불균형 데이터셋을 활용한 선택적 분류 알고리즘 성능 평가

정경희, 이준영, 한동석*
경북대학교
ghjung@knu.ac.kr, ddr37105@knu.ac.kr, *dshan@knu.ac.kr

Selective Classification under Extreme Class Imbalance

Jung Gyeong Hee, Jun Yeong Lee, Dong Seog Han*
Kyungpook National Univ.

요 약

선택적 분류는 모델이 불확실한 입력에 대해 예측을 거부할 수 있도록 설계된 방법으로, 안전성과 정확성이 동시에 요구되는 응용 분야에서 주목받고 있다. 특히 클래스 불균형이 심한 환경에서는 예측 신뢰도에 기반한 선택 전략이 필수적이다. 하지만 기존 연구들은 SelectiveNet, Deep Gambler, SAT(Self-Adaptive Training), TBS(Toward Better Selection) 등 다양한 선택적 분류 알고리즘의 성능을 동일 조건에서의 체계적 비교가 부족하며, 실제 불균형 데이터셋을 활용한 평가도 부족하다. 본 연구에서는 극단적인 클래스 불균형을 갖는 WM811k 데이터셋을 기반으로, 네 가지 선택적 분류 알고리즘을 동일한 실험 설정 하에 비교하고, 커버리지에 따른 선택 정확도와 거부 전략의 특성을 분석한다. 실험 결과는 SelectiveNet이 전체적으로 가장 안정적인 선택 정확도를 보였으며, Deep Gambler는 낮은 커버리지 조건에서의 리스크 회피 능력이 우수했다. SAT와 TBS는 불확실한 샘플 처리와 클래스 선택 균형 측면에서 의미 있는 대안을 제시하였다. 본 연구는 선택적 분류 기법의 실제 적용 가능성을 정량적으로 검증하며, 본 연구는 선택 전략의 실용적 평가 지표를 제안하고, 향후 동적 커버리지 제어나 멀티태스크 응용으로의 확장을 위한 기반을 제공한다.

Ⅰ. 서 론

선택적 분류는 모델이 불확실한 입력에 대해 예측을 보류함으로써 전체 오류를 줄이고, 특히 고위험 환경에서의 신뢰성을 높이는 전략으로 주목받고 있다.[1] 초기 연구에서는 확신 점수가 낮은 경우 거부 클래스로 분류하는 단순 임계값 기반 전략이 사용되었다..[2]

최근에는 학습 가능한 선택 함수나 보류 확률을 별도로 예측하는 모듈화된 구조가 제안되며, 선택 기준의 정밀도와 적응성이 향상되고 있다. 이러한 접근은 크게 보류 클래스를 명시적으로 추가하는 방식과, 선택 헤드를 활용하는 방식으로 나눌 수 있다.[3][4][5][6]

하지만 이들 기법은 대부분 클래스 간 불균형이 적고 이상적인 구조를 가진 비교적 단순한 데이터셋에서만 검증되었으며, 극단적인 클래스 불균형을 갖는 실제 데이터셋에서의 비교는 충분히 이루어지지 않았다.

본 논문에서는 WM811k 데이터셋을 기반으로, SelectiveNet, Deep Gambler, SAT, TBS 네가지 대표적인 선택적 분류 알고리즘을 동일한 실험 조건 하에 비교한다. 커버리지 변화에 따른 선택 정확도와 거부율 등의 주요 성능 지표를 바탕으로 실제 적용 가능성과 한계을 평가하고자 한다. 이와 같은 현실적이고 도전적인 조건은 선택적 분류 알고리즘의 성능을 검증하는 데 적합한 실험 환경을 제공한다.

Ⅱ. 본 론

2.1 선택적 분류 학습 알고리즘

커버리지는 선택적 분류에서 모델이 전체 입력 중 어느 정도 비율에 대해 예측을 수행할지를 정의하는 핵심 지표로, 예측의 신뢰도와 보류율 간의 트레이드오프를 설정하는 기준이 된다. 높은 커버리지는 더 많은 입력에 대해 예측을 시도하되 위험도도 높아지고, 낮은 커버리지는 보다 보수적인 예측을 통해 선택 정확도를 향상시킬 수 있다 . SelectiveNet, Deep Gambler, SAT, 그리고 TBS과 같은 주요 알고리즘은 각기 다른 방식으로 커버리지를 해석하거나 제약하며, 모델이 신뢰할 수 있는 샘플만을 선택적으로 예측할 수 있도록 설계되어 있다.

대부분의 선택적 분류 모델은 SelectiveNet의 구조를 기반으로 설명할 수 있다. SelectiveNet은 예측 헤드와 별도로 선택 헤드를 함께 학습하며, 선택 여부를 결정하는 컨피던스 값을 통해 정해진 커버리지 비율 내에서 신뢰도 높은 샘플만 예측한다. 이 구조는 선택 점수가 명시적으로 출력되며, 커버리지 제약이 손실 함수에 직접 포함되는 것이 특징이다.

반면, Deep Gambler는 선택 헤드를 따로 학습하지 않고, 클래스 예측 로짓 외에 보류 로짓을 추가하여 이를 통해 예측 대신 보류를 선택하는 구조를 사용한다. 모델은 클래스 로짓과 보류 로짓 간의 유틸리티를 최적화하면서 학습되며, 선택 구조가 확률적으로 통합된 형태다.

SAT 또한 보류 로짓을 도입하지만, 그 활용 방식은 다르다. SAT는 예측 분포의 엔트로피를 활용해 불확실한 샘플에 낮은 가중치를 부여하고, 학습 중 손실의 영향을 줄인다. 즉, SAT는 보류를 학습 중 안정화 도구로 활용하며, 선택 구조를 명시적으로 도입하지는 않지만, 결과적으로 선택적 동작을 유도하는 방식이라 할 수 있다.

마지막으로 TBS는 별도의 보류 로짓 없이 소프트맥스 확률 자체를 확신 임계값으로 사용해 선택 기준을 구성한다. 이는 Deep Gambler나 SAT와 달리 보류 로짓이나 엔트로피 기반 가중치 조정 없이도 간결한 방식으로 선택 구조를 형성하며, 선택 기준의 해석 가능성과 구현 효율성을 높인다. 특히 TBS는 SAT 대비 커버리지 정확도 예측의 불안정성을 줄인다.

표 I. SN과 TBS의 클래스별 선택 정확도 비교 (커버리지 = 0.80)

Class	SN Val Acc (%)	TBS Val Acc (%)	Δ Accuracy	Number of Data
Center	27.16	**97.40**	+70.24	346
Donut	69.18	**78.05**	+8.87	41
Edge-Loc	27.71	**70.66**	+42.95	242
Edge-Ring	39.34	**76.17**	+36.83	856
Loc	9.43	**84.57**	+75.14	162
Near-full	**97.89**	80.00	–17.89	5
Random	80.54	**91.80**	+11.26	61
Scratch	1.30	**20.00**	+18.70	50
None (Normal)	50.29	**100.00**	+49.71	110701
Total Accuracy	50.29	**93.30**	+43.01	112464

2.2 실험 구성 요소

실험에서 사용된 웨이퍼 결함 검사 데이터셋은 총 약 81만 장의 웨이퍼 이미지로 구성되어 있으며, 정상 클래스가 전체의 약 95%를 차지하는 극단적인 클래스 불균형을 보인다. 5개 이상 결함 클래스는 각각 1% 미만의 비율로 분포되어 있어, 선택적 분류의 성능을 평가하기에 적합한 환경을 제공한다.

별도의 선택 모듈의 유무가 선택 구조의 불균형 대응 성능을 비교하기 위해, 먼저 SelectiveNet과 TBS 알고리즘을 동일한 커버리지에서 클래스별로 정확도를 비교한다. 그리고 본 실험에서는 선택 모듈과 학습 모듈이 통합된 알고리즘들의 불균형 대응 성능을 다양한 커버리지에서의 검증 정확도를 바탕으로 비교한다.

모든 알고리즘은 동일한 VGG-16 백본을 사용하였으며, 학습률 1e-3, 배치 크기 128, 옵티마이저는 Adam으로 설정하고 총 100 epoch 동안 학습을 진행하였다. 입력 전처리 및 분류 구조는 통일하되, 각 알고리즘 고유의 선택 구조와 손실 함수만을 달리 적용하였다.

2.3 실험 결과

SelectiveNet의 클래스별 선택 성능을 분석한 결과[표 1], 선택이 정상 클래스에 집중되는 경향이 나타났으며, 소수 클래스의 경우 선택 정확도 저하 또는 선택 자체가 이루어지지 않는 현상이 관찰되었다. 이는 선택 헤드가 다수 클래스에 과적합되기 쉬운 구조적 한계를 시사하며, 확신 임계값 기반으로 선택 기준을 형성하는 TBS와 비교했을 때 이러한 편향이 일부 완화됨을 확인할 수 있었다.

추가 실험에서는 SAT, Deep Gambler, TBS에 대해 커버리지 변화에 따른 선택 정확도를 측정하였다[표 2]. Deep Gambler는 커버리지 0.7에서 최고 성능을 보인 반면 이후 급격히 하락하였다. SAT와 TBS는 중간 커버리지에서 비교적 안정적인 정확도를 보였으며, TBS는 간결한 구조에도 불구하고 전체적으로 가장 균형 잡힌 성능을 나타냈다.

Ⅲ. 결 론

본 논문에서는 기존의 학습 보류에 사용되던 알고리즘이 클래스 불균형 상황에서도 강인한지를 확인하기 위해 대표적인 네 가지 선택적 분류 알고리즘의 성능을 확인하였다. 향후에는 커버리지를 동적으로 조절하거나 선택 구조를 멀티태스크 학습과 결합할 예정이다.

표 II. 커버리지 변화에 따른 선택적 분류 성능 비교

Coverage	Gambler	SAT	TBS
1.00	90.38	93.83	91.12
0.90	87.95	95.42	92.34
0.80	83.40	96.75	93.27
0.70	**99.70**	97.63	93.83
0.60	85.25	97.85	93.41
0.50	78.90	98.12	93.24
0.40	71.42	98.35	93.06
0.30	65.10	98.68	92.88
0.20	59.34	99.12	92.63
0.10	52.87	99.55	92.31

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학ICT연구센터(ITRC)의 지원을 받아 수행된 연구임 (IITP-2025-RS-2020-II201808)

참 고 문 헌

[1] E. El-Yaniv and A. Wiener, “On the foundations of noise-free selective classification,” JMLR, vol. 11, pp. 1605 - 1641, 2010.

[2] C. Cortes, M. Mohri, and A. Rostamizadeh, “Boosting with abstention,” in Proc. NeurIPS, 2016, pp. 1660 - 1668.

[3] Y. Geifman and R. El-Yaniv, “SelectiveNet: A deep neural network with an integrated reject option,” in Proc. ICML, 2019, pp. 2151 - 2160.

[4] A. D. Rakhlin, D. Rakhlin, and X. Liu, “Deep Gambler: Learning to abstain with neural networks,” in Proc. NeurIPS, 2018.

[5] Z. Huang, Z. Xu, B. Gong, and S. Wang, “Self-adaptive training: Beyond empirical risk minimization,” in Proc. NeurIPS, 2020.

[6] H. Moon, D. Yun, and H. Kim, “Toward better selection for classification with rejection,” in Proc. ICCV, 2023.