안 개인이 소비하는 물품이라 할 수 있다. 이는 개인의 특징이나 습관에 따라 하루에 소비되는 물품을 효과적으로 예측할 뿐만 아니라 미래의 행동을 예측하는 데도 적용할 수 있다. 따라서 본 연구에서 제안하는 CFS-BR을 적용한 다중 레이블 분류는 경영학에서의 다양한 문제를 새로운 견해로 분석하는데 기초자료를 제공한다.

국내 기업인 카카오는 카카오T를 통해 고객의 다양한 정보를 수집하여 비즈니스에 활용하였다. 카카오T는 1,600만 명의 사용자 데이터를 수집하여 교통 상황이나 택시기사들의 빈 차 상황, 차고지, 주거지, 운행패턴 등을 분석하였다. 그 결과 기존의 이탈고객을 방지하고 고객만족도를 높였다. 넷플릭스 또한 3,000만 명 이상의 시청자 행동을 분석하여 새로운 사업에 투자하는 등 소비자들에게 새로운 서비스를 제공하고, 새로운 비즈니스 기회를 발굴했다. 즉, 인공지능 기반의 머신러닝은 의사결정을 효과적으로 도우며, 나아가 새로운 비즈니스 기회를 창출하므로 모든 산업에 적용될 것이다. 이러한 변화를 고려할 때의 본 연구의 실무적 의의는 다음과 같다. 첫째, CFS-BR은 기업의 데이터 관리비용을 줄이는 데 기여할 수 있다. 기업의 데이터는 기업의 자산이기 때문에 이를 확보하고 저장하는 것은 중요한 업무이다. 그렇지만 모든 데이터를 관리하고 저장하는 것은 무의미하며, 의미 있는 정보만을 관리하고 저장하여야만 한다. CFS-BR은 데이터의 의미 있는 정보만을 저장하고 관리하기 때문에 데이터를 관리하는 비용을 줄일 수 있다. 둘째, CFS-BR을 통한 의사결정은 의사 결정자에게 빠른 분석 결과를 제시하므로 유용하다. 특히, CFS-BR은 의미 있는 정보만 선택하고 관리하므로 데이터 크기가 작아진다. 이에 따라 분석 시간이 줄어들 뿐만 아니라 무엇이 특정 결과와 속성 간의 상관성을 알 수 있다. 마지막으로, 새로운 데이터에 대한 자동 분류에 유용하다. 자동분류는 새로 생성되는 데이터에 대해 자동으로 라벨링 되어야 한다. 그리고 필요한 데이터와 불필요한 데이터를 구분해야 한다. 이러한 기능을 수행하는데 CFS-BR은 유용할 것으로 예상된다. 특히, 추가적인 전처리 작업 없이 데이터를 직접 사용할 수 있으므로 실무적 가치가 높다.

참 고 문 헌

[1] 미아오쉬, 이재성, “음악감성 인식 정확도 향상을 위한 노이즈 제거 기술의 효과 비교 연구”, *인공지능인문학연구*, 제1권, 2018, pp. 97-123, Available at http://dx.doi.org/10.46397/JAIH.1.5.

[2] 민동영, 조성준, “코스피 상장 기업의 다중 레이블 분류를 위한 산업군 키워드 사전의 구축: 단어 임베딩 공간 사이의 선형변환학습을 중심으로”, *대한산업공학회 추계학술대회 논문집*, 2017, pp. 2426-2468.

[3] 임소라, 권용진, “특허문서 필드의 기능적 특성을 활용한 IPC 다중 레이블 분류”, *인터넷정보학회논문지*, 제18권, 제1호, 2017, pp. 77-88.

[4] 임채현, 손민지, 김명호, “다중 레이블 분류를 활용한 안면 피부 질환 인식에 관한 연구”, *정보처리학회논문지/소프트웨어 및 데이터 공학*, 제10권, 제12호, 2021.

[5] 장수진, 위정아, 김영빈, “합성곱 신경망의 멀티 레이블 학습을 통한 한국 영화 포스터의 장르 예측”, *한국 HCI 학회 학술대회*, 2019, pp. 746-749.

[6] 정풀잎, 안현철, 곽기영, “텍스트 마이닝과 소셜 네트워크 분석을 이용한 스마트폰 디자인의 핵심속성 및 가치 식별”, *대한경영학회지*, 제32권, 제1호, 2019, pp. 27-47, Available at http://dx.doi.org/10.18032/kaaba.2019.32.1.27.

[7] Agrawal, R., A. Gupta, Y. Prabhu, and M. Varma, “Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages”, *22nd International Conference on World*

Wide Web, 2013, pp. 13-24, Available at http://dx.doi.org/10.1145/2488388.2488391.

[8] Ahmadi, Z. and S. Kramer, "A label compression method for online multi-label classification", *Pattern Recognition Letters*, 2018, pp. 64-71, Available at http://dx.doi.org/10.1016/j.patrec.2018.04.015.

[9] Azhagusundari, B. and A. S. Thanamani, "Feature selection based on information gain", *International Journal of Innovative Technology and Exploring Engineering*, Vol. 2, No. 2, 2013, pp. 18-21.

[10] Bogaert, M., J. Lootens, D. Van den Poel, and M. Ballings, "Evaluating multi-label classifiers and recommender systems in the financial service sector", *European Journal of Operational Research*, Vol. 279, No. 2, 2019, pp. 620-634, Available at http://dx.doi.org/10.1016/j.ejor.2019.05.037.

[11] Bromuri, S., D. Zufferey, J. Hennebert, and M. Schumacher, "Multi-label classification of chronically ill patients with bag of words and supervised dimensionality reduction algorithms", *Journal of Biomedical Informatics*, Vol. 51, 2014, pp. 165-175, Available at http://dx.doi.org/10.1016/j.jbi.2014.05.010.

[12] Cai, J., J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective", *Neurocomputing*, Vol. 300, 2018, pp. 70-79, Available at http://dx.doi.org/10.1016/j.neucom.2017.11.077.

[13] Chatterjee, A., U. Gupta, M. K. Chinnakotla, R. Srikanth, M. Galley, and P. Agrawal, "Understanding emotions in text using deep learning and big data", *Computers in Human Behavior*, Vol. 93, 2019, pp. 309-317, Available at http://dx.doi.org//10.1016/j.chb.2018.12.029.

[14] Corani, G. and M. Scanagatta, "Air pollution prediction via multi-label classification", *Environmental Modelling & Software*, Vol. 80, 2016, pp. 259-264, Available at http://dx.doi.org/10.1016/j.envsoft.2016.02.030.

[15] Dash, M. and H. Liu, "Feature selection for classification", Intelligent Data Analysis, Vol. 1, No. 3, 1997, pp. 131-156, Available at http://dx.doi.org/10.3233/IDA-1997-1302.

[16] de Morais, J. I., H. Q. Abonizio, G. M. Tavares, A. A. da Fonseca, and S. Barbon, "Deciding among fake, satirical, objective and legitimate news: A multi-label classification system", *The XV Brazilian Symposium on Information Systems*, 2019, pp. 1-8, Available at http://dx.doi.org/10.1145/3330204.3330231.

[17] Doquire, G. and M. Verleysen, "Feature selection for multi-label classification problems", *International Work-Conference on Artificial Neural Networks*, 2011, pp. 9-16.

[18] Doshi, M., "Correlation based feature selection (CFS) technique to predict student performance", *International Journal of Computer Networks & Communications*, Vol. 6, No. 3, 2014, pp. 197-206.

[19] Elisseeff, A. and J. Weston, "A kernel method for multi-labelled classification", *Neural Information Processing Systems*, 2002, pp. 681-687.

[20] Erevelles, S., N. Fukawa, and L. Swayne, "Big Data consumer analytics and the transformation of marketing", *Journal of Business Research*, Vol. 69, No. 2, 2016, pp. 897-904, Available at http://dx.doi.org/10.1016/j.jbusres.2015.07.001.

[21] Folorunso, S. O., S. G. Fashoto, J. Olaomi, and O. Y. Fashoto, "A multi-label learning model for psychotic diseases in Nigeria", *Informatics in Medicine Unlocked*, 2020, Available at http://dx.doi.org/10.1016/j.imu.2020.100326.

[22] Fujii, M., H. Sakaji, S. Masuyama, and H. Sasaki,

"Extraction and classification of risk-related sentences from securities reports", *International Journal of Information Management Data Insights*, Vol. 2, No. 2, 2022, 100096.

[23] George, G., M. R. Haas, and A. Pentland, "Big data and management", 2014, Available at http://dx.doi.org/10.5465/amj.2014.4002.

[24] Gera, M. and S. Goel, "Data mining-techniques, methods and algorithms: A review on tools and their validity", *International Journal of Computer Applications*, Vol. 113, 2015, pp. 22-29.

[25] Ghamrawi, N. and A. McCallum, "Collective multi-label classification", *The 14th ACM international conference on Information and Knowledge Management*, 2005, pp. 195-200.

[26] Giatsoglou, M., M. G. Vozalis, K. Diamantaras, A. Vakali, G. Sarigiannidis, and K. C. Chatzisavvas, "Sentiment analysis leveraging emotions and word embeddings", *Expert Systems with Applications*, Vol. 69, 2017, pp. 214-224, Available at http://dx.doi.org/10.1016/j.eswa.2016.10.043.

[27] Gupta, A., P. Panagiotopoulos, and F. Bowen, "An orchestration approach to smart city data ecosystems", *Technological Forecasting and Social Change*, Vol 153, 2020, 119929, Available at http://dx.doi.org/10.1016/j.techfore.2020.119929.

[28] Gupta, P., T. K. Sharma, and D. Mehrotra, *Label Powerset Based Multi-label Classification for Mobile Applications*, In Soft Computing: Theories and Applications, Springer, Singapore, 2019.

[29] Guyon, I. and A. Elisseeff, "An introduction to variable and feature selection", *Journal of Machine Learning Research*, Vol. 3, No. 3, 2003, pp. 1157-1182.

[30] He, W., F. K. Wang, and V. Akula, "Managing extracted knowledge from big social media data for business decision making", *Journal of Knowledge Management*, 2017.

[31] Henrique, B. M., V. A. Sobreiro, and H. Kimura, "Literature review: Machine learning techniques applied to financial market prediction", *Expert Systems with Applications*, Vol. 124, 2019, pp. 226-251, Available at http://dx.doi.org//10.1016/j.eswa.2019.01.012.

[32] Jabreel, M. and A. Moreno, "A deep learning-based approach for multi-label emotion classification in tweets", *Applied Sciences*, Vol. 9, No. 6, 2019, pp. 1-16, Available at http://dx.doi.org/10.3390/app9061123.

[33] Jiang, A., C. Wang, and Y. Zhu, "Calibrated rank-svm for multi-label image categorization", *IEEE International joint conference on Neural Networks*, 2008, pp. 1450-1455, Available at http://dx.doi.org/10.1109/IJCNN.2008.4633988.

[34] Jungjit, S., M. Michaelis, A. A. Freitas, and J. Cinatl, "Two extensions to multi-label correlation-based feature selection: A case study in bioinformatics", *IEEE International Conference on Systems, Man, and Cybernetics*, 2013, pp. 1519-1524.

[35] Khan, A. U. R., M. Khan, and M. B. Khan, "Naive Multi-label classification of YouTube comments using comparative opinion mining", *Procedia Computer Science*, Vol. 82, 2016, pp. 57-64, Available at http://dx.doi.org/10.1016/j.procs.2016.04.009.

[36] Le, T. H., C. Arcodia, M. A. Novais, A. Kralj, and T. C. Phan, "Exploring the multi-dimensionality of authenticity in dining experiences using online reviews", *Tourism Management*, Vol. 85, 2021, 104292.

[37] Lee, J. and D. W. Kim, "SCLS: Multi-label feature selection based on scalable criterion for large label set", *Pattern Recognition*, Vol. 66, 2017, pp. 342-352, Available at http://dx.doi.org/10.1016/j.patcog.2017.01.014.

[38] Lee, J., I. Yu, J. Park, and D. W. Kim, "Memetic feature selection for multilabel text categorization using label frequency difference", *Information Sciences*, Vol. 485, 2019, pp. 263-280, Available at http://dx.doi.org/10.1016/j.ins.2019.02.021.

[39] Li, J., K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective", *ACM Computing Surveys*, Vol. 50, No. 6, 2017, pp. 1-45, Available at http://dx.doi.org/10.1145/3136625.

[40] Lin, S. C., C. J. Chen, and T. J. Lee, "A Multi-Label Classification With Hybrid Label-Based Meta-Learning Method in Internet of Things", *IEEE Access*, Vol. 8, 2020, pp. 42261-42269.

[41] Liu, L., D. Dzyabura, and N. Mizik, "Visual listening in: Extracting brand image portrayed on social media", *Marketing Science*, Vol. 39, No. 4, 2020, pp 669-686.

[42] Liu, S. M. and J. H. Chen, "A multi-label classification based approach for sentiment classification", *Expert Systems with Applications*, Vol. 42, No. 3, 2015, pp. 1083-1093, Available at http://dx.doi.org/10.1016/j.eswa.2014.08.036.

[43] Liu, W. and I. Tsang, "On the optimality of classifier chain for multi-label classification", *Neural Information Processing Systems*, 2015, pp. 712-720.

[44] Marcheggiani, D., O. Tackstrom, A. Esuli, and F. Sebastiani, "Hierarchical multi-label conditional random fields for aspect-oriented opinion mining", *European Conference on Information Retrieval*, 2014, pp. 273-285.

[45] Miao, J. and L. Niu, "A survey on feature selection", *Procedia Computer Science*, Vol. 91, 2016, pp. 919-926, Available at http://dx.doi.org/10.1016/j.procs.2016.07.111.

[46] Montaes, E., R. Senge, J. Barranquero, J. R. Quevedo, J. J. del Coz, and E. Hllermeier, "Dependent binary relevance models for multi-label classification", *Pattern Recognition*, Vol. 47, No. 3, 2014, pp. 1494-1508, Available at http://dx.doi.org/10.1016/j.patcog.2013.09.029.

[47] Nam, J., J. Kim, E. L. Menca, I. Gurevych, and J. Frnkranz, "Large-scale multi-label text classification?revisiting neural networks", *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 437-452, 2014.

[48] Pereira, R. B., A. Plastino, B. Zadrozny, and L. H. Merschmann, "Correlation analysis of performance measures for multi-label classification", *Information Processing & Management*, Vol. 54, No. 3, 2018, pp. 359-369, Available at http://dx.doi.org/10.1016/j.ipm.2018.01.002.

[49] Phillips-Wren, G., M. Daly, and F. Burstein, "Reconciling business intelligence, analytics and decision support systems: More data, deeper insight", *Decision Support Systems*, Vol. 146, 2021.

[50] Priyadarsini, M. J. P., K. Murugesan, S. R. Inbathini, J. Vishal, S. Anand, and R. N. Nair, "Performance Evaluation of LDA, CCA and AAM", *Research Journal of Applied Sciences, Engineering and Technology*, Vol. 9, No. 9, 2015, pp. 685-699, Available at http://dx.doi.org/10.1109/ICOEI.2018.8553811.

[51] Reyes, O., C. Morell, and S. Ventura, "Scalable extensions of the ReliefF algorithm for weighting and selecting features on the multi-label learning context", *Neurocomputing*, No. 161, 2015, pp. 168-182.

[52] Rokach, L., A. Schclar, and E. Itach, "Ensemble methods for multi-label classification", *Expert Systems with Applications*, Vol. 41, No. 16, 2014, pp. 7507-7523, Available at http://dx.doi.org/10.1016/j.eswa.2014.06.015.

[53] Schlegelmilch, B. B., K. Sharma, and S. Garg,

Employing machine learning for capturing COVID-19 consumer sentiments from six countries: A methodological illustration", *International Marketing Review*, 2022.

[54] Spolar, N., E. A. Cherman, M. C. Monard, and H. D. Lee, "A comparison of multi-label feature selection methods using the problem transformation approach", *Electronic Notes in Theoretical Computer Science*, Vol. 292, 2013, pp. 135-151, Available at http://dx.doi.org/10.1016/j.entcs.2013.02.010.

[55] Stamatescu, G., I. Fagarasan, and A. Sachenko, "Sensing and data-driven control for smart building and smart city systems", *Journal of Sensor*, 2019, Available at http://dx.doi.org/10.1155/2019/4528034.

[56] Sun, L., M. Kudo, and K. Kimura, "Multi-label classification with meta-label-specific features", *23rd International Conference on Pattern Recognition*, 2016, pp. 1612-1617, Available at http://dx.doi.org/10.1109/ICPR.2016.7899867.

[57] Tsoumakas, G. and I. Katakis, "Multi-label classification: An overview", *International Journal of Data Warehousing and Mining*, Vol. 3, No. 3, 2007, pp. 1-13.

[58] Tsoumakas, G., I. Katakis, and I. Vlahavas, *Mining Multi-label Data. In Data mining and Knowledge Discovery Handbook*, Springer, Boston, MA, 2009.

[59] Tsoumakas, G., E. Spyromitros-Xioufis, J. Vilcek, and I. Vlahavas, "Mulan: A java library for multi-label learning", *The Journal of Machine Learning Research*, Vol. 12, 2011, pp. 2411-2414.

[60] Vens, C., J. Struyf, L. Schietgat, S. Deroski, and H. Blockeel, "Decision trees for hierarchical multi-label classification", Machine Learning, Vol. 73, No. 2, 2008, pp. 185-214.

[61] Wang, J. and J. D. Zucker, "Solving multiple-instance problem: A lazy learning approach", *International Conference on Machine Learning*, pp. 1119-1125, 2000.

[62] Wang, J., Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "Cnn-rnn: A unified framework for multi-label image classification", *The IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 2285-2294.

[63] Wang, R., S. Ye, K. Li, and S. Kwong, "Bayesian Network Based Label Correlation Analysis For Multi-label Classifier Chain", 2019, arXiv preprint arXiv:1908.02172.

[64] Wehrmann, J. and R. C. Barros, "Movie genre classification: A multi-label approach based on convolutions through time", *Applied Soft Computing*, Vol. 61, 2017, pp. 973-982, Available at http://dx.doi.org/10.1016/j.asoc.2017.08.029.

[65] Wu, G., R. Zheng, Y. Tian, and D. Liu, "Joint Ranking SVM and Binary Relevance with robust Low-rank learning for multi-label classification", *Neural Networks*, Vol. 122, 2020, pp. 24-39, Available at http://dx.doi.org/10.1016/j.neunet.2019.10.002.

[66] Xu, S., X. Yang, H. Yu, D. J. Yu, J. Yang, and E. C. Tsang, "Multi-label learning with label-specific feature reduction", *Knowledge-Based Systems*, Vol. 104, 2016, pp. 52-61, Available at http://dx.doi.org/10.1016/j.knosys.2016.04.012.

[67] Yassine, A., S. Singh, M. S. Hossain, and G. Muhammad, "IoT big data analytics for smart homes with fog and cloud computing", *Future Generation Computer Systems*, Vol. 91, 2019, pp. 563-573, Available at http://dx.doi.org/10.1016/j.future.2018.08.040.

[68] Zhang, M. L. and Z. H. Zhou, "A k-nearest neighbor based algorithm for multi-label classification", *IEEE International Conference on*

Granular Computing, 2005, pp. 718-721, Available at http://dx.doi.org/10.1109/GRC.2005.1547385.

[69] Zhang, M. L. and Z. H. Zhou, "A review on multi-label learning algorithms", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, No. 8, 2013, pp. 1819-1837, Available at http://dx.doi.org/10.1109/TKDE.2013.39.

[70] Zhang, M. L. and Z. H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning", *Pattern Recognition*, Vol. 40, No. 7, 2007, pp. 2038-2048, Available at http://dx.doi.org/10.1016/j.patcog.2006.12.019.

[71] Zhang, M. L., Y. K. Li, X. Y. Liu, and X. Geng, "Binary relevance for multi-label learning: An overview", *Frontiers of Computer Science*, Vol. 12, No. 2, 2018, pp. 191-202, Available at http://dx.doi.org/10.1007/s11704-017-7031-7.

Information Systems Review
Volume 25 Number 1
February 2023

Exploring the Performance of Multi-Label Feature Selection for Effective Decision-Making: Focusing on Sentiment Analysis

Jong Yoon Won* · Kun Chang Lee**

Abstract

Management decision-making based on artificial intelligence(AI) plays an important role in helping decision-makers. Business decision-making centered on AI is evaluated as a driving force for corporate growth. AI-based on accurate analysis techniques could support decision-makers in making high-quality decisions. This study proposes an effective decision-making method with the application of multi-label feature selection. In this regard, We present a CFS-BR (Correlation-based Feature Selection based on Binary Relevance approach) that reduces data sets in high-dimensional space. As a result of analyzing sample data and empirical data, CFS-BR can support efficient decision-making by selecting the best combination of meaningful attributes based on the Best-First algorithm. In addition, compared to the previous multi-label feature selection method, CFS-BR is useful for increasing the effectiveness of decision-making, as its accuracy is higher.

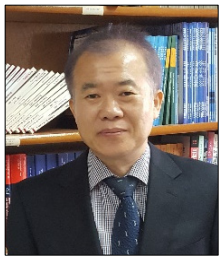
Keywords: Multi-Label, Feature Selection, Sentiment Analysis, Data-mining, Decision-making

* Doctoral Student, SKK Business School, Sungkyunkwan University
** Corresponding Author, Professor, SKK Business School, Sungkyunkwan University

◐ 저 자 소 개 ◑

Jong Yoon Won (yoonjbest1@gmail.com)

He is now pursuing PhD degree at the SKK Business School in Sungkyunkwan University (Seoul, South Korea). He is actively working on neuroscience based decision-making mechanism, business problem solving creativity, Big-Data analytics, and AI-driven analysis of business problem-solving processes, etc.



Kun Chang Lee (kunchanglee@gmail.com)

He is a full professor in Sungkyunkwan University, Seoul, South Korea. He is also affiliated with SAIHST (Samsung Advanced Institute for Health Sciences & Technology). He has published papers at prestigious journals like Decision Support Systems, Journal of MIS, IEEE Transactions on Engineering Management, to name a few.

논문접수일 : 2022년 04월 12일
게재확정일 : 2022년 12월 28일
1차 수정일 : 2022년 06월 27일
2차 수정일 : 2022년 08월 13일
3차 수정일 : 2022년 12월 26일