하이브리드 다중 모델 학습 기법을 이용한 자동 문서 분류

명순희[0] 조형근 김인철
경기대학교 전자계산학과
shmyoung@ysc.ac.kr serice4u@hotmail.com kic@kuic.kyonggi.ac.kr

Automatic Text Classification Using Hybrid Multiple Model Schemes

Soon-Hee Myoung[0] Hyung-Kun Cho In-Cheol Kim
Dept. of Computer Science, Kyonggi University

요 약

본 논문에서는 다중 모델 기계학습 기법을 이용하여 문서 자동 분류의 성능과 신뢰도를 향상시킬 수 있는 연구와 실험 결과를 기술하였다. 기존의 다중 모델 기계 학습법들이 훈련 데이터 또는 학습 알고리즘의 편향에 의한 오류를 극복하고 한 것들인데 비해 본 논문에서 제안한 메타 학습을 이용한 하이브리드 다중 모델 방식은 이 두 가지의 오류 원인을 동시에 해소하고자 하였다. 다양한 문서 집합에 대한 실험 결과, 본 연구에서 제안한 하이브리드 다중 모델 학습법이 전반적으로 기존의 일반 다중모델 학습법들에 비해 높은 성능을 보였으며, 다중 모델의 결합 방식으로서 메타 학습이 투표 방식에 비해 효율적인 것으로 나타났다.

1. 서론

사전 분류된 훈련예를 바탕으로 데이터의 구조적 패턴을 명시적 지식으로 학습하여 신규 사례에 대하여 분류 항목을 제시하는 귀납적 교사학습 기법은 문서 분류에도 높은 정확도를 보인다[1].분류기, 또는 학습된 모델은 훈련 데이터에 대한 학습알고리즘의 적용 결과로 유도된 구조적 패턴으로 데이터 구조를 이용하여 미분류 데이터에 클라스를 매핑하는 기능을 갖는다. 다중 모델 기계학습 기법은 동일 훈련예 집합에서 유도된 다수 모델의 예측를 통합하여 안정성 있는 클라스 예측치를 제시하는 전략이다.본 연구에서는 다중모델 기계학습 기법의 개념을 응용한 하이브리드 형태의 다중모델 학습법을 제안하고 문서집합에 대한 분류 실험을 통해 하이브리드 다중 모델 기법으로 유도된 다중 모델 분류기와 일반 다중 모델 분류기, 그리고 단일 모델을 이용한 분류기의 성능을 비교하였다

2. 문서 모델과 기계학습

기계학습 알고리즘을 적용하기 위해 문서는 특징과 특징 값의 벡터로 표현해야 하며 문서의 특징은 문서 내용을 대표할 수 있는 주요 키워드를 사용한다. 문서는 텍스트의 태그 제거, 불용어 처리, 어미 변화 처리 등의 텍스트 전처리와 특징 추출(feature extraction)을 거쳐 문서의 내용을 대표하는 특징 부분 집합의 벡터로 표현하게 된다. 벡터는 특징 값에 따라 이진벡터(binary vector) 또는 가중치 벡터(weighted vector)로 표현된다. 특징 선택을 위한 방법으로는 일정한 척도(measure)로 특징들을 개별적으로 평가하여 필요한 만큼의 특징을 선택하는 여과방법(filtering)과 내포된 특정 분류기의 성능을 높일 수 있는 특징들의 부분집합을 점진적으로 구해 가는 포장방법(wrapper), 두 가지 접근 방식이 있는데 일반적으로 텍스트 연구에서는 여과 방식이 많이 사용된다[2]. 주요 여과방법으로는 키워드의 문서 내외 발생빈도에 근거한 TFIDF, 특정 특징이 분류된 데이터의 순도에 기여하는 정도를 측정하는 정보 획득(information gain), 특징의 클라스 분포에 기초한 상관성을 나타내는 χ^2- test, 키워드 간의 유사도와 연관성에 의거 추출하는 LSI(Latent Semantic Indexing) 등이 있다. 기계학습법 가운데 문서 분류에 많이 적용되어 온 학습법으로 Naïve Bayesian, k-NN(k Nearest Neighbor), 결정 트리(Decision Tree)를 들 수 있다. Naïve Bayesian방식은 분류 대상 문서가 각 클라스에 속할 조건부 확률을 계산하고 이중 가장 높은 조건부 확률을 갖는 클라스로 분류한다. k-NN 방식은 훈련 예들에 대한 사전 학습을 하지 않고 분류대상 문서가 주어지면 각 훈련문서와의 거리를 계산하여 가장 가까운 k개의 이웃문서를 선택하고 이들의 분류 클라스에 따라 대상문서의 소속 클라스를 정한다. 결정트리는 문서분류를 위한 지식을 트리구조로 표현하는 것인데, 비 단말 노드는 하나의 문서특징을 나타내고, 한 노드에서 분기하는 각 가지는 하나의 특징 값을 표시하며, 단말노드는 하나의 분류 클라스를 나타낸다. 분기의 정점이 되는 각 노드의 선택은 매 분기 시 문서집합의 엔트로피(entropy)를 낮추는데 각 특징이 기여하는 정도를 계산한 정보 획득량(Information Gain, IG)에 따라 결정한다[3].

3. 다중모델 학습 기법

다중 모델 학습 전략은 학습에의 경미한 변화에도 학

습된 분류기가 민감한 성능의 변화를 보이는 불안정성(unstability)을 해소하거나 문제 도메인에 적용된 학습 알고리즘 간의 성능 편차를 상쇄하려는 것이다. 이론적으로는 훈련예의 분산(variance)효과와 기계 학습알고리즘 고유의 편향(bias)에서 비롯되는 오차를 축소하는 것이다. 분산은 훈련예가 실세계의 분포를 반영하기 어려운 한계에서 오는 오차의 원인이며, 편향은 기계학습 고유의 한계에 기인하는 오차율이다. 편향의 원인에는 학습 알고리즘 고유의 지식 및 개념 표현 방식, 탐색 편향 등이 있다. 대표적인 다중모델 기법으로 Bagging[4]과 Boosting은 분산에 기인한 오차를 축소하려는 전략이며, Stacking은 편향의 효과를 상쇄하는 방식이다.

이들 다중학습 기법은 첫째, 모델 생성과정과 둘째, 다중 모델의 분류 예측치를 취합하여 결론을 내는 방식에서 근본적인 차이를 보인다. 다중 모델은 동일 학습 알고리즘에 훈련예를 달리하여 생성한 동질적(homogeneous) 모델과, 다수의 알고리즘에 의한 이질적(heterogeneous) 모델로 구분할 수 있는데, Bagging과 Boosting이 동질적 모델을 다수결로 취합하며 Stacking은 이질적 모델의 분류 예측을 재학습하여 지능적으로 통합한다.

Bagging과 Boosting은 기본 방식은 유사하나 다중 모델을 유도할 훈련예 샘플링 방식과 최종 분류결정 과정이 상이하다. Bagging은 임의의 횟수 만큼 중복을 허용하며 무작위로 부분집합을 추출하고, 다중 모델의 분류 결과를 단순 다수결로 취합하여 클라스를 결정한다.

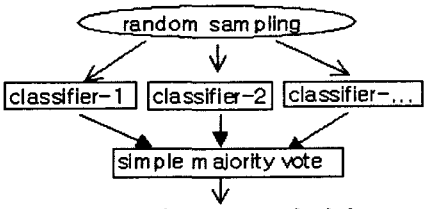


그림1 Bagging의 개념도

Boosting은 모델 생성 과정에서 매회의 분류 결과에 따라 오분류된 데이터의 가중치 분포를 달리하면서 훈련 데이터를 추출, 순차적으로 모델을 생성하는 적응형 방식이다. 신규 사례는 각 모델의 가중치를 적용한 다수결(weighted sum)에 의해 최종 클라스를 결정한다.

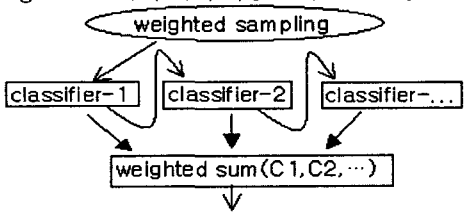


그림 2 Boosting의 개념도

Bagging과 Boosting이 단일 학습 알고리즘의 적용 결과를 다수결로 결정하는 투표 (vote) 알고리즘 방식인데 반하여 Stacking은 다수 알고리즘의 예측 결과를 토대로 상위 단계에서 최종 학습하는 대표적인 메타 학습 알고리즘이다. Stacking은 이질적인 다수 모델을 병합하는 전략으로서 분류알고리즘의 성능을 측정하는데 이용되는 교차검증(cross validation)보다 유연하고 정교한 기법이다. Stacking방식은 기반 단계에서 각 데이터에 대해 클라스 확률 분포를 계산하고 이들 예측치를 결합하여 메타 학습기 생성을 위한 메타 데이터로 사용한다. 신규 사례는 두 단계의 분류를 거쳐 최종 클라스가 제시된다.

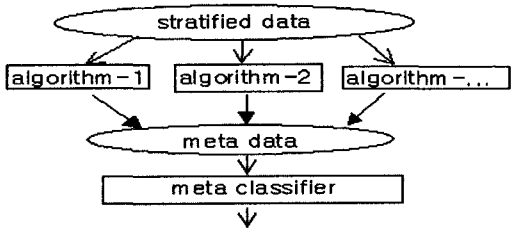


그림3 Stacking의 개념도

4. 하이브리드 다중 모델 학습

위의 기법과 이론에 기반한 여러 가지 파생적 다중모델 학습 알고리즘이 연구되고 있다[5]. 일반적으로 투표 알고리즘의 파생기법은 데이터의 분포의 변화를 통해 분산효과를, Stacking의 파생 기법은 학습 알고리즘의 편향을 해소하려는 접근법이다. 본 연구에서는 확장형 Bagging 및 Boosting과 확장형 Stacking을 제안하였다. 확장형 투표알고리즘은동질적 다중 모델의 통합에 메타 학습을 사용하여 결론 제시에 보다 지능적으로 접근하였다. Stacking의 확장형인 Bagged Stacking과 Boosted Stacking 역시 Bagging과 Boosting 방식으로 모델을 생성하되 예측 결과를 학습하고 메타 분류기를 생성하여 최종 클라스를 결정하는 것이다. 이 방식은 이질적인 학습알고리즘의 편향을 상쇄하는 동시에 훈련예의 변화에 따른 분산의 교정 효과가 있을 것으로 기대하였다.

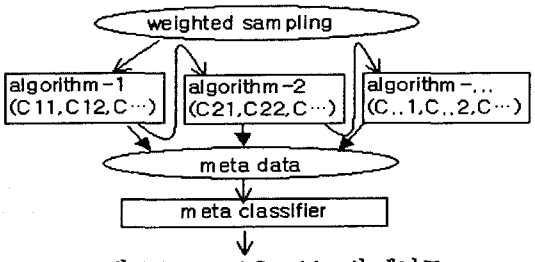


그림 4 Boosted Stacking의 개념도

5. 실험

문서 분류 실험을 통하여 단일모델과 다중모델, 하이브리드 다중 모델기법의 분류기 성능을 비교하고 자 하였다. 특징 선택은 문서 분류에 성능이 우수한 정보 획득(IG) 방식을 채택하고 특징 집합의 크기에 따른 성능 차이를 검토하고자 하였다. 문서 모델은 이진벡터와 가중치 벡터 두 가지로 표현하여 특징 값 표현 방식의 효과를 분석하고자 하였다. 학습 알고리즘은 단일 모델에 결정트리(C4.5), Naïve Bayesian, k-NN을, 투표 알고리즘에 C4.5를, Stacking에는 기반 학습에 C4.5와 Naïve Bayesian을 공통으로 사용하고, 메타 학습에 Decision Stump, C4.5, Naïve Bayesian을 적용하였다. 실험데이터는 MEDLINE 학술 기사 6000건(6개 분류 항목), 텍

스트 분류용으로 공개된 유즈넷 뉴스 5000건(5개)과 웹 문서 4518건(6개)을 사용하였다. 세 종류의 문서 집합 가운데 MEDLINE의 경우 10등분 교차 검증을 거친 분류 정확도는 다음과 같다.

종류	학습기법	50-bin	50-wt	100-bin	100-wt
단일 모델	K-NN	63.38%	63.43%	68.14%	61.62%
	C4.5	69.67%	70.58%	68.24%	67.89%
	NB	76.91%	74.11%	76.78%	74.12%
다중 모델 (투표)	BaggingC4.5	72.51%	69.80%	70.49%	70.59%
	BoostingC4.5	68.48%	67.76%	67.99%	68.09%
다중 모델 (Stacking)	StackingDS	33.18%	30.78%	31.27%	30.25%
	StackingC4.5	76.48%	77.34%	78.07%	78.07%
	StackingNB	76.78%	76.42%	78.92%	78.48%
하이브리드 다중 모델(투표)	ExtBagC4.5	75.58%	75.65%	75.95%	75.67%
	ExtBoostC4.5	75.50%	75.03%	75.48%	75.38%
하이브리드 다중 모델(Stacking)	BagStackC4.5	75.98%	74.20%	76.48%	75.35%
	BoostStackC4.5	76.80%	77.17%	77.68%	76.35%

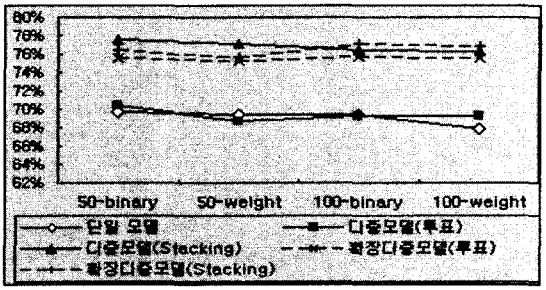
표 1 MEDLINE 문서 집합에 대한 분류 정확도

본 실험에서는 문서 모델의 이진 표현법이 전반적으로 가중치 표현법보다 우수한 성능을 보였다. 문서 특징 부분집합 50 내지 250개 사이에서 성능 차이는 미미하였다. 문서집합 별 분류 정확도는 큰 편차를 보였는데 정확도가 높은 유즈넷 뉴스의 경우 주제 관련성이 높은 컴퓨터 전문용어가 많이 분포된 것이 원인으로 판단된다. 성능이 낮은 Decision Stump(StackingDS)를 메타 학습기로 사용한 경우 정확도가 비현실적으로 낮은데 메타 학습기도 우수한 알고리즘을 이용하는 것이 분류효과에 유리함을 보여준다.위와 같은 실험결과를 종류별로 평균하여(StackDS 제외) 정확도를 그라프로 나타내면 다음과 같다.

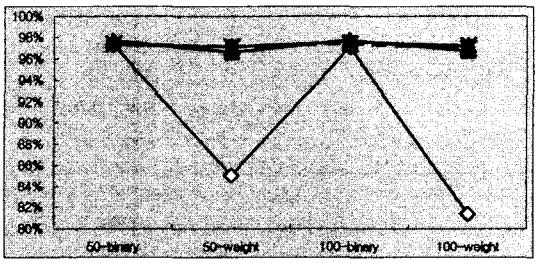
대부분 문서집단에서 하이브리드 학습기의 분류 성능이 단일 모델과 기존의 다중 모델을 능가하였다. 전체적으로 단일 모델 78.36%, 다중 모델 83.55%, 하이브리드 다중 모델이 84.30%의 평균정확도를 보였으며 이러한 결과는 분산과 학습알고리즘의 편향 효과를 동시에 해소 내지 완화가 가능함을 보여준다. 특히 메타학습 기법의 다중 모델 통합 효과가 두드러졌는데 동질 모델과 이질 모델 모두의 통합효과가 높은 것으로 나타났다. 효율적인 통합방법은 동질적 다중모델의 성능 향상에 기여하며, Stacking 종류의 이질적 모델의 우수한 결과는 학습 알고리즘의 편향이 학습 데이터의 분산보다 오차에 미치는 영향이 크기 때문인 것으로 판단된다. 또한 하이브리드 분류기의 결과는 매우 낮은 편차를 보이는데 (단일 0.066, 다중 0.023, 하이브리드 0.01) 이는 하이브리드 분류기의 안정성을 나타내는 것이다.

6. 결론

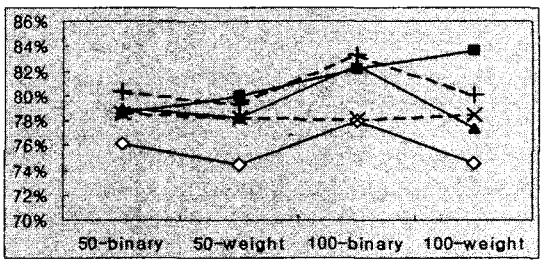
본 논문에서는 다중 모델 기계학습법의 개념과 특징을 고찰하고, 분류 오차의 다양한 원인을 동시에 해소하고자 하는 하이브리드 다중 모델 학습법을 제안하고 텍스트 자동 분류에 적용, 실험하였다. 실험 결과 전반적으로 메타 학습법을 이용한 하이브리드 다중모델기법이 기존의 다중 모델 학습기법들 보다 높은 성능을 보임으로써 문서 분류 도메인에서 그 효과를 입증하였다. 하이브리드 다중 모델 학습은 높은 계산 비용을 요하지만 성능의 향상과 신뢰도의 확보가 중요시 되는 응용분야에서는 그 비용이 정당화될 것으로 판단된다.



(a) MEDLINE 문서 집합



(b)유즈넷 뉴스 문서 집합



(c) 웹 문서 집합

그림 5 문서 집합별 분류 정확도

참고 문헌

[1]Chen, H., " Machine learning for information retrieval." JASIS,Vol.46, No.3, pp.194-216, 1995.

[2]Han, Jiawei and Kamber, M., Data Mining. Kaufmann, New York, 2001.

[3]Mitchell, Tom, Machine Learning, McGraw-Hill, New York, 1997.

[4]Breiman, Leo., " Bagging predictors," Machine Learning, Vol.24, No.1, pp. 49-64, 1996

[5]Bauer, Eric and Kohavi, Ron,, " An empirical comparison of voting classification algorithms," Machine Learning, Vol.36, pp.105-142, 2000.