ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2018 September: 24(3): 21~44 http://dx.doi.org/10.13088/jiis.2018.24.3.021

이질성 학습을 통한 문서 분류의 정확성 향상 기법

윌리엄
국민대학교
(*williamwong@kookmin.ac.kr*)

현윤진
국민대학교
(*yoonjin0630@kookmin.ac.kr*)

김남규
국민대학교
(*ngkim@kookmin.ac.kr*)

최근 인터넷 기술의 발전과 함께 스마트 기기가 대중화됨에 따라 방대한 양의 텍스트 데이터가 쏟아져 나오고 있으며, 이러한 텍스트 데이터는 뉴스, 블로그, 소셜미디어 등 다양한 미디어 매체를 통해 생산 및 유통되고 있다. 이처럼 손쉽게 방대한 양의 정보를 획득할 수 있게 됨에 따라 보다 효율적으로 문서를 관리하기 위한 문서 분류의 필요성이 급증하였다. 문서 분류는 텍스트 문서를 둘 이상의 카테고리 혹은 클래스로 정의하여 분류하는 것을 의미하며, K-근접 이웃(K-Nearest Neighbor), 나이브 베이지안 알고리즘(Naïve Bayes Algorithm), SVM(Support Vector Machine), 의사결정나무(Decision Tree), 인공신경망(Artificial Neural Network) 등 다양한 기술들이 문서 분류에 활용되고 있다. 특히, 문서 분류는 문맥에 사용된 단어 및 문서 분류를 위해 추출된 형질에 따라 분류 모델의 성능이 달라질 뿐만 아니라, 문서 분류기 구축에 사용된 학습데이터의 질에 따라 문서 분류의 성능이 크게 좌우된다. 하지만 현실세계에서 사용되는 대부분의 데이터는 많은 노이즈(Noise)를 포함하고 있으며, 이러한 데이터의 학습을 통해 생성된 분류 모형은 노이즈의 정도에 따라 정확도 측면의 성능이 영향을 받게 된다. 이에 본 연구에서는 노이즈를 인위적으로 삽입하여 문서 분류기의 견고성을 강화하고 이를 통해 분류의 정확도를 향상시킬 수 있는 방안을 제안하고자 한다. 즉, 분류의 대상이 되는 원 문서와 전혀 다른 특징을 갖는 이질적인 데이터소스로부터 추출한 형질을 원 문서에 일종의 노이즈의 형태로 삽입하여 이질성 학습을 수행하고, 도출된 분류 규칙 중 문서 분류기의 정확도 향상에 기여하는 분류 규칙만을 추출하여 적용하는 방식의 규칙 선별 기반의 앙상블 준지도학습을 제안함으로써 문서 분류의 성능을 향상시키고자 한다.

주제어 : 텍스트 마이닝, 문서 분류, 이질성 학습, 준지도 학습, 앙상블 학습

논문접수일 : 2018년 7월 16일 논문수정일 : 2018년 8월 13일 게재확정일 : 2018년 8월 28일
원고유형 : 일반논문 교신저자 : 김남규

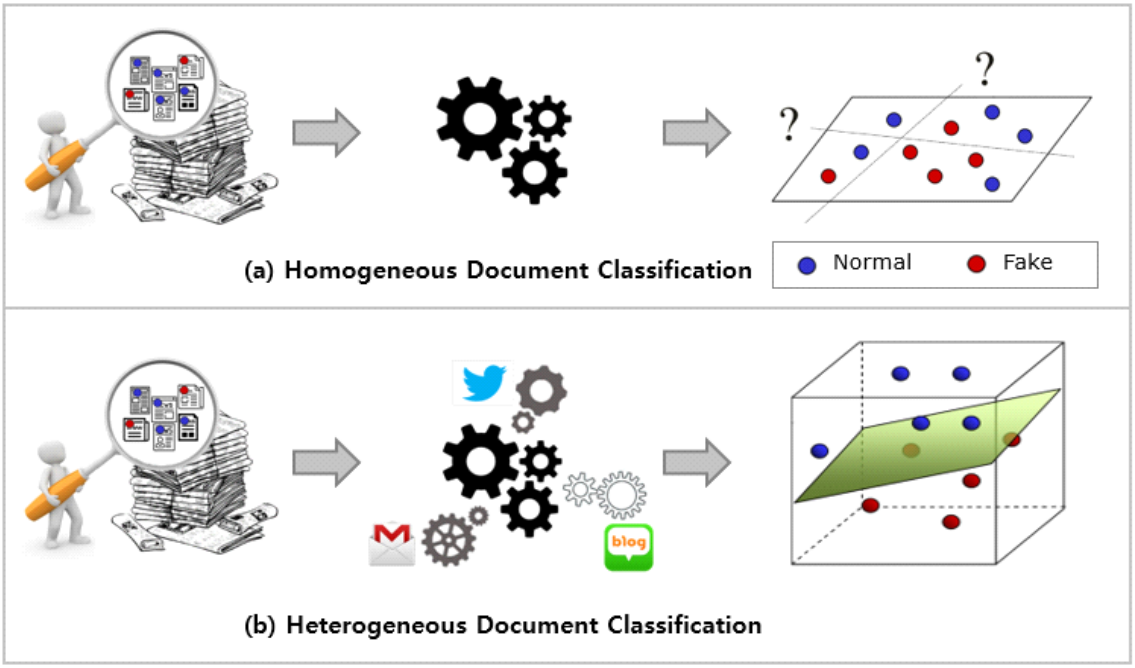
1. 개요

최근 인터넷 기술의 발전과 함께 스마트 기기가 대중화됨에 따라 방대한 양의 텍스트 데이터가 쏟아져 나오고 있으며, 이러한 텍스트 데이터는 뉴스, 블로그, 소셜미디어 등 다양한 미디어 매체를 통해 생산 및 유통되고 있다. 이처럼 손쉽게 방대한 양의 정보를 획득할 수 있게 됨에 따라 보다 효율적으로 문서를 관리하기 위한 문서 분류의 필요성이 급증하였다. 문서 분류는 텍스트 문서를 둘 이상의 카테고리 혹은 클래스로 정의하여 분류하는 것을 의미하며, K-근접 이웃(K-Nearest Neighbor), 나이브 베이지안 알고리즘(Naïve Bayes Algorithm), SVM(Support Vector Machine), 의사결정나무(Decision Tree), 인공신경망(Artificial Neural Network) 등 다양한 기계학습 기술들이 문서 분류에 활용되고 있다.

문서 분류는 문맥에 사용된 단어 및 문서 분류

를 위해 추출된 형질에 따라 분류 모델의 성능이 달라지기 때문에, 이를 활용하여 문서 분류의 정확도를 향상시키기 위한 시도들이 활발히 이루어져 왔다. 대표적으로 Angelova and Weikum (2006)은 정확도 향상을 위한 그래프 기반의 분류 알고리즘을 제안하였고, Mitra et al.(2007)은 LSSVM(Least Square Support Vector Machine) 기반의 LSI(Latent Semantic Index) 계수를 제안함으로써 기존의 선형 SVM 기반 분류기 및 신경망 기반 분류기에 비해 문서 분류의 성능을 크게 향상시켰다. 이처럼 많은 연구들이 문서 분류를 위한 새로운 알고리즘 제안 및 기존 알고리즘 수정을 통해 문서 분류의 성능 향상을 시도하였으며, 이러한 접근 방식은 이미 많은 발전을 이루어 더 이상의 개선이 어려울 정도로 충분히 성숙한 상태이다. 이에 본 연구는 새로운 알고리즘을 제안하거나 기존 알고리즘을 수정하는 접근이 아닌, 분류 모델 구축에 필요한 학습데이터의 활용 방식을 개선하여 문서 분류의 정확도를 향상시키는 방안을 제안하고자 한다.

문서 분류의 성능이 분류기 구축에 사용된 학습데이터의 질에 따라 크게 좌우된다는 것은 널리 알려진 사실이다(Wu and Zhu, 2008). 하지만 현실세계에서 사용되는 대부분의 데이터는 많은 노이즈(Noise)를 포함하고 있으며, 이러한 데이터의 학습을 통해 생성된 분류 모형은 노이즈의 정도에 따라 정확도 측면의 성능이 영향을 받게 된다. 따라서 노이즈에 대한 효율적인 처리를 통해 분류기의 성능을 향상시키기 위한 시도가 꾸준히 이루어져 왔다(Kim et al., 2011; Liu et al., 2015; Sáez et al., 2013). 대부분의 연구는 노이즈가 문서 분류에 미치는 부정적인 영향을 최소화



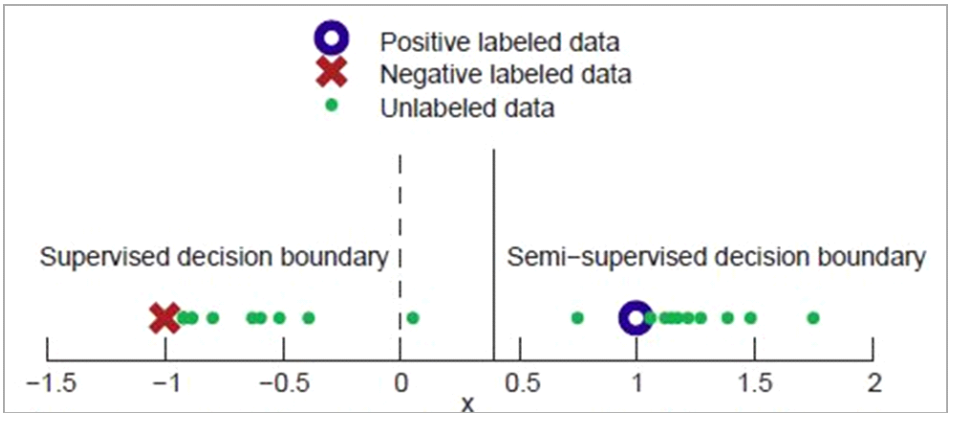
〈Figure 1〉 Comparison between Homogeneous and Heterogeneous Document Classification

하기 위한 방안에 집중하고 있으나, 본 연구에서는 노이즈를 인위적으로 삽입하여 문서 분류기의 견고성을 강화하고 이를 통해 분류의 정확도를 향상시킬 수 있는 방안을 제안하고자 한다. 즉 분류의 대상이 되는 원 문서와 전혀 다른 특징을 갖는 이질적인 데이터소스로부터 형질을 추출하고, 이러한 형질을 원 문서에 일종의 노이즈의 형태로 삽입하여 이질성 학습을 수행하고자 한다(Figure 1).

<Figure 1>은 뉴스 기사의 카테고리를 분류하기 위해 구축한 일반 문서 분류기와 이형질 학습 기반 문서 분류기의 예를 나타내고 있다. <Figure 1>의 (a)는 뉴스 기사에 대한 학습을 통해 분류 모형을 구축하고 이를 뉴스 기사 분류에 적용한 예이며, (b)는 뉴스 기사에 대한 학습 과정에서 이질적 데이터소스의 형질을 삽입하고, 그 결과로 구축된 모형을 뉴스 기사의 분류에 적용한 예이다. 전통적인 분류기 (a)는 대상 데이터와 학습 데이터가 뉴스 데이터로 동일하기 때문에, 분류의 기준이 뉴스 기사에 포함된 어휘로부터만 생성된다는 한계를 갖는다. 따라서 새로운 이슈를 포함하고 있는 뉴스 기사가 등장했을 경우, 기존의 분류 기준으로는 새로 주어진 문서를 정확하게 분류하는 것이 어렵게 된다. 반면 문서 분류기 (b)는 기존의 뉴스 데이터뿐만 아니라 블로그, 소셜미디어 등 다양한 이질적 데이터를 활용해 이형질 학습을 수행하기 때문에, 문서 분류기 (a)에 비해 다양한 관점에서 학습을 진행함으로써 보다 정교한 분류 기준을 가질 수 있게 된다.

본 연구에서는 문서 분류기의 학습 과정에서 이질적 데이터소스를 추가한 이질성 학습을 위해 준지도학습(Semi-Supervised Learning)의 자기훈련(Self-Training) 기법을 활용하고자 한다. 구체적으로는 분류 및 학습의 대상이 되는 원 데이터를 분류 데이터로 사용하고, 이질적 데이터소스로부터 발췌한 데이터를 미분류 데이터로 적용하여 자기훈련을 수행함으로써 이질성 학습을 실현하고자 한다.

하지만 준지도학습이 항상 기존의 기계학습에 비해 우수한 성능을 나타내지는 않으며, 이는 준지도학습의 기본 원리를 나타낸 <Figure 2>를 통해 확인할 수 있다. <Figure 2>에서 각 점은 미분



〈Figure 2〉 Basic Concept of Semi-supervised Learning
(Zhu, X. and Goldberg, A.B., 2009)

류 데이터를, O와 X는 각각 Positive와 Negative로 분류된 데이터를 나타낸다. 이 때 분류 데이터 두 개만을 반영하여 분류 기준을 설정하게 되면 기준선은 점선과 같이(x=0) 형성되게 되며, 미분류 데이터를 모두 반영하여 분류 기준을 설정하면 기준선은 실선과 같이(x는 약 0.4) 형성되게 된다. 즉 점선과 실선 사이에 존재하고 있는 미분류 데이터의 경우 기존의 분류기는 Positive로, 준지도학습 기반 분류기는 Negative로 서로 다르게 결정함을 알 수 있다. 이 때 이 미분류 데이터의 원래 값이 Negative였다면 준지도학습을 통해 분류 정확도가 높아졌겠지만, 원래 값이 만약 Positive였다면 준지도학습을 통해 오히려 분류 정확도가 낮아지는 결과가 초래된다.

따라서 본 연구에서는 이형질 학습을 위해 준지도학습을 활용하되, 위에서 소개한 준지도학습의 한계를 극복하기 위해 규칙 선별 기반의 앙상블 준지도학습(Rule Selection-based Ensemble Semi-supervised Learning: RSESL) 알고리즘을 제안한다. 구체적으로는 (1) 이질적 데이터인 뉴스, 블로그, 트위터 데이터로부터 형질을 추출하여 이질성 학습을 수행하고, (2) 도출된 분류 규칙 중 문서 분류기의 정확도 향상에 기여하는 분류 규칙만을 추출하여 적용하는 방식으로 문서 분류의 성능을 향상시키고자 한다.

본 논문의 이후 부분은 다음과 같이 구성된다. 다음 장인 2장에서는 본 연구와 관련된 선행 연구들을 요약하고, 3장에서는 전체 연구 개요와 제안 방법론을 설명한다. 이후 4장에서는 제안 방법론을 실제 데이터에 적용한 실험 결과를 분석하고, 마지막 5장에서는 본 연구의 기여와 향후 연구 방향을 제시한다.

2. 관련 연구

2.1 데이터 이질성

인터넷과 모바일 서비스 등 최신 기술의 발달로 인해 다양한 매체로부터 방대한 양의 구조화된 정형 데이터 및 구조화되지 않은 비정형 데이터를 획득할 수 있게 됨에 따라 이질적 데이터를 통합하여 사용하려는 시도들이 이루어지고 있다. 이러한 이질적 데이터는 (1) 데이터 유형, 파일 형식, 데이터 인코딩 방식, 데이터 모델 등의 차이를 의미하는 구문 이질성(Syntactic Heterogeneity), (2) 서로 다른 관점에 따라 구성된 데이터로 인한 해석의 차이를 의미하는 의미적 이질성(Semantic Heterogeneity), (3) 데이터의 통계적 특성의 차이를 의미하는 통계적 이질성(Statistical Heterogeneity)의 3가지 범주로 구분할 수 있다(L'Heureux et al., 2017). 대부분 노이즈에 대한 효율적 처리에 관한 연구가 주를 이루고 있으며, 대표적으로 노이즈 데이터를 임의로 삽입하여 학습시킴으로써 기존의 결함 예측(Fault Prediction) 방법론들의 노이즈 허용 오차를 조사한 연구(Kim et al., 2011), 노이즈가 있는 소프트웨어의 결함 예측을 위한 클러스터 기반의 형질 선택 방법론에 관한 연구가 있다(Liu et al., 2015). 또한 Sáez et al.(2013)은 다양한 다중 분류 시스템(MCSs)에 대한 실험을 통해 각각의 다중 분류 시스템들이 노이즈 데이터에 대해 갖는 견고성(Robustness)을 확인하고, 노이즈를 포함한 분류기의 새로운 측정 지표를 제안하였다. 특히, 텍스트 데이터의 경우에는 문서에 포함된 어휘들에 의해 형질이 결정되기 때문에 데이터의 관점에 따라 서로 다른 형질을 갖게 되며, 이는 의미적 이질성의 특징을 갖는다고 볼 수 있다. 하

지만 이러한 이질적 데이터를 활용하여 문서 분류기의 성능을 향상시키고자 하는 시도는 찾아보기 힘들다. 이에 본 연구에서는 문서 분류의 성능을 향상시키기 위해 서로 다른 특징을 갖는 이질적 데이터를 학습데이터로 사용하고자 한다. 특히, 텍스트 데이터의 경우에는 문서에 포함된 어휘들에 의해 형질이 결정되기 때문에 데이터의 관점에 따라 서로 다른 형질을 갖게 되며, 이는 의미적 이질성의 특징을 갖는다고 볼 수 있다. 이에 본 연구에서는 문서 분류의 성능을 향상시키기 위해 서로 다른 특징을 갖는 이질적 데이터를 학습데이터로 사용하고자 한다.

2.2 준지도학습

기존의 기계학습 알고리즘은 충분한 수의 레이블을 갖고 있는 분류 데이터가 있을 때 효과적인 방법이나, 레이블이 없는 미분류 데이터에 레이블을 부여하기 위한 전문가의 판단이 필요할 뿐만 아니라 레이블을 갖는 분류 데이터를 확보를 위해 상대적으로 많은 비용과 시간을 필요로 한다는 어려움이 있다. 반면 레이블이 없는 미분류 데이터의 경우에는 상대적으로 적은 비용과 시간을 투자하여 손쉽게 수집이 가능하기 때문에 이를 활용할 수 있는 준지도학습 알고리즘이 각광을 받고 있다. 준지도학습은 레이블을 갖는 분류 데이터 수가 충분하지 않을 경우, 레이블이 없는 미분류 데이터를 학습데이터로 사용함으로써 더 높은 정확도를 갖는 양질의 분류기를 구축한다. 효과적인 준지도학습을 위해 EM(Empectation Maximization) 기반의 방법론(Nigam et al., 2000; Shahshahani and Landgrebe, 1994), 자기훈련 (Li and Zhou, 2005; Rosenberg et al., 2005; Tanha et al., 2017; Triguero et al., 2014; Yarowsky, 1995), 상호훈련(Co-Training)(Blum and Mitchell, 1998; Tanha et al., 2011), TSVM(Transductive Support Vector Machine) (Joachims, 1999), S3VM(Semi-Supervised SVM)(Bennett and Demiriz, 1999), 그래프 기반 방법론(Belkin et al., 2006; Zhu et al., 2005), 부스팅 기반 방법론(Mallapragada et al., 2009) 등 다양한 기술들이 활용되고 있다.

특히 자기훈련은 준지도학습의 가장 대표적인 기술로써, 자연어 처리(Natural Language Processing) (Ando and Zhang, 2005; McClosky et al., 2006; Yarowsky, 1995), 객체 탐지(Rosenberg et al., 2005), 원격 감지 영상(remote sensing imagery)의 분류(Maulik and Chakraborty, 2011), 문서 분류 (Kim and Kim, 2016) 등 다양한 분야에서 활용되고 있다. 자기훈련은 적은 양의 레이블을 갖는 분류 데이터를 대상 데이터로 하여 학습을 수행하고, 이를 활용하여 레이블이 없는 미분류 데이터를 예측하여 분류하게 된다. 이후 분류된 데이터의 예측값 중 가장 높은 확률값을 갖는 데이터를 학습데이터에 추가하게 되며, 일련의 과정을 반복함으로써 분류기를 구축한다. 자기훈련의 경우, 학습데이터를 추가할 때 특별한 가정을 하지는 않지만 초기 분류 데이터의 학습을 통해 미분류 데이터를 예측한 결과를 가장 정확하다고 판단하여 학습을 수행한다(Triguero et al., 2015). 하지만 초기학습에 사용되는 레이블을 갖는 분류 데이터가 희소하기 때문에 분류기의 신뢰도 확보가 어렵다는 한계가 존재한다. 특히, 자기훈련은 레이블이 없는 미분류 데이터를 점진적 학습을 통해 레이블을 부여하고 추가 학습데이터로 사용하기 때문에 분류기의 성능 저하를 초래할 수 있다. 따라서 본 연구에서는 이를 극복하기 위해 규칙 선별 기반의 앙상블 준지도학습 알

고리즘을 제안하고, 이를 활용하여 이질성 학습을 수행한다.

2.3 앙상블 학습(Ensemble Learning)

앙상블 학습은 여러 개의 분류기를 구축하고 해당 예측 결과들을 결합함으로써 새로운 가설(Hypothesis)을 학습하는 기법으로, 단일 분류기의 성능을 향상시키기 위해 널리 사용되고 있는 기법이다(Dietterich, 2000). 대표적으로 Dasarathy and Sheela(1979)은 2개 이상의 분류기를 활용한 분류기 구축 방안을 제안하였고, Hansen and Salamon(1990)은 인공신경망과 유사하게 구성된 앙상블 기법을 통해 기존 인공신경망 기반 분류기의 성능을 향상시킬 수 있음을 보였다. 또한 Schapire(1990)는 에이다부스트(AdaBoost) 알고리즘의 전신인 부스팅(Boosting)을 통해 성능이 낮은 분류기들을 결합함으로써 더 높은 성능을 갖는 분류기 구축 방안을 제안하였다.

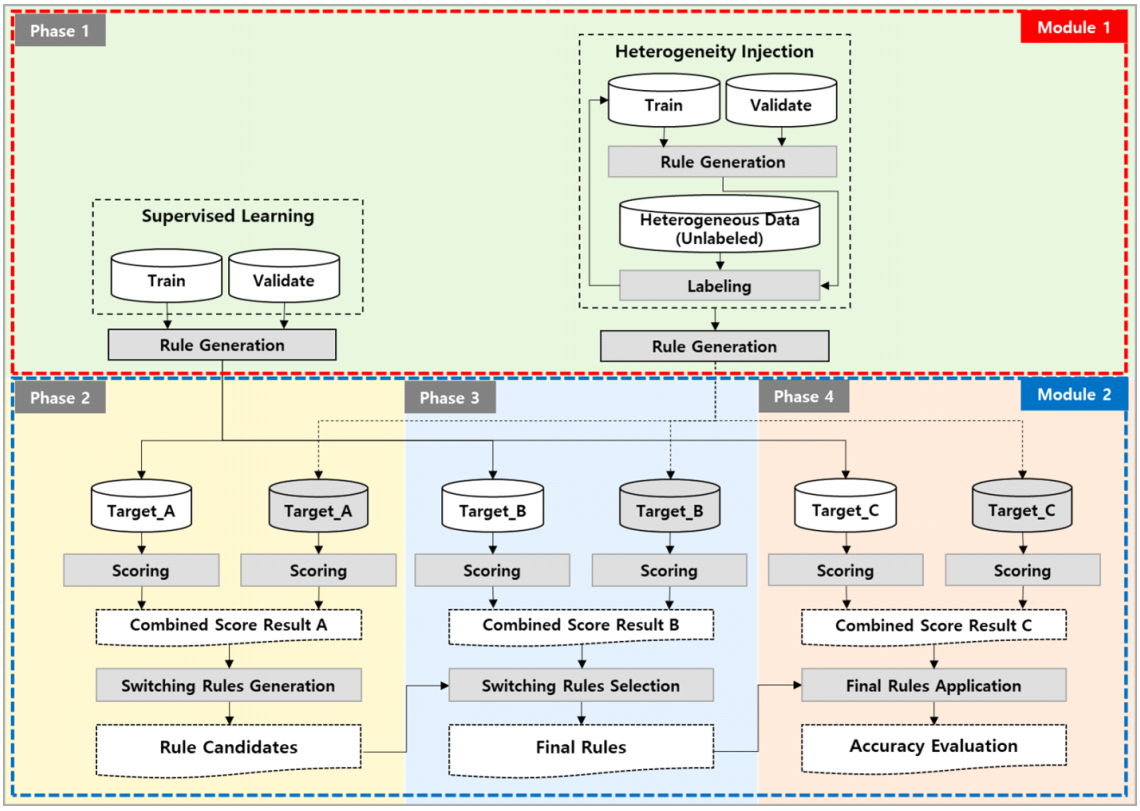
이러한 앙상블 학습을 위해서 중요하게 고려되어야 할 사항은 크게 2가지로, (1) 가능한 다양한 분류기를 통해 (2) 예측 정확도가 높은 분류기 결과를 결합함으로써 앙상블 학습이 이루어져야 한다(Polikar, 2006). 이 때, 다양한 분류기의 구축은 학습데이터, 형질, 파라미터(Parameter) 설정, 분류기의 유형 등을 달리함으로써 가능하며, 보다 많은 분류기를 구축함으로써 해당 결과들 중 단일 분류기보다 성능이 향상된 결과를 결합하는 방식으로 이루어진다. 대표적인 앙상블 학습 기법으로는 배깅(Bagging)(Breiman, 1996; Min, 2014), 부스팅(Freund and Schapire, 1996, Schapire, 1990), 에이다부스트(Freund and Schapire, 1997; Kim, 2012), 스태킹(Stacked Generalization)(Wolpert, 1992), 혼합 전문가(Mixture of Experts) 알고리즘(Jacobs et al., 1991; Jordam and Jacobs, 1994; Jordan and Xu, 1995) 등이 있다. 이에 본 연구에서는 이질적 데이터인 뉴스, 블로그, 트위터 데이터로부터 형질을 추출하여 이질성 학습을 수행한 후 도출된 분류 규칙 중 분류기의 정확도 향상에 기여하는 분류 규칙만을 추출하여 적용하는 방식의 규칙 선별 기반의 앙상블 준지도학습 알고리즘을 제안한다.

3. 규칙 선별 기반 앙상블 준지도학습을 통한 문서 분류 방법론

3.1 연구 개요

본 절에서는 이질적 데이터를 활용한 규칙 선별 기반의 앙상블 준지도학습 방법론을 통해 문서 분류의 성능을 향상시키는 방안에 대해 소개한다. 여기서 이질적 데이터란 레이블이 없는 미분류 데이터를 의미한다. 제안 방법론은 다양한 이질적 데이터소스로부터 새로운 형질을 추출하고, 자기훈련 기법을 활용하여 이질성 학습을 수행한다. 이때, N개의 이질적 데이터소스를 활용하여 이질성 학습을 수행할 경우, N개의 이질성 분류기가 구축되며, 구축된 이질성 분류기들의 결과를 결합하여 가장 높은 예측값을 갖는 분류규칙을 선정함으로써 이질성 학습 기반의 분류규칙을 생성하게 된다. 이렇게 생성된 이질성 기반 분류 규칙은 기존의 기계학습 기반 분류 규칙과 함께 규칙 선별 기반의 앙상블 준지도학습 문서 분류를 위해 활용된다.

<Figure 3>은 본 연구의 전체 개요도를 나타내며, 원통형으로 표시된 부분은 분류 및 학습의 대상이 되는 뉴스 데이터(Train, Validate,



〈Figure 3〉 Research Overview

Target_A, Target_B, Target_C), 레이블이 없는 미분류 데이터인 이질성 데이터(Heterogeneous Data) 등의 데이터소스를 나타낸다. 또한 직사각형으로 표시된 부분은 주요 프로세스를 나타내며, 점선으로 표시된 도형은 각 프로세스의 산출물을 나타낸다.

제안 방법론은 Module 1 이질성 주입(Phase 1)과 Module 2 분류 규칙 선별(Phase 2~4)의 2가지 모듈로 구성된다. Phase 1에 해당하는 Module 1의 이질성 주입은 본 연구의 핵심 부분으로, 분류 및 학습의 대상이 되는 원 데이터에 이질성을 인위적으로 주입시키기 위하여 이질성 학습을 수행한다. 이질성 학습은 준지도학습의 자기훈련 기법을 활용하여 원 데이터에 이질적 데이터로부터 추출한 새로운 형질을 학습데이터로 추가하는 방식으로 이루어진다. 구체적으로는 원 데이터를 학습하여 초기 분류기를 구축하고, 이를 이질적 데이터에 적용함으로써 가장 높은 예측값을 갖는 데이터만을 학습데이터에 추가하게 된다. 이때, 활용되는 이질적 데이터의 원천 소스의 수에 따라 각각의 이질성 분류기가 생성되기 때문에 기존 앙상블 학습 이론을 적용하여 해당 분류기들의 예측 결과를 결합하여 가장 높은 예측값을 갖는 분류 규칙을 선정함으로써 이질

성 학습 기반의 분류 규칙을 생성한다. 이렇게 도출된 이질성 학습 기반 분류 규칙은 원 데이터를 대상으로 하여 기존의 기계학습 알고리즘을 통해 도출된 기계학습 기반 분류 규칙과 함께 이후 Module 2에서 수행될 규칙 선별 기반 앙상블 준지도학습에 활용된다. Phase 2 ~ Phase 4에 해당하는 Module 2의 분류 규칙 선별은 Module 1을 통해 도출된 기계학습 기반 분류 규칙과 이질성 학습 기반 분류 규칙을 활용하여 분류 규칙을 선별하고, 이에 기반하여 최종 문서 분류기를 구축하는 과정이다. 이때, 타겟 데이터를 A, B, C 3개의 데이터집합(뉴스 데이터)로 분리하여 사용하는 것은 문서 분류를 위한 분류 규칙을 선별하고, 검증 과정을 거쳐 최종 선정된 분류 규칙을 테스트하여 문서 분류기의 성능을 확인하기 위함이다. 구체적으로는, Phase 2에서 타겟 데이터 A를 대상으로 기계학습 기반 분류 규칙과 이질성 학습 기반 분류 규칙을 적용하여 각각 스코어링을 수행하고, 해당 결과를 결합함으로써 가장 높은 예측값을 갖는 규칙들 만을 선별하여 분류 규칙 후보군을 생성한다. 다시 말해, 분류 대상인 타겟 데이터가 갖는 형질에 따라 가장 적합한 분류 규칙들을 산출함으로써 문서 분류기 구축에 활용 가능한 분류 규칙 후보군을 생성하게 된다. Phase 3은 Phase 2를 통해 도출된 분류 규칙 후보군 중에서 실제 문서 분류기 성능을 향상시키는데 기여하는 분류 규칙들을 선별하는 단계이다. 타겟 데이터 B를 대상으로 하여 Phase 2와 같은 방식으로 스코어링을 수행하고, 해당 결과를 결합한 후 Phase 2를 통해 산출된 분류 규칙 후보군을 적용하여 타겟 데이터를 정확하게 분류해낸 분류 규칙들 만을 선별하여 문서 분류기 구축을 위한 최종 분류 규칙으로 선정한다. 이후 Phase 4에서 Phase 3을 통해 최종 선정된 분류 규칙의 예측 정확도를 평가함으로써 문서 분류기를 구축할 수 있다. 즉, 타겟 데이터 C를 대상으로 Phase2, 3과 같은 방식으로 스코어링을 수행하고, 해당 결과를 결합한 후 Phase 3을 통해 최종 선정된 분류 규칙을 적용한 예측 정확도를 평가하여 문서 분류기를 구축하게 된다. 제안 방법론에 대한 보다 자세한 설명은 이어지는 3.2절과 3.3절에서 다루도록 한다.

3.2 Module 1: 이질성 주입(Heterogeneity Injection)

본 절에서는 제안 방법론의 핵심인 Module 1의 이질성 주입을 위한 이질성 학습에 대해 소개하며, 구체적으로는 준지도학습의 자기훈련 기법을 통해 서로 다른 이질적 데이터로부터 새로운 형질을 추출하여 학습데이터로 추가함으로써 이질성 학습을 수행하는 과정을 설명한다.

$$TF-IDF(d,t) = TF(d,t) \times IDF(t)$$

$$TF(d,t) = \begin{cases} 0 & if\ freq(d,t) = 0 \\ 1 + \log\left(1 + \log(freq(d,t))\right) & otherwise \end{cases}$$

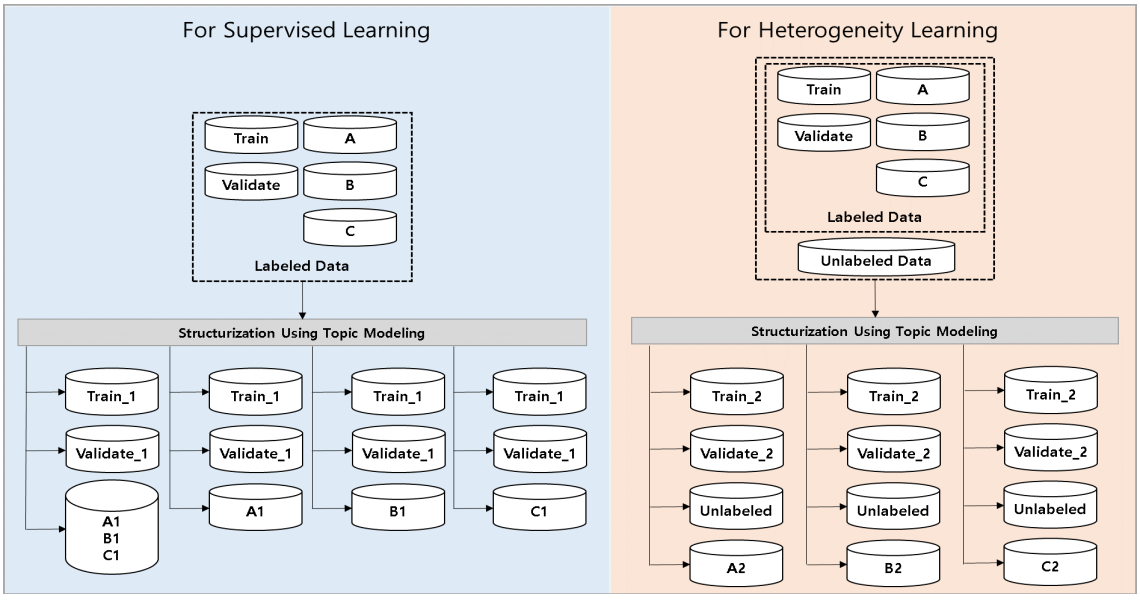
$$IDF(t) = \log \frac{|d|}{|d_t|}$$

제안 방법론의 수행에 앞서 우선적으로 선행되어야 하는 것은 데이터의 구조화 작업이다. 텍스트 데이터의 경우, 구조화되어 있지 않은 비정형 데이터이기 때문에 데이터 분석 가능한 형태로의 변환이 필수적이며, 본 연구에서는 이를 위해 텍스트 마이닝의 대표적 기법인 토픽 모델링(Topic Modeling)을 활용한다. 토픽 모델링은 각 문서에 포함된 용어의 빈도수에 근거하여 유사 문서를 그룹화한 뒤 각 그룹을 대표하는 주요 용

어들을 추출하여 해당 그룹의 토픽 키워드 집합을 제시하는 방식으로 이루어지며(Blei et al., 2003; Hofmann, 2001), 하나의 문서가 여러 토픽에 동시에 대응될 수 있다는 특징을 갖는다. 토픽 모델링의 주요 이론적 배경 중 하나인 TF-IDF는 여러 문서에서 자주 출현하는 일반적인 단어에 대해 가치를 낮게 부여 하고 특정 문서에서만 출현하는 특수한 단어에 대해 가치를 높게 부여하는 계산 방식으로, 각 문서는 용어 수만큼의 차원과 TF-IDF를 값으로 갖는 벡터로 표현된다. 이때, TF-IDF 값은 다음과 같은 방법으로 측정된다.

위 식에서 freq(d, t)는 문서 d에서 용어 t가 출현한 빈도수, |d|는 전체 문서의 수, |d ∈|는 용어 t를 포함하는 문서의 수를 나타낸다. 이때, 용어 t가 다른 문서에서는 출현하지 않고 특정 문서 d에서만 주로 출현한다면, 용어 t가 문서 d에서 나타내는 TF-IDF 값인 TF-IDF(d, t) 값이 높게 나타난다. 토픽 모델링 과정은 이미 기존의 연구들에서 많이 소개되었을 뿐 아니라 일반적인 사용 분석 도구를 활용하여 쉽게 수행할 수 있기 때문에 본 논문에서는 자세히 소개하지 않는다.

데이터 구조화는 각각의 분류기가 적용되는 대상 데이터를 통합하여 토픽 모델링을 수행함으로써 이루어지며, N개의 분류기가 구축될 경우 N번의 토픽 모델링을 통해 데이터 구조화를 수행한다(Figure 4). 이때, 데이터 구조화 프로세스를 분류기에 따라 분리하여 수행하는 것은 이질적 데이터로부터 추출된 형질의 차이가 토픽 가중치에 영향을 주기 때문이다. 예를 들어, 뉴스 데이터를 원 데이터로 한 기계학습 기반 분류기와 트위터, 블로그의 이질적 데이터를 활용한 이질성 분류기를 구축한다고 가정하면, 총 3개의 분류기가 구축되기 때문에 이에 따른 데이터 구조화도 3번의 토픽 모델링 수행을 통해 이루어진다.

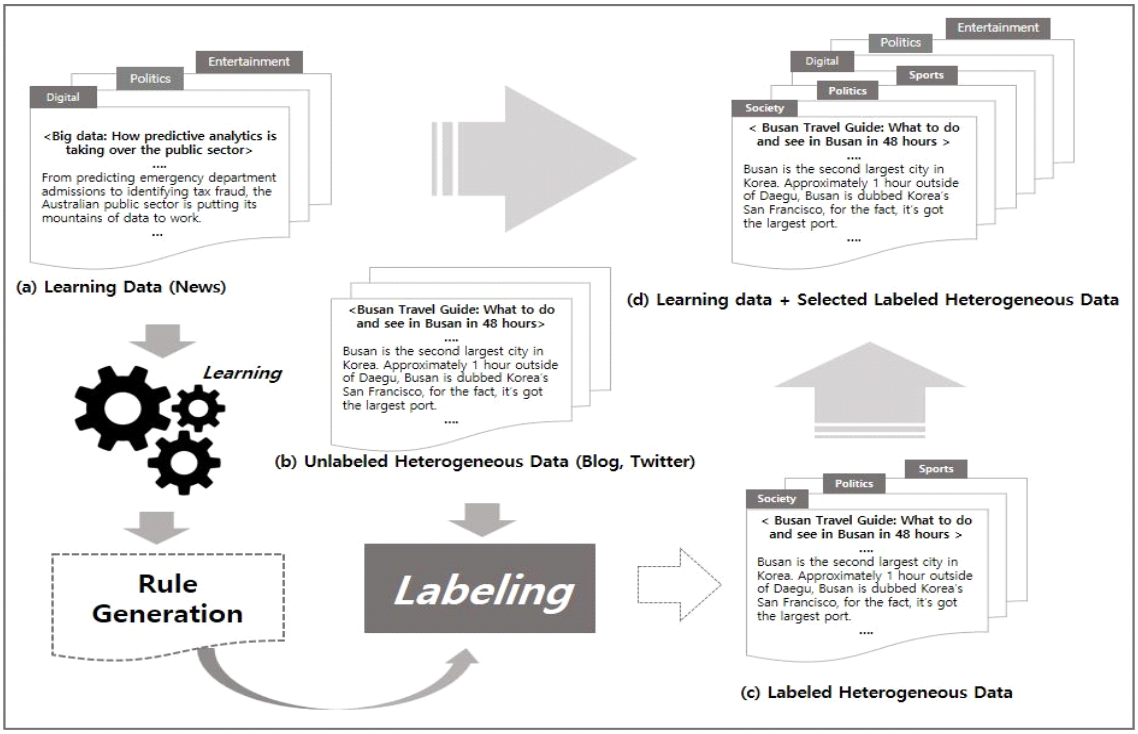


〈Figure 4〉 Data Structurization for Learning

<Table 1>은 데이터 구조화 결과의 예시를 나타내고 있으며, 이처럼 데이터를 구조화한 이후에 해당 데이터를 학습데이터로 활용하여 기존의 기계학습 기반 분류기와 자기훈련 기법을 통한 이질성 학습 기반의 이질성 분류기를 구축한다. <Figure 5>는 이질성 학습을 통한 이질성 주

〈Table 1〉 Example of data structurization result

Doc_No	Topic_1	Topic_2	Topic_3	Topic_4	Topic_5	Label
Doc_1	0.004	0.000	0.155	0.003	0.000	DIGITAL
Doc_2	0.003	0.000	0.274	-0.001	0.004	ENTERTAIN
Doc_3	0.004	0.000	0.074	-0.004	0.000	ENTERTAIN
Doc_4	-0.004	0.000	0.163	-0.001	0.009	DIGITAL
Doc_5	0.004	-0.001	0.002	-0.005	0.011	ENTERTAIN
Doc_6	0.003	-0.001	0.002	-0.003	-0.003	POLITICS
Doc_7	0.009	0.000	0.060	-0.004	0.012	SPORTS
Doc_8	0.002	-0.001	0.021	0.000	0.008	SPORTS
Doc_9	0.006	0.009	0.010	0.004	0.000	POLITICS
Doc_10	0.007	0.000	-0.025	0.000	0.000	DIGITAL



〈Figure 5〉 Example of Heterogeneity Injection Process

입 과정의 예를 나타내고 있다.

<Figure 5>에서 (a) 원 데이터인 뉴스 데이터를 학습시켜 초기 분류 규칙을 생성하여 (b) 미분류된 이질적 데이터인 블로그와 트위터 데이터에 적용함으로써 (c) 예측값에 따라 이질적 데이터에 레이블이 부여된다. 이때, (c)의 결과 중 높은 예측값을 갖는 이질적 데이터들을 선별하여 학습데이터에 추가하여 다시 학습을 시킴으로써 새로운 분류 규칙을 생성하는 방식으로 반복학습을 통해 최종 분류 규칙을 선정하여 문서 분류기를 구축한다. 이처럼 미분류된 이질적 데이터를 학습에 활용해 원 데이터에 이질성을 주입함으로써 학습데이터의 재구성을 통해 이질성 분류기를 구축할 수 있다. 이때, 이질적 데이터 원천 소스의 수에 따라 여러 개의 이질성 분류기가 생성되는데, 해당 분류기들의 예측 결과들을 결합하여 가장 높은 예측값을 갖는 분류 규칙을 선정함으로써 최종적인 이질성 학습 기반의 분류 규칙을 도출한다. 이와 더불어 기존의 기계학습 알고리즘을 통해 원 데이터를 학습시켜 기계학습 기반 분류 규칙을 도출한다.

3.3 Module 2: 분류 규칙 선별

본 절은 <Figure 3>의 Phase 2 ~ Phase 4에 해당하는 부분으로, Module 1의 산출물인 이질성 학습 기반의 분류 규칙과 기계학습 기반의 분류 규칙을 활용하여 규칙 선별 기반의 앙상블 학습 알고리즘을 통한 문서 분류기 구축 과정을 소개한다(Phase 2 ~ Phase 4). 다시 말해, 이질성 학습 기반의 분류 규칙과 기계학습 기반의 분류 규칙을 활용하여 문서 분류기 구축을 위한 분류 규칙 후보군을 선별하고, 검증을 통해 최종 분류 규칙을 선정하여 테스트함으로써 문서 분류기의 성능을 측정하는 과정을 설명한다.

<Table 2>는 <Figure 3>의 타겟 데이터 A를 대상으로 이질성 학습 기반의 분류 규칙(HC)과 기계학습 기반의 분류 규칙(SC)을 적용해 스코어링한 결과를 결합하여 나타낸 예로, 결합된 스코어링 결과에 기반하여 각 분류 규칙에 의한 예측값(Confidence)의 차이(Difference)가 산출된 것을 확인할 수 있다. 이때, 기계학습 기반의 분류 규칙에 따른 예측값과 레이블(카테고리)를 기준으로 하여, 이질성 학습 기반의 분류 규칙에 따른 예측값과의 차이가 양수일 경우 이질성 학습 기반의 분류 규칙에 의해 부여된 레이블이 기준 레이블을 대체하게 된다. 예를 들어, <Table 2>의 문서 1001과 1003이 이질성 학습 기반의 분류 규칙과 기계학습 기반의 분류 규칙 예측값의 차이가 양수를 보임에 따라 “SPORTS”는 “DIGITAL”로 대체되고, “DIGITAL”은 “POLITICS”로 대체되게 되며, 해당 결과는 <Table 3>에 나타나 있다.

〈Table 2〉 Example of combined score result for target dataset A

Doc_No	SC_Confidence	SC_Category	HC_Confidence	HC_Category	Difference
1001	0.404	SPORTS	0.863	DIGITAL	0.458
1002	0.844	ENTERTAIN	0.514	DIGITAL	-0.330
1003	0.682	DIGITAL	0.977	POLITICS	0.296
1004	0.989	DIGITAL	0.682	DIGITAL	-0.307

〈Table 3〉 Example of combined score result with original category

Doc_No	Ori_Category	SC_Confidence	SC_Category	HC_Confidence	HC_Category	Difference	Decision
1001	DIGITAL	0.404	SPORTS	0.863	DIGITAL	0.458	DIGITAL
1002	DIGITAL	0.844	ENTERTAIN	0.514	DIGITAL	-0.330	ENTERTAIN
1003	DIGITAL	0.682	DIGITAL	0.977	POLITICS	0.296	POLITICS
1004	DIGITAL	0.989	DIGITAL	0.682	DIGITAL	-0.307	DIGITAL

하지만 <Table 3>의 문서 1003의 경우, 실제 레이블(Ori_Category)이 “DIGITAL”인데 위의 분류 규칙에 따라 “POLITICS”로 잘못 분류되는 경우가 발생한다. 이처럼 예측값의 차이에만 근거하여 분류 규칙을 생성해 문서를 분류할 경우, 잘못된 분류로 인한 성능의 저하가 초래될 수 있다. 따라서 분류 규칙에 의해 대체된 레이블을 원 데이터의 실제 레이블과 비교 분석하여 레이블 정확도를 산출할 필요가 있다. 따라서 본 연구에서는 문서 분류의 성능 향상을 위한 분류 규칙을 생성하기 위해 (1) 예측값의 차이에 따른 임계값과 (2) 대체된 레이블이 정확도를 활용한 규칙 선별 알고리즘을 제안한다. 규칙 선별 알고리즘은 다음과 같이 수행된다.

```
Loop SC & HC Confidence Result Rows
    If Diff < 0 OR HC_Cat == SC_Cat OR (ORG_Cat != HC_Cat and ORG_Cat != SC_Cat) then
        Continue
    Endif
    If ORG_Cat == HC_Cat and ORG_Cat != SC_Cat then
        CorrectCount = CorrectCount + 1
    Elseif ORG_Cat <> HC_Cat and ORG_Cat == SC_Cat then
        IncorrectCount = IncorrectCount + 1
    Endif
    CurrentDiffGain = CorrectCount - IncorrectCount
    If (Current Row Rule == Next Row Rule) AND (Current Row Diff == Next Row Diff) then
        continue
    Elseif (Current Row Rule != Next Row Rule) then
        NetGain = CurrentDiffGain
    Elseif (Current Row Rule == Next Row Rule) AND (Current Row Diff != Next Row Diff) then
        If NetGain < CurrentDiffGain then
            NetGain = CurrentDiffGain
        Endif
    Endif
    If NetGain > 0
        Print Selected Rule
        NetGain = 0
        CorrectCount = 0
        IncorrectCount = 0
    Endif
End Loop
```

<Table 4>는 제안 알고리즘을 통해 생성된 분류 규칙 후보군의 예를 나타내며, 실제 레이블과 비교하여 제대로 분류된 개수(Correct)와 잘못 분류된 개수(Incorrect)에 따라 순이득(Net Gain)이 산출된다. 이때, 순이득이 0보다 큰 분류 규칙이 최종 분류 규칙 선정을 위한 분류 규칙 후보군으로 선정된다. 위의 과정을 통해 도출된 분류 규칙 후보군 R1, R2, R5, R6, R7은 <Figure 3>의 타겟 데이터 B에 적용되어 Phase 2와 같은 방식으로 분류 규칙 후보군에 대한 검증이 이루어지며, 그 결과의 예가 <Table 5>에 나타나있다.

분류 규칙 후보 중 R7의 경우, <Table 4>에서는 예측값의 차이(Threshold)가 0.4로 양수이기 때문에 “SPORTS”가 “DIGITAL”로 대체되었으나 <Table 5>에서는 타겟 데이터 B에서는 순이득이 0보다 작아 유효하지 않은 분류 규칙으로 구분되어 최종 분류 규칙에서는 제외된 것을 확인할 수 있다. 이렇게 검증 과정을 거쳐 최종적

〈Table 4〉 Example of switching rules generation

Rule No	Rule	Threshold	Correct	Incorrect	Net Gain	Selected Gain (>0)
R1	DIGITAL->ENTERTAIN	0	12	0	12	TRUE
R2	DIGITAL->POLITICS	0	42	14	28	TRUE
R3	ENTERTAIN->DIGITAL	0	21	38	-17	FALSE
R4	ENTERTAIN->POLITICS	0.1	8	9	-1	FALSE
R5	ENTERTAIN->POLITICS	0	12	5	7	TRUE
R6	ENTERTAIN->SPORTS	0	39	20	19	TRUE
R7	SPORTS->DIGITAL	0.4	1	0	1	TRUE
R8	SPORTS->DIGITAL	0.3	1	1	0	FALSE
R9	SPORTS->DIGITAL	0.2	1	3	-2	FALSE
R10	SPORTS->DIGITAL	0.1	2	4	-2	FALSE
R11	SPORTS->DIGITAL	0	7	7	0	FALSE
R12	SPORTS->ENTERTAIN	0	13	13	0	FALSE

〈Table 5〉 Example of rule candidates after switching rules selection

Rule No	Rule	Threshold	Correct	Incorrect	Net Gain	Selected Gain (>0)
R1	DIGITAL->ENTERTAIN	0	11	6	5	TRUE
R2	DIGITAL->POLITICS	0	30	12	18	TRUE
R5	ENTERTAIN->POLITICS	0	15	3	12	TRUE
R6	ENTERTAIN->SPORTS	0	43	18	25	TRUE
R7	SPORTS->DIGITAL	0.4	3	5	-2	FALSE

으로 선정된 분류 규칙은 <Figure 3>의 타겟 데이터 C에 적용되어 문서 분류기를 구축함으로써 제안 방법론의 예측 정확도를 평가하게 된다.

4. 실험

4.1 실험 데이터

본 연구의 실험을 위해서는 뉴스, 트위터, 블로그 3가지 유형의 데이터 소스가 필요하며, 이를 위해 2014년 6월 22일부터 7월 5일까지의 뉴스 기사 총 387,018건, 블로그 데이터 327,554건, 트위터 데이터 14,000,000건을 수집하였다. 뉴스 데이터의 경우 한국의 포털사이트 “D”사의 뉴스 기사 중 디지털, 연예, 정치, 스포츠 4개의 카테고리(레이블)를 대상으로 수집하였다. 실제 실험에서는 카테고리 간 형평성 유지를 위해 8,250건으로 가장 적은 수의 기사를 포함하고 있는 디지털 카테고리를 기준으로, 각 카테고리별로 8,250건의 기사를 추출하여 총 33,000건의 뉴스 기사에 대한 분석을 수행하였다. 이후, 불용어 사전을 사용해 대상 데이터의 전처리를 수행하였으며, 전처리된 뉴스 데이터를 대상으로 랜덤 샘플링을 통해 레이블이 부여된 학습데이터 1,000건, 레이블이 없는 미분류 데이터 20,000건을 추출하였고, 규칙 선별 기반의 생성 및 검증 과정을 위한 타겟 데이터로 나머지 12,000건(각 카테고리별 3,000건)의 뉴스 데이터를 사용하였다. 또한 미분류 뉴스 데이터와 함께 블로그, 트위터 데이터 데이터 모두 레이블이 없는 미분류 데이터로 정의하여 실험에 사용하였다.

4.2 실험 과정 및 결과

본 제안 방법론의 적용 가능성을 알아보기 위해 4.1절에서 소개한 실제 실험 데이터를 대상으로 실험을 수행하였다. 실험을 위해 수집된 데이터를 대상으로 실험 데이터를 선정하였으며, 레이블이 부여된 뉴스 데이터 1000건을 6:4의 비율로 Train 데이터 600건, Validate 데이터 400건으로 구분하여 사용하고, 미분류 데이터로 레이블이 없는 뉴스 데이터 20,000건, 블로그 데이터 20,000건, 트위터 데이터 200,000건을 사용하였다. 또한 타겟 데이터 A, B, C는 뉴스 데이터 12,000건을 4,000건씩 랜덤샘플링하여 사용하였다. 실험은 <Figure 3>의 전체 연구개요도와 같은 흐름으로 이루어지며, 추가로 Module 2의 반복학습을 통해 규칙 선별 과정의 유효성을 검증하고, 전통적 기계학습 기반의 분류기와 정확도를 비교분석을 통해 본 제안 방법론의 성능을 검증하였다.

4.2.1 Module 1: 이질성 주입

실험에 앞서 구축할 분류기에 따라 데이터집합을 재구성하였다. 분류기의 구축을 위해 학습데이터와 대상 데이터로 구성된 데이터 집합이 공통적으로 필요하나, 분류기의 특징에 따라 학습데이터 집합의 차이가 존재한다. 기계학습 기반의 분류기는 학습데이터로 레이블이 있는 분류 데이터를 사용하지만 이질성 학습 분류기는 초기 학습을 위한 분류 데이터와 이질성 학습을 위한 미분류 데이터 집합을 학습데이터로 필요로 한다. 이에 따라 본 실험에서는 기계학습 분류기를 위한 데이터 집합, 미분류 뉴스 데이터 기반의 이질성 분류기를 위한 데이터 집합, 블로그 데이터 기반의 이질성 분류기를 위한 데이터

집합, 트위터 데이터 기반의이질성 분류기를 위한 데이터 집합 총 4가지 유형의 데이터 집합을 구성하여 <Figure 4>와 같은 방식으로 각각 구조화한 후, 실험을 수행하였다.

기계학습 기반 분류기의 경우, 인공신경망 알고리즘을 통해 분류 규칙을 도출하였으며, 이질성 분류기의 경우, <Figure 5>와 같은 방식으로 미분류 뉴스 데이터, 블로그, 트위터 각각의 이질적 데이터소스를 활용해 원 데이터에 이질성을 주입하기 위한 이질성 학습을 통해 총 3개의 이질성 분류기를 구축하였다. 그 결과, 미분류 뉴스 데이터 기반의 이질성 분류기의 경우, 임계값이 0.9이상인 이질적 데이터가 18,263건으로 해당 미분류 데이터 중 91% 비율로 학습데이터에 추가되었고, 동일한 임계값 조건 하에서 블로그 기반의 이질성 분류기는 16,737건(84%), 트위터 기반의 이질성 분류기는 101,100건(50%)의 데이터가 각 분류기의 학습데이터로 추가되었다. 이후, 이질적 데이터의 추가로 재구성된 학습데이터를 대상으로 이질성 학습을 수행하여 각각의 이질성 분류기를 구축하였으며, 이질성 학습 기반의 분류 규칙을 도출하기 위해 각 분류기들의 결과를 결합하여 예측값이 가장 높은 분류 규칙을 선별하였다. <Table 6>은 해당 결과의 일부를 나타내며, 이질성 분류기별 예측값과 해당 분류기들의 예측 결과들을 결합하여 가장 높은 예측값을 갖는 분류 규칙을 보여주고 있다.

4.2.2 Module 2: **분류 규칙 선별**

이후 4.2.1의 Module 1을 통해 도출된 기계학습 기반 분류 규칙과 이질성 학습 기반 분류 규칙을 타겟 데이터 A, B, C 각각에 적용하여 <Figure 3>의 Phase 2 ~ Phase 4와 같은 방식으로, 분류 규칙 후보군을 생성하고(<Table 7> 참조), 해당 분류 규칙 후보군의 검증을 거쳐 최종적으로 문서 분류기에 사용될 최종 규칙을 선정

〈Table 6〉 Results of heterogeneity injection through heterogeneity learning (Part)

No	News		Blog		Twitter		Heterogeneity	
	Confidence	Category	Confidence	Category	Confidence	Category	Confidence	Category
1028	0.973	DIGITAL	0.422	SPORTS	0.795	DIGITAL	0.973	DIGITAL
1305	0.976	DIGITAL	0.995	DIGITAL	0.876	DIGITAL	0.995	DIGITAL
1337	0.977	DIGITAL	0.985	DIGITAL	0.696	DIGITAL	0.985	DIGITAL
2047	0.953	ENTERTAIN	0.880	SPORTS	0.994	SPORTS	0.994	SPORTS
2050	1.000	ENTERTAIN	0.864	DIGITAL	0.621	ENTERTAIN	1.000	ENTERTAIN
2059	0.695	SPORTS	0.749	SPORTS	0.994	SPORTS	0.994	SPORTS
3085	0.977	POLITICS	0.816	POLITICS	0.967	POLITICS	0.977	POLITICS
3274	0.971	POLITICS	0.623	SPORTS	0.470	SPORTS	0.971	POLITICS
3276	0.980	POLITICS	0.951	POLITICS	0.992	POLITICS	0.992	POLITICS
4931	0.972	DIGITAL	1.000	ENTERTAIN	0.389	SPORTS	1.000	ENTERTAIN
4762	0.874	SPORTS	0.691	SPORTS	0.994	SPORTS	0.994	SPORTS
4928	0.991	SPORTS	0.933	SPORTS	0.994	SPORTS	0.994	SPORTS

한 후 테스트를 통해 문서 분류기의 예측 정확도를 측정하였다. 이때, 분류 규칙 후보군을 타겟 데이터 B에 적용하여 순이득이 1보다 작은 분류 규칙 후보는 탈락시키고, 1보다 큰 분류 규칙을 최종 분류 규칙으로 선정하여 타겟 데이터 C에 적용해 테스트를 수행하여 문서 분류기의 예측 정확도를 평가하였다. 본 연구에서는 제안 방법론의 규칙 선별 과정의 타당성을 증명하기 위해 타겟 데이터 A, B, C의 순서를 변화하여 타겟 데이터 B, C, A와 C, A, B의 순으로 동일한 실험을 수행하였으며, 해당 결과를 활용하여 전통적 기계학습 기반의 문서 분류기와 성능 비교를 통해 본 제안 방법론의 검증을 수행하였다.

4.2.3 검증

제안 방법론(RSESL Classifier)과 전통적 기계학습 기반 문서 분류기의 성능을 비교 분석한 결과가 <Table 8>에 나타나있다. <Table 8>에서 전통적 기계학습 기반 문서 분류기와 제안 방법론을 통해 예측 및 부여된 레이블을 12,000건의 대상 데이터의 실제 레이블과 비교한 결과, 전통적

〈Table 7〉 Rule candidates extracted results (Part)

No	Ori_Category	SC_Confidence	SC_Category	HC_Confidence	HC_Category	Difference
1028	DIGITAL	0.908	SPORTS	0.973	DIGITAL	0.065
1305	DIGITAL	0.948	POLITICS	0.995	DIGITAL	0.047
1337	DIGITAL	0.942	POLITICS	0.985	DIGITAL	0.043
2047	ENTERTAIN	0.976	ENTERTAIN	0.994	SPORTS	0.019
2050	ENTERTAIN	0.849	DIGITAL	1.000	ENTERTAIN	0.151
2059	ENTERTAIN	0.998	SPORTS	0.994	SPORTS	-0.004
3085	POLITICS	0.955	POLITICS	0.977	POLITICS	0.022
3274	POLITICS	0.401	SPORTS	0.970	POLITICS	0.569
3276	POLITICS	0.794	ENTERTAIN	0.992	POLITICS	0.198
4931	SPORTS	0.642	SPORTS	1.000	ENTERTAIN	0.358
4762	SPORTS	0.946	ENTERTAIN	0.994	SPORTS	0.048
4928	SPORTS	0.836	ENTERTAIN	0.994	SPORTS	0.158

〈Table 8〉 Accuracy comparison of supervised classifier and RSESL classifier

		Traditional Classifier		RSESL Classifier	
Category	Actual	Predicted	Correct	Predicted	Correct
ENTERTAIN	3000	2947	2724	2908	2747
SPORTS	3000	3169	2792	3066	2829
DIGITAL	3000	2831	2566	2912	2642
POLITICS	3000	3053	2754	3114	2790
Total	12000	12000	10836	12000	11008

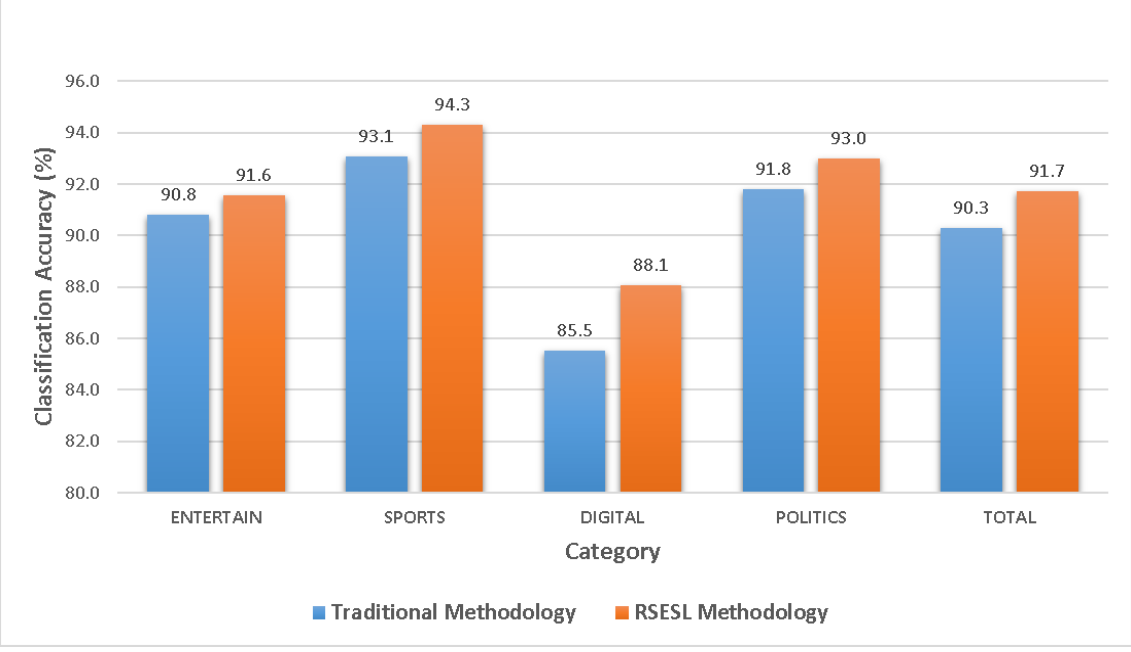
기계학습 기반 문서 분류기는 10,836건, 제안 방법론은 11,008건의 문서를 정확히 예측 분류해냄으로써 제안 방법론이 보다 정확한 예측 분류를 통해 문서 분류기의 성능을 향상시켰음을 확인하였다.

<Figure 6>은 전통적 기계학습 기반 문서 분류기와 제안 방법론의 예측 정확도를 각 카테고리별로 비교 분석한 결과를 그래프로 나타낸 것으로, 연예 카테고리는 0.8%p, 스포츠 카테고리는 1.2%p, 디지털 카테고리는 2.6%p, 정치 카테고리는 1.2%p로 모든 카테고리에 대해 예측 정확도가 향상됨을 확인하였으며, 전체 카테고리 기준으로 전통적 기계학습 기반 문서 분류기는 90.3%, 제안 방법론은 91.7%로 전통적 문서 분류기 대비 문서 분류의 예측 정확도를 1.4%p 향상시킴을 보였다.

5. 결론

본 연구는 문서 분류의 정확도를 향상시키기 위한 방안으로 규칙 선별 기반의 앙상블 준지도 학습 알고리즘을 제안하였다. 제안 방법론은 이질적 데이터인 뉴스, 블로그, 트위터 데이터로부터 새로운 형질을 추출하여 이질성 학습을 수행함으로써 원 데이터에 이질성을 주입하고, 이를 활용하여 도출된 분류 규칙 중 문서 분류기의 정확도 향상에 기여하는 분류 규칙만을 추출하여 적용하는 방식으로 이루어지며, 전통적 기계학습 기반 문서 분류기에 비해 예측 정확도가 1.4%p 증가함을 보임으로써 제안 방법론을 통해 문서 분류의 성능을 향상시킬 수 있음을 증명하였다.

제안 방법론은 다음의 측면에서 학술적, 실무적 차원의 기여를 갖는다. 우선 학술적 측면에서



〈Figure 6〉 Classification Accuracy Comparison for Each Category

제안 방법론은 서로 다른 형질을 갖는 이질적 데이터의 활용을 통해 이질성 학습을 수행함으로써 문서 분류의 성능을 향상시키는 방안을 제안했다는 점에서 의의를 갖는다. 이는 기준 문서와 동일한 데이터 소스의 문서뿐만 아니라, 상이한 형질을 갖는 데이터 소스로부터 이질적 데이터를 추출하여 이를 학습데이터로 보강함으로써 문서 분류기 구축에 활용한 것은 매우 새로운 시도로 인정받을 수 있다고 판단된다. 또한 기존의 기계학습 기반 분류기와 이질성 분류기를 통한 분류 규칙 도출을 통해 데이터가 갖는 형질에 따라 적합한 분류 규칙을 선별적으로 적용함으로써 보다 정확한 문서 분류를 가능하게 했다는 점에서 그 기여를 인정 받을 수 있다. 한편 실무적 측면에서 제안 방법론은 이질적 데이터를 활용해 기존의 동질적 분류 데이터를 활용한 분류 규칙을 보완함으로써 실시간으로 생겨나는 방대한 양의 텍스트 데이터를 효율적으로 분류하고 관리할 수 있다는 점에서 그 기여를 크게 인정받을 수 있을 것으로 기대한다.

하지만 본 연구는 향후 다음의 측면에서 보완이 필요하다. 본 연구는 1,000건의 학습데이터를 사용해 실험이 수행되었으며, 이질성 데이터를 활용한 분류 규칙의 정확도가 향상됨을 보였으나, 향상 폭이 상대적으로 작다는 한계가 있다. 따라서 향후 연구에서는 이질성 주입 과정에서 이질성 데이터의 양이 분류 규칙 정확도에 미치는 영향을 파악할 필요가 있다. 또한 레이블이 없는 미분류 데이터의 유형에 따라 실험 결과가 달라질 수 있기 때문에 다른 잠재적 이질성 데이터를 활용하여 반복 실험을 수행할 필요가 있으며, 제안 방법론의 확장성을 높이기 위해 본 연구에서 수동으로 수행되었던 분석 단계들에 대한 자동화가 필요하다.

참고문헌(References)

Ando, R. K. and T. Zhang, "A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data," *Journal of Machine Learning Research*, Vol. 6 (2005), 1817~1853.

Angelova, R. and G. Weikum, "Graph-Based Text Classification: Learn from Your Neighbors," *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (2006), 485~492.

Belkin, M., P. Niyogi, and V. Sindhwani, "Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples," *Journal of Machine Learning Research*, Vol. 7(2006), 2399~2434.

Bennett, K. P. and A. Demiriz, "Semi-Supervised Support Vector Machines," *Advances in Neural Information Processing Systems*, Vol. 11(1999), 368~374.

Blei, D.M., A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, Vol. 3, No. Jan(2003), 993~1022.

Blum, A. and T. Mitchell, "Combining Labeled and Unlabeled Data with Co-Training," *Proceedings of the eleventh annual conference on Computational learning theory*, (1998), 92~100.

Breiman, L., "Bagging Predictors," *Machine learning*, Vol. 24, No. 2(1996), 123~140.

Dasarathy, B. V. and B. V. Sheela, "A Composite Classifier System Design: Concepts and Methodology," *Proceedings of the IEEE*, Vol. 67, No. 5(1979), 708~713.

Dietterich, T.G., "Ensemble Methods in Machine Learning," *Multiple Classifier Systems*, Vol. 1857(2000), 1~15.

Freund, Y. and R. E. Schapire, "Experiments with a New Boosting Algorithm," *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, (1996),148~156.

Freund, Y. and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, Vol. 55, No. 1(1997), 119~139.

Hansen, L. K. and P. Salamon, "Neural Network Ensembles," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 12, No. 10(1990), 993~1001.

Hofmann, T., "Unsupervised Learning by Probabilistic Latent Semantic Analysis," *Machine learning*, Vol. 42, No. 1-2(2001), 177~196.

Jacobs, R. A., M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive Mixtures of Local Experts," *Neural Computation*, Vol. 3, No. 1(1991), 79~87.

Joachims, T., "Transductive Inference for Text Classification using Support Vector Machines," *International Conference on Machine Learning*, Vol. 99(1999), 200~209.

Jordan, M. I. and L. Xu, "Convergence Results for the EM Approach to Mixtures of Experts Architectures," *Neural Networks*, Vol. 8, No. 9(1995), 1409~1431.

Jordan, M. I. and R. A. Jacobs, "Hierarchical Mixtures of Experts and the EM Algorithm," *Neural Computation*, Vol. 6, No. 2(1994), 181~214.

Kim, M., "Ensemble Learning with Support Vector Machines for Bond Rating," *Journal of Intelligence and Information Systems*, Vol. 18, No. 2(2012), 29~45.

Kim, D., and N. Kim, "Mapping Categories of Heterogeneous Sources using Text Analytics," *Journal of Intelligence and Information Systems*, Vol. 22, No. 4(2016), 193~215.

Kim, S., H. Zhang, R. Wu, and L. Gong, "Dealing with Noise in Defect Prediction," *Proceedings of the 33rd International Conference on Software Engineering*, (2011), 481~490.

L'Heureux, A., K. Grolinger, H. F. ElYamany, and M. Capretz, "Machine Learning with Big Data: Challenges and Approaches," *IEEE Access*, Vol. 5(2017), 7776~7797.

Li, M. and Z. H. Zhou, "SETRED: Self-Training with Editing," *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Vol. 3518(2005), 611~621.

Liu, W., S. Liu, Q. Gu, X. Chen, and D. Chen, "Fecs: A Cluster based Feature Selection Method for Software Fault Prediction with Noises," *IEEE 39th Annual Computer Software and Applications Conference (COMPSAC)*, Vol. 2(2015), 276~281.

Mallapragada, P. K., R. Jin, A. K. Jain, and Y. Liu, "Semiboost: Boosting for Semi-Supervised Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 31, No. 11(2009), 2000~2014.

Maulik, U. and D. Chakraborty, "A Self-Trained Ensemble with Semisupervised SVM: An Application to Pixel Classification of Remote

Sensing Imagery," *Pattern Recognition*, Vol. 44, No. 3(2011), 615~623.

McClosky, D., E. Charniak, and M. Johnson, "Effective Self-Training for Parsing," *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, (2006), 152~159.

Min, S., "Bankruptcy Prediction using an Improved Bagging Ensemble," *Journal of Intelligence and Information Systems*, Vol. 20, No. 4(2014), 121~139.

Mitra, V., C. J. Wang, and S. Banerjee, "Text Classification: A Least Square Support Vector Machine Approach," *Applied Soft Computing,* Vol. 7, No. 3(2007), 908~914.

Nigam, K., A. K. McCallum, S. Thrun, and T. Mitchell, "Text Classification from Labeled and Unlabeled Documents using EM," *Machine Learning*, Vol. 39, No. 2(2000), 103~134.

Provost, F. and T. Fawcett, Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking, O'Reilly Media, Inc., California, 2013.

Polikar, R., "Ensemble based Systems in Decision Making," *IEEE Circuits and Systems Magazine*, Vol. 6, No. 3(2006), 21~45.

Rosenberg, C., M. Hebert, and H. Schneiderman, "Semi-Supervised Self-Training of Object Detection Models," *Seventh IEEE Workshops on Application of Computer Vision*, Vol. 1(2005), 29~36.

Sáez, J.A., M. Galar, J. Luengo, and F. Herrera, "Tackling the Problem of Classification with Noisy Data using Multiple Classifier Systems: Analysis of the Performance and Robustness," *Information Sciences*, Vol. 247(2013), 1~20.

Salton, G. and C. Buckley, "Term Weighting Approaches in Automatic Text Retrieval," Technical Report, Cornell University, 1987.

Schapire, R.E., "The Strength of Weak Learnability," *Machine Learning*, Vol. 5, No. 2(1990), 197~227.

Shahshahani, B.M. and D. A. Landgrebe, "The Effect of Unlabeled Samples in Reducing the Small Sample Size Problem and Mitigating the Hughes Phenomenon," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 32, No. 5(1994), 1087~1095.

Tanha, J., M. van Someren, and H. Afsarmanesh, "Disagreement-based Co-Training," *23rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, (2011), 803~810.

Tanha, J., M. van Someren, and H. Afsarmanesh, "Semi-Supervised Self-Training for Decision Tree Classifiers," *International Journal of Machine Learning and Cybernetics*, Vol. 8, No. 1(2017), 355~370.

Triguero, I., J. A. Sáez, J. Luengo, S. García, and F. Herrera, "On the Characterization of Noise Filters for Self-Training Semi-Supervised in Nearest Neighbor Classification," *Neurocomputing*, Vol. 132(2014), 30~41.

Triguero, I., S. García, and F. Herrera, "Self-Labeled Techniques for Semi-Supervised Learning: Taxonomy, Software and Empirical Study," *Knowledge and Information Systems*, Vol. 42, No. 2(2015), 245~284.

Wolpert, D.H., 1992. "Stacked Generalization,"

Neural Networks, Vol. 5, No. 2(1992), 241~259.

Wu, X. and X. Zhu, "Mining with Noise Knowledge: Error-Aware Data Mining," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, Vol. 38, No. 4(2008), 917~932.

Yarowsky, D., "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods," *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, (1995), 189~196.

Zhu, X., "Semi-Supervised Learning Literature Survey," Computer Sciences TR 1530, University of Wisconsin, 2008. Available at http://pages.cs.wisc.edu/;jerryzhu/pub/ssl_survey.pdf

Zhu, X. and A. B. Goldberg, "Introduction to Semi-Supervised Learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, Vol. 3, No. 1(2009), 1~130.

Zhu, X., J. Lafferty, and R. Rosenfeld, "Semi-Supervised Learning with Graphs," *Doctoral Dissertation*, Language Technologies Institute, Carnegie Mellon University, 2005.

Abstract

Improving the Accuracy of Document Classification by Learning Heterogeneity

William Xiu Shun Wong* · Yoonjin Hyun* · Namgyu Kim**

In recent years, the rapid development of internet technology and the popularization of smart devices have resulted in massive amounts of text data. Those text data were produced and distributed through various media platforms such as World Wide Web, Internet news feeds, microblog, and social media. However, this enormous amount of easily obtained information is lack of organization. Therefore, this problem has raised the interest of many researchers in order to manage this huge amount of information. Further, this problem also required professionals that are capable of classifying relevant information and hence text classification is introduced. Text classification is a challenging task in modern data analysis, which it needs to assign a text document into one or more predefined categories or classes. In text classification field, there are different kinds of techniques available such as K-Nearest Neighbor, Naïve Bayes Algorithm, Support Vector Machine, Decision Tree, and Artificial Neural Network.

However, while dealing with huge amount of text data, model performance and accuracy becomes a challenge. According to the type of words used in the corpus and type of features created for classification, the performance of a text classification model can be varied. Most of the attempts are been made based on proposing a new algorithm or modifying an existing algorithm. This kind of research can be said already reached their certain limitations for further improvements. In this study, aside from proposing a new algorithm or modifying the algorithm, we focus on searching a way to modify the use of data. It is widely known that classifier performance is influenced by the quality of training data upon which this classifier is built. The real world datasets in most of the time contain noise, or in other words noisy data, these can actually affect the decision made by the classifiers built from these data. In this study, we consider that the data from different domains, which is heterogeneous data might have the characteristics of noise which can be utilized in the classification process.

* Kookmin University
** Corresponding Author: Namgyu Kim
Kookmin University
77 Jeongneung-ro, Seongbuk-gu, Seoul, 02707, Korea
Tel: +82-2-910-5425, Fax: +82-2-910-4017, E-mail: ngkim@kookmin.ac.kr

In order to build the classifier, machine learning algorithm is performed based on the assumption that the characteristics of training data and target data are the same or very similar to each other. However, in the case of unstructured data such as text, the features are determined according to the vocabularies included in the document. If the viewpoints of the learning data and target data are different, the features may be appearing different between these two data. In this study, we attempt to improve the classification accuracy by strengthening the robustness of the document classifier through artificially injecting the noise into the process of constructing the document classifier.

With data coming from various kind of sources, these data are likely formatted differently. These cause difficulties for traditional machine learning algorithms because they are not developed to recognize different type of data representation at one time and to put them together in same generalization. Therefore, in order to utilize heterogeneous data in the learning process of document classifier, we apply semi-supervised learning in our study. However, unlabeled data might have the possibility to degrade the performance of the document classifier. Therefore, we further proposed a method called Rule Selection-Based Ensemble Semi-Supervised Learning Algorithm (RSESLA) to select only the documents that contributing to the accuracy improvement of the classifier. RSESLA creates multiple views by manipulating the features using different types of classification models and different types of heterogeneous data. The most confident classification rules will be selected and applied for the final decision making. In this paper, three different types of real-world data sources were used, which are news, twitter and blogs.

Key Words : Text Mining, Text Classification, Heterogeneity Learning, Semi-Supervised Learning, Ensemble Learning

Received : July 16, 2018 Revised : August 13, 2018 Accepted : August 28, 2018
Publication Type : Regular Paper Corresponding Author : Namgyu Kim

저 자 소 개

William Xiu Shun Wong
현재 국민대학교 비즈니스IT전문대학원 박사과정에 재학 중이다. 말레이시아과학대학교 컴퓨터공학과에서 학사 학위를 취득하고, 국민대학교 비즈니스IT전문대학원에서 비즈니스IT를 전공하여 경영정보학 석사학위를 취득하였다. 주요 관심분야는 텍스트 마이닝, 준지도학습, 데이터 마이닝 및 소셜네트워크분석 등이다.

현윤진
현재 국민대학교 비즈니스IT전문대학원 박사과정에 재학 중이다. 국민대학교 경영정보학부에서 학사 학위를 취득하고, 본교 비즈니스IT전문대학원에서 비즈니스IT를 전공하여 경영정보학 석사학위를 취득하였다. 주요 관심분야는 텍스트 마이닝, 데이터 마이닝 및 소셜네트워크분석 등이다.

김남규
현재 국민대학교 경영정보학부 부교수로 재직 중이다. 서울대학교 컴퓨터공학과에서 학사 학위를 취득하고, KAIST 테크노경영대학원에서 Database와 MIS를 전공하여 경영공학 석사 및 박사학위를 취득하였다. 한국지능정보시스템학회 부회장, 한국정보기술응용학회 부회장, 한국경영학회 상임이사, 한국경영정보학회 이사, 한국인터넷정보학회 이사, 한국CRM학회 이사를 역임하였다. 주요 관심분야는 Text Mining, Data Mining, Data Modeling 등이다.