JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 25 No. 11, pp. 41-50, November 2020
https://doi.org/10.9708/jksci.2020.25.11.041

Exploring the Feature Selection Method for Effective Opinion Mining: Emphasis on Particle Swarm Optimization Algorithms

Kyun Sun Eo*, Kun Chang Lee**

*PhD Student, SKK Business School, Sungkyunkwan University, Seoul, Korea
**Professor, SKK Business School/SAIHST (Samsung Advanced Institute of Health Sciences & Technology), Sungkyunkwan University, Seoul, Korea

[Abstract]

Sentimental analysis begins with the search for words that determine the sentimentality inherent in data. Managers can understand market sentimentality by analyzing a number of relevant sentiment words which consumers usually tend to use. In this study, we propose exploring performance of feature selection methods embedded with Particle Swarm Optimization Multi Objectives Evolutionary Algorithms. The performance of the feature selection methods was benchmarked with machine learning classifiers such as Decision Tree, Naive Bayesian Network, Support Vector Machine, Random Forest, Bagging, Random Subspace, and Rotation Forest. Our empirical results of opinion mining revealed that the number of features was significantly reduced and the performance was not hurt. In specific, the Support Vector Machine showed the highest accuracy. Random subspace produced the best AUC results.

▸ **Key words**: Sentiment analysis, Feature selection, Particle Swarm Optimization, Multi Objective Evolutionary Algorithm, Machine learning

[요 약]

감성분석 연구에서는 문장에 내포된 감성을 결정짓는 단어를 찾는 것으로부터 시작된다. 경영자는 소비자가 주로 사용하는 단어를 분석함으로써 시장의 반응을 이해할 수 있다. 본 연구에서는 감성분류의 성능에 영향을 미치는 단어를 찾기 위하여 입자군집최적화 탐색방법과 다목적진화 알고리즘이 적용된 속성선택 방법을 제안한다. 속성선택 방법은 기존 머신러닝 분류기를 벤치마킹함으로써 성능이 비교된다. 벤치마킹된 분류기는 의사결정나무, 나이브 베이지안 네트워크, 서포터 벡터 머신, 랜덤포레스트, 배깅, 랜덤 서브스페이스, 로테이션 포레스트이다. 연구결과에 따르면, 입자군집최적화 알고리즘이 적용된 속성선택방법으로 선택된 속성을 사용한 경우에 속성의 수를 상당히 줄일 수 있었고, 분류기의 성능을 유지시킬 수 있었다. 특히, 정확도 결과에서는 입자군집 최적화 탐색방법으로 선택된 속성을 사용한 경우의 서포터 벡터 머신의 성능이 가장 높게 나타났다. AUC 결과에서는 랜덤 서브스페이스가 가장 높게 나타났다. 본 연구의 결과는 해당 탐색방법과 분류기를 적용함으로써 오피니언 마이닝 모델의 성능을 효율적으로 유지 및 개선시키도록 도움을 준다.

▸ **주제어**: 감성분석, 속성선택, 입자군집최적화, 다목적진화알고리즘, 머신러닝

• First Author: Kyun Sun Eo, Corresponding Author: Kun Chang Lee
*Kyun Sun Eo (eokyunsun@gmail.com), SKK Business School, Sungkyunkwan University
**Kun Chang Lee (kunchanglee@gmail.com), SKK Business School/SAIHST (Samsung Advanced Institute of Health Sciences & Technology), Sungkyunkwan University
• Received: 2020. 10. 13, Revised: 2020. 11. 09, Accepted: 2020. 11. 10.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

웹 2.0의 성장으로 인해 소셜 미디어에는 사용자가 게시하는 게시물들이 대량으로 업로드되고 있다[1]. 이 게시물들은 사람들의 관심, 생각, 감성, 행동들을 담고 있다. 시간이 흐름에 따라 오랜기간 동안 소셜 미디어에 축적되었다.

오피니언 마이닝(Opinion Mining)은 상품 및 서비스에 대한 소비자의 리뷰를 긍정 또는 부정적인지를 머신러닝 방법으로 분류하는 감성분석 방법이다[2]. 출시된 상품 및 서비스에 대한 소비자의 의견을 파악하는 것은 기업 운영자에게 소중한 자원이 된다. 소비자의 반응이 긍정적이면 이후에 연관된 상품을 출시할 수 있고 부정적이면 의견을 바탕으로 개선 방안을 강구할 수 있다[3].

소비자의 의견은 소비자가 중요하게 생각하는 단어를 토대로 유추할 수 있다. 이는 도메인마다 천차만별이다. 영화 리뷰에는 영화의 장르 및 내용과 관련된 단어들이 포함되고, 전자제품의 경우에는 디자인 또는 성능을 설명하는 단어가 사용된다. 따라서 오피니언 마이닝 대상에 따라 유의미하게 사용되는 단어를 추출하는 과정은 오피니언 마이닝 모델 구축과정에서 필수적이다.

일반적으로 감성분석 연구에서는 텍스트 데이터를 처리하는 백 오브 워즈(Bag of Words) 방법이 사용된다[4]. 백 오브 워즈 방법은 텍스트를 단어의 조합으로 속성 셋이 구성된다. 백 오브 워즈 방법으로 구성된 속성벡터 셋은 단어가 많으면 많을수록 고차원의 속성 셋이 생성되는데, 이를 위해 불필요한 속성의 수를 감소시킬 수 있는 속성선택(Feature Selection) 방법이 요구된다.

본 연구에서는 입자군집 최적화(Particle Swarm Optimization) 방법과 다목적진화 알고리즘(Multi Objective Evolutionary Algorithm) 방법이 적용된 속성선택 방법을 이용하여 불필요한 속성을 줄이고 오피니언 마이닝 모델의 성능을 개선시키고자 한다. 속성선택방법은 머신러닝 분류기에 사용되는 속성의 수를 줄여 학습효율을 높이고 분류기의 성능을 향상 및 유지시키기 위해 사용된다.

본 연구에서는 속성선택 방법으로 선택된 속성으로 구성된 모델의 성능을 객관적으로 검증하기 위해 다수의 머신러닝 분류기를 벤치마킹하였다. 벤치마킹한 분류기는 의사결정나무(Decision Tree), 나이브 베이지안 네트워크(Naive Bayesian Network), 서포터 벡터머신(Support Vector Machine), 랜덤포레스트(Random Forest), 배깅(Bagging), 랜덤서브스페이스(Random Subspace), 마지막으로 로테이션 포레스트이다(Rotation Forest)[5-10]. 이에 따라, 본 연구에서 제시하는 연구질문(Research Question)은 다음과 같다.

RQ1: 오피니언 마이닝 모델에 속성선택 방법을 적용할 경우 모델의 성능은 향상 및 유지되는가?

RQ2: 속성선택 방법으로 선택된 속성을 사용할 경우, 가장 높은 성능을 나타내는 분류기는 무엇인가?

본 연구의 구성은 다음과 같다. 2장에서는 오피니언 마이닝에 대한 선행연구와 입자군집 최적화 알고리즘과 다목적 진화 알고리즘이 적용된 속성선택 방법 연구를 설명하고, 3장에서는 본 연구에서 제안되는 연구방법, 4장에서는 연구 결과 및 토의, 마지막으로 5장에서는 결론을 다루고 한계점과 향후 연구에 대해 논의될 것이다.

II. Preliminaries

1. A study on the sentiment analysis based on machine learning

머신러닝은 감성분석 연구에서 중요한 역할을 갖는다. 머신러닝 알고리즘은 긍정적이거나 부정적, 혹은 중립적인 의견으로 구성된 분류문제에 사용된다. 기존에 학습된 알고리즘은 새로 추가된 데이터를 분류에 사용된다. 서포터 벡터머신과 나이브 베이즈와 같은 머신러닝 분류방법은 과거 많은 연구에서 다양한 데이터 셋을 벤치마킹하여 높은 정확도 성과를 나타내었다. 감성분석에 관한 선행연구는 다음 Table 1에 요약되었다.

Ye et al. (2009)의 연구에서는 나이브 베이즈와 서포터 벡터 머신으로 N-그램을 기반으로 구축된 모델이 사용되었다. 미국과 유럽의 유명지를 소개하는 여행블로그의 리뷰에 대한 감성분류 성과를 비교하였다[11]. 서포터 벡터머신과 N-그램 모델은 나이브 베이즈 모델보다 우수했지만 많은 양의 리뷰를 모델에 학습시킬 때에는 세 모델 모두 80%이상의 정확도에 도달하였다.

감성분석 분류를 해결하기 위한 다른 방법으로는 인공신경망 기반 모델이 제안되었다. 인공신경망은 사람 뇌를 구성하는 뉴런의 작동원리를 모방하여 만든 알고리즘이다. 인공신경망은 입력 값을 비선형 변환을 통해 처리되고, 결과 값은 네트워크의 다음 층에 연결된 신경망에 전달된다. Moraes et al. (2013)에서는 문서 단위의 감성분석을 위하여 서포터 벡터머신과 인공신경망이 비교되었다. 여기서 감성점수는 이진 변수로 분류모델에 사용되었다[12].

감성분석은 다양한 분야에 응용되지만, 그 중에서도 소셜 미디어에 대한 분석이 매우 활발하다. 소셜 미디어에는 전 세계 사람들에 의해 게시글이 활발하게 업로드되고 있다. 특히, 트위터는 지인 혹은 친구에게 짧은 문장으로 의

Table 1. A Study on the Sentiment Analysis Based on Machine Learning

author	year	Data	study method	BOW	FS
Ye et al.	2009	Travel blog	SVM, Naive Bayes	O	Information Gain
Moraes et al.	2013	Movie, Amazon	SVM, Naive Bayes, ANN	O	Expert Knowledge Minimum Fequency Information Gain Chi-Square
Ghiassi et al.	2013	Twitter	DAN2, SVM	O	Term Frequency
Da Silva et al.	2014	Twitter	BOW, FH	O	-
Wang et al.	2014	Amazon review	Bagging, Boosting, Random Subspace	O	-
Yoo et al.	2018	twitter	Prediction framework for users' sentimental trajectories for events detect in real time	O	-
Garcia-Pablos et al.	2018	SemEval2016	W2VLDA(Word2vec with LDA)	-	-
This study	**2020**	**SemEval2017**	**Comparison performance of FS method (PSO, MOEA)**	**O**	**PSO, MOEA**

사소통하는 도구로써 자리 잡았다[13,14]. 트위터는 전 세계의 5억명이 넘은 사람들이 사용하는 마이크로 블로그 서비스이다. 트위터에 게시되는 짧은 트윗들은 서포터 벡터 머신과 인공신경망 모델로 작성자의 의견을 분류하기 위하여 학습대상으로 사용되었다. 트윗에 포함된 반복적이고 무의미한 단어들은 단어의 빈도수에 따라 선택되어 제거 대상으로 삼는다[15].

속성 재표현 방법에는 백 오브 워즈 방법과 피쳐해싱 방법이 있다. 정확도 측면에서는 백 오브 워즈 방법이 피쳐해싱 방법보다 높은 성과를 나타내는 것으로 확인되었다[4].

앙상블 분류기는 오피니언 마이닝에서 단일 분류기보다 높은 성능을 나타내는 것으로 제안된다. 배깅 랜덤서브 스페이스, 부스팅 모델이 적용된 앙상블 분류기는 단일 분류기보다 감성분석에서 우수한 성능을 나타냈다[16].

Yoo et al., (2018) 의 연구에서는 소셜미디어의 콘텐츠를 실시간으로 관리 가능한 감성 예측 시스템이 제안되었다. 실시간 감성분석 시스템은 사용자의 감성변화에 따른 경로를 분석하고 예측할 수 있다. 이 시스템은 특정 이벤트를 발견하고 실시간으로 관리가 가능하다[17].

Garcia Pablos et al. (2018)에서는 코사인 유사도를 분석하는 Word2Vec 방법과 문서의 키워드를 추출하는 LDA(Latent Dirichlet Allocation) 토픽모델링 방법이 결합된 방법을 제안하였다[18]. 이는 W2VLAD로 명칭되어 사전학습이 필요 없고, 특정 도메인에 대한 소스 없이도 감성을 분류하는 시스템이다.

2. A study on the feature selection with evolutionary algorithms

유전알고리즘의 진화 연산은 자연 진화방식을 이용한다. 자연 진화방법에 따르면, 해결하고자 하는 문제는 개인이 사는 환경의 역할을 하며, 개인은 문제에 대한 가능한 해결방안을 나타낸다. 개인이 환경에 적응하는 정도는 피트니스 함수에 의해 그 적정성이 측정된다. 자연에서의 진화처럼 진화 알고리즘은 해당문제에 대한 점진적인 해결방안을 만들어 나간다. 유전 알고리즘은 초기에는 무작위 해결방안의 모집단에서 시작되어 각 반복마다 교차 및 돌연변이와 같은 변이 연산자를 사용하여 최고의 개인을 선정하고 결합하여 다음 세대가 구성된다. 이 과정은 정지 기준이 충족될 때까지 반복된다. 반면, 두 가지 이상의 문제 목적 사이에 본질적인 충동이 있을 경우의 문제에는 다목적 알고리즘이 필요하다.

2.1 Particle Swarm Optimization(PSO)

입자군집 최적화는 사회적 행동의 시뮬레이션으로부터 영감을 받은 최적화 기법이다. 입자군집 최적화의 원래 개념은 하늘을 나는 새들의 군집 움직임과 이에 필요한 정보교환 수단을 바탕으로 시뮬레이션하여 문제를 해결하는 것이다[19,20]. 입자군집 최적화는 전통적으로 연속 값에 대한 공간 검색에 적용되어 왔다. 입자군집 최적화는 여러 입자가 해결책(Solution) 공간의 다른 부분을 탐색할 수 있기 때문에 최적의 해결책이 발견될 때까지 진행된다. 입자군집최적화 방법이 사용된 속성선택 연구는 다음 Table 2과 같다.

Xue et al. (2012)에서는 두 가지 입자 군집 최적화 방법이 제안되었다[21]. 첫 번째는 비지배적정렬 방법이 적용된 입자군집 최적화 (Non-Dominance sorted PSOFS, NSPSOFS) 방법이고, 두 번째는 파레토 우선 해 기반 입자군집 최적화 방법에 군중심리, 돌연변이, 지배의 개념이 적용된 CMDPSOFS (PSOFS-Crowding, Mutation, and Dominance) 방법이다. 연구결과에 따르면, 12가지 벤치마킹 데이터 셋을 통해 K-최근접이웃(K-Nearest

Table 2. A Study on the Feature Selection of the Particle Swarm Optimization Algorithm

Author	year	Data	Model learners	FS search algorithm
Xue et al.	2012	Wine, Australian, Zoo, Vehicle, German, WBCD, Ionosphere, Lung Cancer, Hillvalley, Musk1, Madelon, Isolet5	KNN	NSPSOFS, CMDPSOFS, PSO-MO
Cervante et al.	2012	Chess, Splice, Spect, Lymph	DT	BPSOFS
Krisshna et al.	2014	Cambridge ORL, UMIST, Extended YaleB, CMUPIE, Color FERET, FEI, HP	Euclidean	ThBPSO
Kushwaha, & Pant	2018	Reuter-21578, TDT2, TR11	k-means clustering	LBPSO
Amoozegar & Minaei-Bidgoli	2018	Vehicle, German, Sonar, Hillvalley, Musk1, LSVT, Madelon, Isolet5, CNAE	kNN	RFPSOFS, CMDPSOFS, HMPSOFS, MOPSO, NSGAII
This study	**2020**	**SemEval2017**	**DT, NBN, SVM, RF, BA, RS, Rof**	**PSO, MOEA**

Neighbor)의 성능으로 비교한 결과, 두 번째 방법이 다른 방법보다 더 나은 결과를 나타냈다.

Cervante et al. (2012)에서는 이진 입자군집 최적화 방법을 바탕으로 선택된 속성을 사용하여 의사결정나무에 학습시켰다[22]. 이항 입자군집 최적화를 통해 적절한 가중치를 사용하여 속성의 수를 상당히 줄일 수 있고, 더 높은 분류 정확도를 달성할 수 있다.

Krisshna et al., (2014)의 연구에서는 얼굴 인식 시스템의 성능 향상을 위하여 듀얼 서브밴드 주파수 도메인 속성추출 방법과 임계값 기반 이진 입자군집 최적화 방법(ThBPSO)이 제안되었다[23]. 결과에 따르면, 제안된 방법은 얼굴 인식률을 유의하게 증가시킬 수 있었고, 얼굴인식에 필요한 속성의 수를 상당히 감소시킬 수 있었다.

Kushwaha, & Pant, (2018)의 연구에서는 비지도 텍스트 클러스터링 속성선택(Link Based Particle Swarm Optimization, LBPSO) 방법이 제안되었다[24]. 이 방법은 중요한 속성을 선택하기 위하여 새로운 근접선택 전략이 사용된다. 비지도 텍스트 클러스터링 속성선택방법으로 Reuter 21578, TDT2 and Tr11의 세가지 데이터셋을 분석한 결과에 따르면 비지도 텍스트 클러스터링 방법은 우수한 결과를 나타냈다.

2.2 Multi Objective Evolutionary Algorithm(MOEA)

다목적진화 알고리즘은 유전알고리즘의 일종으로 파레토 최적해 생성에 따른 탐색방법이다. 유전알고리즘은 자연 진화 과정으로부터 영감을 받은 메타 휴리스틱 최적화 방법으로 공간 탐색 방법에 해당된다[25]. 다목적진화 알고리즘을 이용한 속성선택 방법은 두 개 이상의 목적에 따른 최적의 속성을 선택하는 것에 사용된다. 이를 바탕으로 속성선택 방법은 분류기에 학습되는 속성의 수를 최소화하고 분류기의 정확도를 향상시키기 위해 사용된다. 다목적진화 알고리즘이 사용된 속성선택방법 연구는 다음 Table 3과 같다.

Gaspar-Cunha (2010)에서는 단일 광양자 방사 컴퓨터 단층촬영(Single Proton Emission Computed Tomography)을 통한 심장 진단 문제를 다목적진화 알고리즘과 서포터벡터 머신으로 입증되었다[26]. Vignolo et al. (2013)의 연구에서는 얼굴인식에 가장 관련 있는 속성

Table 3. A Study on the Feature Selection of Multi-Objective Evolutionary Algorithm

Author	year	Data	Model learners	FS search algorithm
Gaspar-Cunha	2010	Cardiac SPECT Diagnosis	SVM	MOEA
Vignolo et al.	2013	Essex Face Database	KNN	MOEA-wraper, Aggregative GA, MOGA
Mlakar et al.	2017	CK, MMI, JAFFE	SVM	DEMOFS
Jimenez et al.	2017	Online Sales data set	RF, M5Rules, PCR, Relaxo, Foba, LeapForward, Penalized Ppr, Ridge	ENORA, NSGA-II
Zhang et al.	2020	20 UCI repository data	OPS	MOFS-BDE, NSGAFS, DEMOFS, MOPSOFS
This study		**SemEval2017**	**DT, NBN, SVM, RF, BA, RS, Rof**	**MOEA, PSO**

셋을 선택하기 위해 유전 알고리즘 기반 다목적 래퍼 방법이 제안되었다[27]. 속성의 수를 최소화하기 위해 사용된 다목적 래퍼 알고리즘은 동시에 차별화를 최대화 할 수 있는 공간을 탐색한다. 얼굴인식 분야에서는 다목적 차분진화(Differential Evolution Multi Objective, DEMO) 알고리즘이 적용된 속성선택 방법을 이용하여 얼굴표정 분류 방법이 제안되었다[28].

ENORA(Evolutionary NOn-Dominated Radial Slots based Algorithm)에 랜덤포레스트로 구동되는 래퍼 검색 전략이 적용된 속성선택 방법은 정확한 판매예측 회귀모델을 구성하기 위한 속성을 선택하는데 사용되었다. ENORA는 NSGA-II(Non-dominated Sorted Genetic Algorithm), 후방탐색방법, 순환 속성제거(Recursive Feature Elimination) 방법, 속성선택 전 데이터 셋과 비교되어 평균적으로 높은 성능을 나타냈다[29].

최근 연구에서는 셀프 학습을 통한 바이너리 차등 진화 알고리즘(Multi Objective Feature Selection-The Binary Differential Evolution with Self-learning, MOFS-BDE)이라고 하는 새로운 다목적 속성선택 방법이 제안되었다[30].

다목적알고리즘의 메카니즘은 분류기의 정확도를 극대화하고 속성 수를 최소화해야하는 속성선택의 특성과 부합된다. 다목적 진화 알고리즘은 다목적 최적해결책을 병렬로 탐색하고 최종 모집단에서 한 번의 실행으로 최적의 해결책 집합을 찾을 수 있기 때문에 다목적 최적화에 적합하다. 따라서 다목적 탐색 알고리즘의 목적은 파레토 우선(Pareto front)에 대한 좋은 근사치인 해결책 집단을 발견하는 것이다. 다목적 속성 선택의 경우에는 해결책으로 정확도와 모델 복잡성 사이에 관련된 속성의 부분집합을 나타낼 수 있다.

진화알고리즘이 사용된 선행연구를 살펴본 결과, 이미지 인식, 오픈 데이터 셋 등 다양한 분야에서 진화알고리즘이 적용되었다. 그러나 텍스트 감성분석 분야에서는 진화 알고리즘을 이용한 속성선택 연구가 미흡하다.

본 연구에서는 트위터 텍스트 데이터를 바탕으로 구성된 SemEval2017 오픈 데이터가 사용되었다. 나아가 진화알고리즘을 기반으로 만들어진 입자군집 최적화 알고리즘과 다목적 진화 알고리즘이 적용된 속성선택 방법이 사용된다.

III. The Proposed Scheme

본 연구에서 제안된 연구방법의 순서는 다음 Fig 1과 같다.

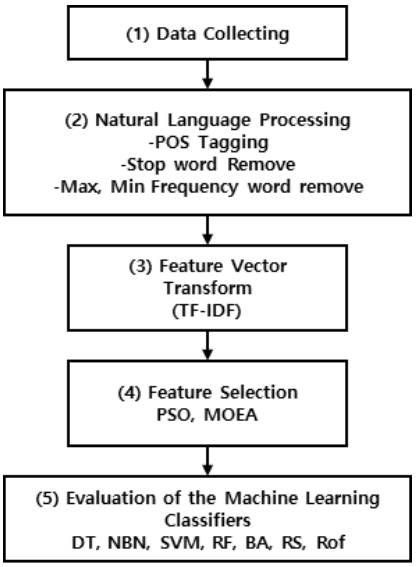


Fig. 1. Procedure of this study

(1) 데이터 수집

본 연구에서는 효과적인 오피니언 마이닝 분류를 위하여 입자군집 최적화 방법과 다목적진화 알고리즘이 적용된 속성선택 방법을 제안한다. 본 연구를 위하여 SemEval2017 오픈 데이터를 이용하였다[31]. SemEval2017 데이터는 SemEval2016의 트위터 데이터를 추출한 트윗들로 구성되어 있다. 총 2013, 2014, 2015, 2016년도 기간동안 수집된 트위터의 짧은 텍스트 문장과 문장의 긍정, 부정, 중립 등과 같은 감성 값으로 이루어져 있다.

(2) 자연어 처리 단계

자연어 처리단계에서는 첫째, 문장을 단어 별로 분리하고, 단어에 대한 형태소를 태깅한다. 문장 속의 단어는 명사, 형용사, 동사, 부사 등으로 구성되어 있어 각각의 단어마다 형태소 태깅이 필요하다. 둘째, 문장 속에 있는 단어 중에서 불필요하게 중복된 단어를 제거한다. 마지막으로 셋째, 너무 빈번하게 등장하는 단어나 극히 드물게 나타나는 단어는 제거한다.

(3) 속성 벡터 변화 단계

자연어처리 과정를 거친 데이터는 TF-IDF(Term Frequency-Inverse Document Frequency) 방법을 이용해 가중치가 산출된다[32]. TF-IDF는 TF(단어의 빈도), IDF(역 단어가 등장하는 문서의 수)의 곱으로 단어의 중요도를 나타낸다. 산출된 데이터 셋은 재표현 방법을 통해 속성벡터 셋으로 변환된다.

(4) 속성 선택 단계

본 연구는 상관관계 속성선택 방법이 이용되고 속성을 탐색하기 위한 탐색방법으로 다목적진화 알고리즘방법과 입자군집 최적화 방법이 사용되었다. 속성선택 방법을 이용하여 선택된 속성 개수는 다음 Table 4와 같다.

속성선택 전에는 2013은 4,254개, 2014는 4,804개, 2015는 3,480개, 2016은 5,128개이다. 다목적진화 알고리즘 적용 후에는 2013, 2014, 2015, 2016년도가 각각 16개, 55개, 30개, 8개로 줄었으며, 입자군집최적화 방법이 적용된 후에는 1,633개, 1,315개, 983개, 2,123개로 줄었다.

(5) 머신러닝 분류기 학습 & 검증

본 연구에서는 속성선택방법의 효과성을 검증하기 위해 다음과 같이 다수의 머신러닝 분류기가 사용되었다.

본 연구에서 벤치마킹된 분류기는 의사결정트리(DT), 나이브 베이지안 네트워크(NBN), 서포터 벡터 머신(SVM), 랜덤포레스트(RF), 배깅(BA), 랜덤 서브스페이트(RS), 로테이션 포레스트(Rof)이다[25-30].

분류기 모델의 편향된 훈련 및 테스트를 피하기 위해 10-겹 교차검증(10-Fold Cross Validation)이 사용되었다[33]. 10-겹 교차검증은 전체 데이터 셋에서 균등하게 10분할한 다음, 분할된 9할의 데이터를 훈령용으로 사용되고 남은 1할 데이터를 검증용에 사용된다. 이를 9번 더 순차적으로 실시하여 모든 데이터를 평가하는 방법이다.

본 연구에서는 정확도(Accuracy), AUC(Area Under the ROC Curve)을 이용하여 모델의 성능을 평가하였다. 성능평가 지표에서 정확도는 전체 분류 건수에서 긍정감성인지 부정감성인지 맞게 분류한 비율이다. 정확도의 수식은 다음과 같다. TP(True Positive)는 긍정을 긍정으로, TN(True Negative)는 부정을 부정으로 맞게 분류한 것이고, FP(False Positive)는 부정을 긍정으로, FN(False Negative)는 긍정을 부정으로 잘못 분류한 것이다.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$True\ Positive\ Rate = \frac{TP}{TP+FN} \quad (2)$$

$$False\ Positive\ Rate = \frac{FP}{FP+TN} \quad (3)$$

AUC는 신호탐지 이론에서 주로 사용되고, y축은 적중확률(Sensitivity, True Positive Rate), x 축은 오경보확률(1-Specificity, False Positive Rate)을 나타내는 그래프이다. AUC는 0.5부터 1까지의 범위로 예측확률을 나타내고 1에 근접할수록 높은 예측확률을 갖는다.

IV. Results and Discussion

본 연구의 분석 결과는 다음 Table 5, 6과 같다. 정확도의 결과는 Table 5, AUC의 결과는 Table 6에 나타낸다. 표에 나타나는 MOEA는 다목적 진화알고리즘을 이용한 속성선택방법이고, PSO는 입자 군집 최적화이다. DT는 의사결정나무, NBN은 나이브 베이지안 네트워크, SVM은 서포터 벡터머신, RF는 랜덤포레스트, BA는 배깅, RS는 랜덤서브스페이스이다. Rof는 로테이션 포레스트를 의미한다. 본 연구의 연구질문에 대한 결과는 다음과 같다.

RQ1: 오피니언 마이닝 모델에 속성선택 방법을 적용할 경우 모델의 성능은 향상 및 유지되는가?

속성선택을 사용하지 않은 경우의 정확도 결과는 서포터 벡터머신(SVM)이 가장 높은 성과를 나타냈다. AUC의 결과는 twitter-2013, 15에서 랜덤포레스트(RF)가 가장 높았고, twitter-14, 16에서 랜덤서브스페이스(RS)의 성능이 가장 높았다.

입자군집 최적화 방법을 이용하여 선택된 속성을 사용한 경우의 정확도 결과에서는 twitter-2013, 2014, 2015의 분석에서는 성능이 유지되었지만 2016에서는 하락하였다. 이는 10겹-교차검증의 평균값이다. 결과에 대한 t-검정결과(p^*<0.05)는 다음 Table 7과 같다. 속성선택 전의 결과와 속성선택 후의 결과 그룹의 비교에서는 2013(0.6968), 2014(0.7001), 2015(0.4428), 2016(0.0159*), 전년도 2013-16(0.6003)이다. 2016년도에서만 통계적으로 하락했으나 전체 년도 데이터에서는 속성선택 전과 비교하여 속성선택 후의 결과는 통계적으로 유의미하지 않았다.

Table 4. Number of Variables

	Number of Features			Number of Class			
	Before	MOEA	PSO	Positive	Neutral	Negative	Total
twitter-2013	4,254	16	1,633	1,475	1,513	559	3,547
twitter-2014	4,804	55	1,315	982	669	202	1,853
twitter-2015	3,480	30	983	365	1038	987	2,390
twitter-2016	5,128	8	2,123	7,059	10,342	3,231	20,632

Table 5. Accuracy Results

	ACC	DT	NBN	SVM	RF	BA	RS	Rof
Before	twitter-2013	53.17	51.90	**59.88**	58.73	57.57	58.19	55.54
	twitter-2014	54.02	54.67	**58.40**	54.72	57.69	55.91	55.86
	twitter-2015	51.05	53.10	**58.54**	57.99	56.23	56.19	54.10
	twitter-2016	60.46	55.06	**62.28**	60.91	60.95	60.71	61.81
MOEA	twitter-2013	49.68	49.68	49.99	**50.27**	50.07	49.82	49.68
	twitter-2014	55.32	55.37	55.32	55.10	**55.59**	53.43	55.48
	twitter-2015	43.89	45.73	46.19	**46.44**	45.86	45.40	45.77
	twitter-2016	51.94	51.94	51.91	**51.95**	51.93	51.94	51.92
PSO	twitter-2013	52.38	50.63	**60.33**	57.37	56.07	57.12	53.82
	twitter-2014	57.10	56.34	**58.88**	55.42	57.86	56.02	56.83
	twitter-2015	49.58	50.17	**57.32**	54.98	52.89	54.14	51.67
	twitter-2016	59.40	54.46	**61.19**	57.43	59.59	59.07	59.82

Table 6. AUC Results

	AUC	DT	NBN	SVM	RF	BA	RS	Rof
Before	twitter-2013	0.65	0.67	0.68	**0.74**	0.71	0.72	0.69
	twitter-2014	0.60	0.64	0.60	0.64	0.66	**0.68**	0.63
	twitter-2015	0.64	0.71	0.65	**0.75**	0.72	0.73	0.68
	twitter-2016	0.67	0.65	0.65	0.69	0.70	**0.71**	0.70
MOEA	twitter-2013	0.56	0.56	**0.57**	**0.57**	**0.57**	**0.57**	0.56
	twitter-2014	0.57	0.57	0.55	0.57	**0.58**	0.57	0.58
	twitter-2015	0.51	**0.54**	0.52	**0.54**	0.53	0.53	**0.54**
	twitter-2016	**0.52**	**0.52**	**0.52**	**0.52**	**0.52**	**0.52**	**0.52**
PSO	twitter-2013	0.65	0.67	0.69	0.71	0.70	**0.72**	0.67
	twitter-2014	0.61	0.64	0.60	0.64	**0.66**	**0.66**	0.63
	twitter-2015	0.62	0.70	0.64	**0.72**	0.71	**0.72**	0.66
	twitter-2016	0.65	0.63	0.64	0.65	**0.68**	**0.68**	**0.68**

Table 7. T-Test Results

Before-PSO	ACC
twitter-2013	0.6968
twitter-2014	0.7001
twitter-2015	0.4428
twitter-2016	0.0159*
Full Year	0.6003

RQ2: FS방법에서 선택된 속성을 사용할 경우, 최적의 분류기는 무엇인가?

정확도를 살펴보면 입자군집 최적화 방법을 이용하여 선택된 속성을 사용한 경우에 서포터 벡터 머신(SVM)으로 분석한 결과가 2013년도에 60.33, 2014년도에 58.88, 2015년도에 57.32, 2016년도에 62.28로 가장 높게 나타났다.

AUC에서는 2013년도에는 랜덤 서브스페이스(RS)가 가장 높게 나타났고, 2014년도에는 배깅(BA)과 랜덤 서브스페이스(RS), 2015년도에는 랜덤 포레스트(RF)와 랜덤 서브스페이스(RS), 마지막으로 2016년도에는 배깅(BA), 랜덤 서브스페이스(RS), 로테이션 포레스트(Rof가 가장 높게 나타났다. 공통적으로 가장 높게 나타나는 분류기는 랜덤 서브스페이스(RS)이다. Fig 2는 2015년도에 PSO방법이 적용된 속성을 랜덤 서브스페이스(RS)로 분석한 AUC의 결과(0.7164) 이미지이다.

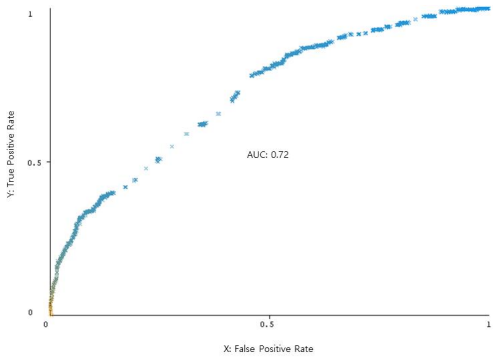


Fig. 2. AUC Image

V. Conclusions

본 연구에서는 오피니언 마이닝을 위한 다목적진화 알고리즘과 입자군집최적화 알고리즘이 적용된 속성선택 방법이 사용되었다. 본 연구에서 사용된 속성선택 방법의 객관적인 성능평가를 위해 기존의 머신러닝 분류기가 벤치마킹 되었다. 벤치마킹 된 분류기는 의사결정나무, 나이브 베이지안 네트워크, 서포터 벡터머신, 랜덤포레스트, 배깅, 랜덤서브스페이스이스, 마지막으로 로테이션 포레스트이다.

연구결과에 따르면, 입자군집최적화 알고리즘이 사용된 속성선택방법으로 선택된 속성을 사용한 경우에 속성의 수를 대폭 줄일 수 있었고 분류기의 성능을 유지할 수 있었다. 정확도의 결과에서는 서포터 벡터 머신의 성능이 가장 높게 나타났고, AUC의 결과에서는 랜덤 서브스페이스가 공통적으로 가장 높게 나타났으나, 배깅, 랜덤포레스트 또한 가장 높게 나타났다. AUC는 부정에서 부정을 맞춘 비율에 비해 긍정에서 긍정을 맞춘 비율이 클수록 1에 가까운 값이 나온다. 따라서, 해당 트리계열의 앙상블 분류기는 부정감성 예측보다 긍정감성 예측에서 높은 결과를 나타내는 것으로 사료 된다.

이와 달리, 다목적진화 알고리즘이 사용된 속성선택방법으로 선택된 속성을 사용한 경우에는 속성선택 전과 비교하여 속성의 수를 두 자리 개수 또는 한 자리 개수로 줄일 수 있었지만, 그만큼 분류기의 성능은 현저히 하락하였다.

본 연구의 결과에 따른 학술적 의의는 다음과 같다.

속성선택 방법을 이용하여 속성 수를 줄여 분류기의 분석 리소스 및 시간을 줄일 수 있다. 기존의 속성 수는 2013년이 4,254개, 2014년이 4,804개, 2015년이 3,480개. 2016이 5,128이었다. 입자군집 최적화 알고리즘이 적용된 속성선택을 적용하여 분류기의 성능을 향상 시킬 수 있었다. 입자군집 최적화가 적용된 속성선택을 사용한 경우에는 각각 데이터마다 2013년은 1,633개, 2014년은 1,315개, 2015년은 3,480개, 마지막으로 2016년은 2,123으로 줄이고도 성능을 유지시킬 수 있었다. 이는 속성선택 전과 비교하여 2013년도는 32.11%, 2014년도는 27.37%, 2015년도는 28.24%, 2016년도는 41.40%에 해당 된다. 입자군집최적화 알고리즘이 적용된 속성선택 방법과 서포트 벡터머신, 랜덤서브스페이스 분류기를 이용하여 우수한 오피니언 마이닝 분류모델을 구축할 수 있다.

본 연구의 실무적 의의는 다음과 같다.

소셜미디어 서비스는 실시간으로 수백만 건의 게시글이 게재된다. 이와 같이 대량의 데이터를 처리하는 경우에는 입자군집 최적화 알고리즘이 적용된 속성선택 방법을 이용하여 오피니언 마이닝 분류모델을 구축할 수 있다. 입자군집최적화 알고리즘이 사용된 모델은 대량의 텍스트 데이터에서 변환된 수많은 속성을 선별하여 데이터 연산량을 줄일 수 있다. 따라서 머신러닝 분류기의 학습 및 검증에 필요한 시간과 메모리 사용량을 줄일 수 있다.

본 연구의 한계점은 다음과 같다. 첫째, 본 연구에서는 트위터 데이터를 사용하여 입자군집 최적화와 다목적 진화 알고리즘을 적용한 속성 셋을 사용하였다. 영화 리뷰 및 아마존 상품 리뷰와 같은 다른 주제의 데이터를 사용하여 속성선택 방법을 검증하여야 한다. 둘째, 본 연구에서 사용된 입자군집 최적화 방법과 다목적 진화 알고리즘이지만 선행 연구에서는 정보획득, 카이제곱 검증 등 방법을 사용하였다. 이에 따라, 다양한 속성선택 방법들을 사용하여 검증하여 상대적인 본 논문에서 사용된 방법의 상대적인 성과를 파악하여야 한다.

향후 연구에서는 오피니언 마이닝에서 백 오브 워즈 방법과 같이 활발하게 사용되고 있는 워드 임베딩 방법에 대한 연구가 필요하다. 워드 임베딩 방법으로 구성된 속성 셋를 입자군집최적화 방법이 사용된 속성선택방법으로 구축된 모델을 검증하여야 한다. 나아가 LSTM(Long-Short Term Memory)과 같은 더욱 발전된 딥러닝 분류 모델과 입자군집 최적화 알고리즘이 적용된 속성선택 방법으로 구축하여 오피니언 마이닝 모델을 한층 더 발전시킬 수 있을 것이다.

REFERENCES

[1] K. Xu, G. Qi, J. Huang, T. Wu, and X. Fu, “Detecting Bursts in Sentiment-Aware Topics from Social Media,” Knowledge-Based Systems, Vol. 141, pp. 44-54, Feb 2018. DOI: 10.1016/j.knosys.2017.11.007

[2] A. Yadollahi, A. G. Shahraki, and O. R. Zaiane, “Current State of Text Sentiment Analysis from Opinion to Emotion Mining,” ACM Computing Surveys (CSUR), Vol. 50, No.2, pp. 1-33, May 2017. DOI: 10.1145/3057270

[3] K. S. Eo, and K. C. Lee, “Exploring an Optimal Feature Selection Method for Effective Opinion Mining Tasks,” Journal of the Korea Society of Computer and Information, Vol. 24, No. 2, pp. 171-177, Feb 2019. DOI: 10.9708/jksci.2019.24.02.171

[4] N. F. Da Silva, E. R. Hruschka, and E. R. Hruschka Jr, “Tweet Sentiment Analysis with Classifier Ensembles,” Decision Support Systems, Vol. 66, pp. 170-179, Oct 2014. DOI: 10.1016/j.dss.2014.07.003

[5] M. A. Friedl, and C. E. Brodley, “Decision Tree Classification of Land Cover from Remotely Sensed Data,” Remote sensing of

environment, Vol. 61, No. 3, pp. 399-409, Sep 1997. DOI: 10.1016/S0034-4257(97)00049-7

[6] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian Network Classifiers," Machine Learning, Vol. 29, No. 2-3, pp. 131-163, Nov 1997. DOI: 10.1023/A:1007465528199

[7] J. A. Suykens, and J. Vandewalle, "Least Squares Support Vector Machine Classifiers,". Neural Processing Letters, Vol. 9, No. 3, pp. 293-300, Jun 1999. DOI: 10.1023/A:1018628609742

[8] L. Breiman, "Random Forests," Machine Learning, Vol. 45, No. 1, pp. 5-32, Oct 2001. DOI: 10.1023/A:1010933404324

[9] L. Breiman, "Bagging Predictors," Machine Learning, Vol. 24, No. 2, pp. 123-140, Aug 1996. DOI: 10.1007/BF00058655

[10] T. K. Ho, "The Random Subspace Method for Constructing Decision Forests," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 8, pp. 832-844, Aug 1998. DOI: 10.1109/34.709601

[11] Q. Ye, Z. Zhang, and R. Law, "Sentiment Classification of Online Reviews to Travel Destinations by Supervised Machine Learning Approaches," Expert Systems with Applications, Vol. 36, No. 3, pp. 6527-6535, Apr 2009. DOI: 10.1016/j.eswa.2008.07.035

[12] R. Moraes, J. F. Valiati, and W. P. G. Neto, "Document-Level Sentiment Classification: An Empirical Comparison Between SVM and ANN," Expert Systems with Applications, Vol. 40, No. 2, pp. 621-633, Feb 2013. DOI: 10.1016/j.eswa.2012.07.059

[13] F. Corea, "Can Twitter Proxy the Investors' Sentiment? The Case for The Technology Sector," Big Data Research, Vol. 4, pp. 70-74, Jun 2016. DOI: 10.1016/j.bdr.2016.05.001

[14] Y. Ruan, A. Durresi, and L. Alfantoukh, "Using Twitter Trust Network for Stock Market Analysis," Knowledge-Based Systems, Vol. 145, pp. 207-218, Apr 2018. DOI: 10.1016/j.knosys.2018.01.016

[15] M. Ghiassi, J. Skinner, and D. Zimbra, "Twitter Brand Sentiment Analysis: A hybrid System Using N-Gram Analysis and Dynamic Artificial Neural Network," Expert Systems with Applications, Vol. 40, No. 16, pp. 6266-6282, Nov 2013. DOI: 10.1016/j.eswa.2013.05.057

[16] G. Wang, J. Sun, J. Ma, K. Xu, and J. Gu, "Sentiment Classification: The Contribution of Ensemble Learning," Decision Support Systems, Vol. 57, pp. 77-93, Jan 2014. DOI: 10.1016/j.dss.2013.08.002

[17] S. Yoo, J. Song, & O. Jeong, "Social Media Contents Based Sentiment Analysis and Prediction System," Expert Systems with Applications, Vol. 105, pp. 102-111, Sep 2018. DOI: 10.1016/j.eswa.2018.03.055

[18] A. García-Pablos, M. Cuadros, & G. Rigau, "W2VLDA: Almost Snsupervised System for Aspect Based Sentiment Analysis," Expert Systems with Applications, Vol. 91, pp. 127-137, Jan 2018. DOI: 10.1016/j.eswa.2017.08.049

[19] J. B. Park, K. S. Lee, J. R. Shin, and K. Y. Lee, "A Particle Swarm Optimization for Economic Dispatch with Nonsmooth Cost Functions," IEEE Transactions on Power Systems, Vol. 20, No. 1, pp. 34-42, Jan 2005. DOI: 10.1109/TPWRS.2004.831275.

[20] M. Amoozegar, and B. Minaei-Bidgoli, "Optimizing Multi-Objective PSO Based Feature Selection Method Using a Feature Elitism Mechanism," Expert Systems with Applications, Vol. 113, pp. 499-514, Dec 2018. DOI: 10.1016/j.eswa.2018.07.013

[21] B. Xue, M. Zhang, and W. N. Browne, "Particle Swarm Optimization for Feature Selection in Classification: A Multi-Objective Approach," IEEE Transactions on Cybernetics, Vol. 43, No. 6, pp. 1656-1671, Dec 2012. DOI: 10.1109/TSMCB.2012.2227469.

[22] L. Cervante, B. Xue, M. Zhang, and L. Shang, "Binary Particle Swarm Optimisation for Feature selection: A Filter Based Approach," In 2012 IEEE Congress on Evolutionary Computation, pp. 1-8, 2012. DOI: 10.1109/CEC.2012.6256452.

[23] N. A. Krisshna, V. K. Deepak, K. Manikantan, and S. Ramachandran, "Face Recognition Using Transform Domain Feature Extraction And PSO-Based Feature Selection," Applied Soft Computing, Vol. 22, pp. 141-161, Sep 2014. 10.1016/j.asoc.2014.05.007

[24] N. Kushwaha, and M. Pant, "Link Based BPSO for Feature Selection in Big Data Text Clustering," Future Generation Computer Systems, Vol. 82, pp. 190-199, May 2018. DOI: 10.1016/j.future.2017.12.005

[25] Z. Wang, M. Li, and J. Li, "A Multi-Objective Evolutionary Algorithm for Feature Selection Based on Mutual Information with a New Redundancy Measure," Information Sciences, Vol. 307, pp. 73-88, Jun 2015. DOI: 10.1016/j.ins.2015.02.031

[26] A. Gaspar-Cunha, "Feature Selection Using Multi-Objective Evolutionary Algorithms: Application to Cardiac SPECT Diagnosis," In Advances in Bioinformatics, pp. 85-92, 2010. DOI: 10.1007/978-3-642-13214-8_11

[27] L. D. Vignolo, D. H. Milone, and J. Scharcanski, "Feature Selection for Face Recognition Based on Multi-Objective Evolutionary Wrappers," Expert Systems with Applications, Vol. 40, No. 13, pp. 5077-5084, Oct 2013. DOI: 10.1016/j.eswa.2013.03.032

[28] U. Mlakar, I. Fister, J. Brest, and B. Potočnik, "Multi-Objective Differential Evolution for Feature Selection in Facial Expression Recognition Systems," Expert Systems with Applications, Vol. 89, pp. 129-137, Dec 2017. DOI: 10.1016/j.eswa.2017.07.037

[29] F. Jiménez, G. Sánchez, J. M. García, G. Sciavicco, and L. Miralles, "Multi-Objective Evolutionary Feature Selection for Online Sales Forecasting," Neurocomputing, Vol. 234, pp. 75-92, Apr 2017. DOI: 10.1016/j.neucom.2016.12.045

[30] Y. Zhang, D. W. Gong, X. Z. Gao, T. Tian, & X. Y. Sun, "Binary Fifferential Evolution with Self-Learning for Multi-Objective Feature Selection," Information Sciences, Vol. 507, pp. 67-85,

Jan 2020. DOI: 10.1016/j.ins.2019.08.040

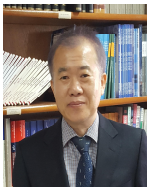
[31] S. Rosenthal, N. Farra, and P. Nakov, “SemEval-2017 Task 4: Sentiment Analysis in Twitter,” arXiv preprint arXiv:1912.00741, 2019. DOI: 10.18653/v1/S17-2088

[32] A. Aizawa, “An Information-Theoretic Perspective of Tf-Idf Measures,” Information Processing & Management, Vol. 39, No. 1, pp. 45-65, Jan 2003. DOI: 10.1016/S0306-4573(02)00021-3

[33] S. Arlot, and A. Celisse, “A Survey of Cross-Validation Procedures for Model Selection,”. Statistics Surveys, Vol. 4, pp. 40-79, 2010. DOI: 10.1214/09-SS054

Authors

Kyun Sun Eo is a Ph.D. student in SKK Business School at Sungkyunkwan University. He is interested in data mining, machine learning, sentiment analysis, and artificial intelligence.



Kun Chang Lee is a full professor of MIS in SKK Business School at Sungkyunkwan University. His recent research interests include emotion mining, health informatics, Human-Robot Interaction (HRI), and artificial intelligence techniques in decision making analysis.