

지도학습 기반 암상 분류 시 클래스 간 자료 불균형을 고려한 평가지표 개발

김도완 · 최준환 · 변중무*

한양대학교 자원환경공학과

Development of Evaluation Metrics that Consider Data Imbalance between Classes in Facies Classification

Dowan Kim, Junhwan Choi, and Joongmoo Byun*

Dept. of Earth Resources and Environmental Engineering, Hanyang Univ.

요약: 머신러닝을 이용한 분류 모델 훈련에서 학습자료의 양과 질은 학습한 모델의 성능을 좌우하므로 학습자료 생성이 매우 중요한 역할을 한다. 그러나 자료 생성에 높은 비용이 들어 이상적인 학습자료 생성이 어려울 때에는 클래스 간 자료 불균형 문제가 발생한다. 만약 학습자료로 사용될 탐사자료가 클래스 간 불균형하게 얻어지면, 클래스 별로 균형을 이루는 학습이 이루어지기 힘들다. 따라서 데이터가 상대적으로 적은 클래스는 재현율이 현저히 떨어지게 된다. 그 뿐만 아니라 정확도와 정밀도 등의 평가지표들에 대한 신뢰도가 떨어지게 된다. 따라서 이 연구에서는 두 단계에 걸쳐 자료 불균형 문제를 해소하고자 하였다. 첫 번째로 기존의 정확도와 정밀도를 개선하여 자료 불균형을 고려할 수 있는 새로운 평가지표로 가중정확도와 가중정밀도를 고안하였다. 다음으로 클래스 간의 가중정밀도와 재현율의 균형을 맞추어 주도록 오버샘플링을 수행하였다. 개발한 알고리즘을 물리검층 자료를 이용한 암상 및 공극유체 규명 문제에 적용함으로써 검증하였다. 그 결과 다수 클래스와 소수 클래스들 간의 불균형이 상당 부분 완화되었고, 클래스 간의 경계를 보다 명확하게 확인할 수 있었다.

주요어: 머신러닝, 자료 불균형, 평가지표, 오버샘플링, 암상 분류

Abstract: In training a classification model using machine learning, the acquisition of training data is a very important stage, because the amount and quality of the training data greatly influence the model performance. However, when the cost of obtaining data is so high that it is difficult to build ideal training data, the number of samples for each class may be acquired very differently, and a serious data-imbalance problem can occur. If such a problem occurs in the training data, all classes are not trained equally, and classes containing relatively few data will have significantly lower recall values. Additionally, the reliability of evaluation indices such as accuracy and precision will be reduced. Therefore, this study sought to overcome the problem of data imbalance in two stages. First, we introduced weighted accuracy and weighted precision as new evaluation indices that can take into account a data-imbalance ratio by modifying conventional measures of accuracy and precision. Next, oversampling was performed to balance weighted precision and recall among classes. We verified the algorithm by applying it to the problem of facies classification. As a result, the imbalance between majority and minority classes was greatly mitigated, and the boundaries between classes could be more clearly identified.

Keywords: machine learning, imbalance problem, evaluation metrics, oversampling, facies classification

서론

탄성과 자료 해석에서 암상 및 공극유체 분류 작업은 저류층을 암석의 종류와 공극 내 유체의 종류에 따라 목적에 맞게 분류하는 작업을 말한다. 암상 분류 결과는 관심 영역의 퇴적 환경을 이해하는 데 도움을 주기 때문에 탐사단계에서 저류층 특성화의 일환으로써 중요한 역할을 담당한다. 나아가 탄화수소 매장량을 추정하고 경제성 있는 퇴적분지를 선정하는 데 사용될 수 있다.

Received: 10 June 2020; Revised: 24 August 2020;

Accepted: 26 August 2020

*Corresponding author

E-mail: jbyun@hanyang.ac.kr

Address: Dept. of Earth Resources and Environmental Engineering, Hanyang University, 222 Wangsimni-ro, Seongdong-gu, Seoul, Korea

©2020, Korean Society of Earth and Exploration Geophysicists

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

암상 및 공극유체의 특성은 탄성과 탐사자료에서 진폭의 크기 혹은 진폭의 변화로부터 추정했다(Gao, 2003; Chopra and Marfurt, 2005; Yenugu *et al.*, 2010). 지층 정보를 가지고 있는 검층자료가 존재하는 경우, 검층자료로부터 계산된 탄성물성들을 함께 활용하면 좋은 고해상도의 참고자료가 될 수 있다. 검층자료를 이용하는 경우에는 탄성과 탐사자료와 검층자료의 해석상 연결고리를 구축하는 작업이 필요하다. 이를 위하여 AVO (Amplitude Variation with Offset) 역산 혹은 임피던스 역산 등을 통해 계산된 탄성물성 값이 이용될 수 있다 (Goodway *et al.*, 1997; Chi and Han, 2009; Nieto *et al.*, 2013; Pendrel *et al.*, 2017; Choi *et al.*, 2019).

최근에는 머신러닝의 많은 발전으로 인하여 암상 및 공극유체 분류에도 머신러닝을 접목한 연구가 활발히 진행되고 있다. Jin (2018)은 암상 분류 문제에 다양한 머신러닝 기법을 적용하여 성능을 비교하였으며, Naeini (2019)은 검층자료를 이용하여 딥러닝 모델을 학습시킨 후 스위트 스팟(sweet spot)을 예측하였다. 나아가 Saleem *et al.* (2019)은 레이블이 있는 검층자료 뿐만 아니라 시추공 주변 자료까지 학습에 이용하는 준지도학습 기법을 적용한 바도 있다.

하지만 암상 및 공극유체 규명 문제에 머신러닝을 적용하는 경우, 제한된 환경에서 취득된 검층자료를 활용하기 때문에 이상적인 학습자료를 구축하기 어렵다는 문제점을 가지고 있다. 특히 지역마다 암상 및 공극유체에 따른 분포 비율이 다르게 되는데 이러한 경우 심각한 자료 불균형 문제가 발생할 수 있다. 머신러닝 분야에서 자료 불균형 문제는 학습자료 내 클래스에 따른 자료의 비율이 상이한 경우 발생한다. 예를 들어 석유 및 가스전 탐사 시, 공극 내 탄화수소를 함유한 영역에서 취득된 검층자료보다 탄화수소를 함유하지 않은 영역에서 취득되는 검층자료의 양이 더 많은 경우가 이에 해당된다. 이러한 경우 자료가 상대적으로 풍부한 클래스, 즉 탄화수소가 존재하지 않는 경우의 클래스로 과하게 치중된 모델이 학습될 수 있다.

학습자료 내에 불균형 문제를 완화하기 위해 학습자료를 개선시키는 전처리 기법이 연구되었다. 전처리 기법은 크게 언더샘플링 기법과 오버샘플링 기법으로 구분된다. 언더샘플링 기법은 자료의 개수가 상대적으로 많은 클래스(다수 클래스, majority class)에서 일부 샘플만 취득해 사용하는 방법이다. 하지만 이 방법은 자료를 일부 취득하는 과정에서 중요한 정보를 잃어버릴 수 있어, 학습자료가 부족한 환경일수록 적용하기 어려운 기법이다. 반면 오버샘플링 기법은 언더샘플링과는 반대로 자료의 양이 적은 클래스(소수 클래스, minority class)에 가상 자료를 추가함으로써 클래스 간의 균형을 맞추어주는 방법이다.

이에 Kim *et al.* (2019)는 오버샘플링 기법을 적용하여 암상 및 공극유체 규명 시의 자료 불균형 문제를 완화하고자 하였다. 오버샘플링 기법을 적용하는 경우에는 합성자료의 개수

를 적정 수준으로 결정하는 작업이 반드시 동반되어야 한다. 만약 너무 적은 자료를 합성하게 되면 불균형 문제를 충분히 해결하기 어렵고, 반대로 너무 많은 자료를 합성하게 되면 오히려 소수 클래스에 대한 과적합이 발생하게 된다(He and Garcia, 2009). 일반적으로 머신러닝 분야에서 사용되는 평가 지표인 정확도(accuracy), 정밀도(precision) 등은 자료 불균형 정보가 고려되지 않아 클래스 간의 균형 여부를 모니터링하기에 적합하지 않다. 따라서 이 연구에서는 자료 불균형을 고려한 평가지표를 제안하고, 제안된 평가지표를 모니터링함으로써 적정 수준의 합성자료를 생성하고자 하였다. 또한 합성자료를 생성하여 오버샘플링을 수행한 후 암상 및 공극유체 분류를 수행하여 분류 결과의 정확도 개선 여부를 비교 분석하고자 하였다.

오버샘플링 기법 리뷰

무작위 오버샘플링은 가장 간단하게 적용할 수 있는 오버샘플링 기법이다. 해당 방법은 기존의 소수 클래스 내 샘플들을 학습에 한 번만 사용하는 것이 아니라 수차례 반복적으로 무작위 추출하여 사용한다.

$$T = N + P, \quad (1)$$

$$P' = P + E, \quad (2)$$

$$T' = N + P + E, \quad (3)$$

여기서 T 는 전체 자료, N 과 P 는 각각 다수 클래스와 소수 클래스를 의미한다. 또한 P' 는 오버샘플링 후의 소수 클래스를 의미하고, E 는 소수 클래스에서 무작위로 추출된 샘플들을 의미한다. 따라서 새로운 학습자료 T' 는 오버샘플링된 소수 클래스 P' 와 다수 클래스 N 의 합으로 새롭게 정의된다. 무작위 오버샘플링은 소수 클래스에 대하여 학습 시 손실 함수에 가중치를 주는 것과 비슷한 효과를 얻게 되는데, 이는 소수 클래스에 대해서 쉽게 과적합 될 수 있음을 의미한다.

이러한 문제를 완화하고자 Chawla *et al.* (2002)은 단순히 기존 자료를 복제하는 무작위 오버샘플링 기법과는 다르게 기존 자료들을 합성하여 새로운 자료를 생성하는 SMOTE (Synthetic Minority Oversampling Technique) 기법을 제안하였다.

$$synthetic_j = p_i + r_j dif_j, \quad j = 1, 2, \dots, s, \quad (0 \leq s \leq k) \quad (4)$$

여기서 p_i 는 소수 클래스에서 무작위로 추출된 샘플을 의미하며 $synthetic_j$ 는 p_i 를 이용하여 새롭게 합성된 자료를 의미한다. 다음으로 소수 클래스 내에서 p_i 를 중심으로 k 개의 최근접 이웃점을 계산하였을 때, dif 는 p_i 와 이웃점들 사이의 거리를 의미한다. 이 때 s 는 0 이상 k 이하의 무작위 정수로 결정된다. 마지막으로 r_j 는 0과 1 사이의 난수로써 합성자료

를 새롭게 만들 때마다 새롭게 갱신된다.

SMOTE 기법은 기존의 샘플을 그대로 사용하지 않고 샘플들의 조합으로 새로운 자료를 합성한다는 점에서 과적합을 완화할 수 있지만, p_i 가 단순히 무작위로 추출된다는 점에서 두 가지 문제점이 발생할 수 있다. 먼저 선택된 p_i 가 잡음에 해당하는 경우 잡음의 영향만 오히려 강조될 수 있다. 반대로 p_i 가 클래스 간 경계부로부터 멀리 떨어진 위치에 분포하는 경우, 합성자료는 계산량만 증대시킬 뿐 클래스의 경계를 학습시키는데 크게 도움이 되지 않는다.

따라서 Han *et al.* (2005)은 SMOTE를 개선한 Borderline-SMOTE를 제안하였다. Borderline-SMOTE는 p_i 에 대해서 SMOTE를 적용하기에 앞서 전체 영역 T 에서 임의로 결정된 m 개의 최근접 이웃점들을 탐색한다. 이후 소수 클래스를 세 개의 그룹으로 세분화한다. 만약 m 개의 이웃점들이 모두 다수 클래스에 속하는 경우 잡음 영역, 과반수 이상 m 개 미만의 이웃점들이 다수 클래스에 속하는 경우 위험 영역, 과반수 미만의 이웃점만이 다수 클래스에 속하는 경우 안전 영역으로 정의한다. 그 중 위험 영역에 속한 p_i 에 대해서만 SMOTE를 적용하는 것이 Borderline-SMOTE이다.

Fig. 1은 클래스 간 불균형 문제가 있는 임의의 자료에 SMOTE 기법과 Borderline-SMOTE를 적용하였을 때 생성된 합성자료의 비교이다. 이 때 k 와 m 은 각각 3, 4로 설정하였다. SMOTE를 적용한 경우(Fig. 1a) 잡음에 영향을 받은 합성자료가 1번 영역에서 생성된 것을 볼 수 있다. 또한 다수 클래스와의 경계부에서 멀리 떨어진 2번 영역에서도 다수의 합성자료가 생성되었다. 반면 Borderline-SMOTE를 적용한 경우인 Fig. 1b를 보면 1번 영역에서 합성자료가 전혀 생성되지 않았으며 2번 영역에서도 합성자료가 거의 만들어 지지 않은 것을 확인할 수 있다. 따라서 이 연구에서는 Borderline-SMOTE를 적용하여 학습자료 내 자료 불균형 문제를 해소하고자 하였다.

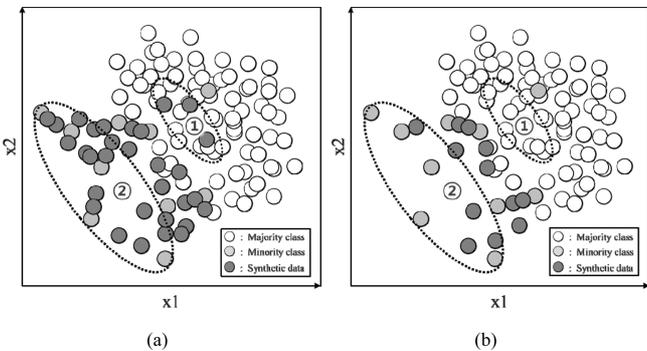


Fig. 1. (a) Synthetic data produced by applying the SMOTE (Synthetic minority oversampling technique). Some synthetic data in Area 1 are produced by noise, and the synthetic data in Area 2 are not conducive to training a boundary. (b) Synthetic data formed using the Borderline-SMOTE technique.

클래스 간 불균형을 고려한 평가지표 개발

머신러닝을 이용한 분류 모델의 성능을 평가하기 위해 가장 보편적으로 사용하는 평가지표에는 정확도(accuracy), 정밀도(precision), 재현율(recall) 등이 있다. Table 1은 두 개의 클래스를 분류할 때의 오차행렬(confusion matrix)이다. 이 때 정확도, 정밀도, 재현율은 다음과 같다.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}, \quad (5)$$

$$Precision = \frac{TP}{TP+FP}, \quad (6)$$

$$Recall = \frac{TP}{TP+FN}. \quad (7)$$

즉, 정확도는 전체 자료 중 올바르게 예측한 비율을 나타내고, 정밀도는 특정 클래스로 예측한 자료 중 올바르게 예측한 비율, 재현율은 실제 특정 클래스에 속한 자료 중 올바르게 해당 클래스로 예측한 비율을 의미한다. 세 개의 평가지표는 모두 다른 정보를 나타내기 때문에, 학습의 목적에 따라 평가지표들을 복합적으로 사용하는 것이 중요하다.

하지만 만약 클래스 간의 자료 불균형 문제가 있는 경우, 재현율과는 다르게 정확도와 정밀도는 평가지표로서 사용에 어려움이 발생한다. 재현율의 경우 수식을 구성하는 TP와 FN은 모두 실제 동일한 클래스를 의미하기 때문에 클래스 간 불균형에 영향을 받지 않는다. 그러나 정확도와 정밀도는 서로 다른 클래스 간의 비율로써 표현이 되기 때문에 클래스 간 자료 불균형이 존재하는 경우에는 다수 클래스에 편향된 결과를 얻게 된다.

Fig. 2는 클래스 간 자료의 불균형이 존재하지 않는 경우와 불균형이 존재하는 경우 분류 모델의 학습 결과가 어떻게 달라지는지를 보여준다. 학습 자료로는 사이킷런(Scikit-Learn)에서 제공하는 Two-moon 자료를 사용했으며, 결정 경계는 서포트 벡터 머신을 이용하여 학습시켰다. Fig. 2a는 자료의 불균형이 존재하지 않는 경우로 클래스 A와 클래스 B 모두 1,000개의 샘플을 가지고 있는 반면, Fig. 2b에서 클래스 A는 Fig. 2a와 동일하게 1,000개의 샘플을 가지고 있지만 클래스 B에는 100개의 샘플만을 가지고 있다. 두 경우의 학습된 결정 경계를 비교해 보면 불균형이 존재하는 Fig. 2b의 경우가 불균형이 존재하지 않는 Fig. 2a의 경우보다 다수 클래스인 클래스 A를 더 많이 예측하도록 결정 경계가 이동한 것을 확인할 수 있다. 이

Table 1. Confusion matrix for binary classification.

		Predicted	
		Positive	Negative
Actual	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

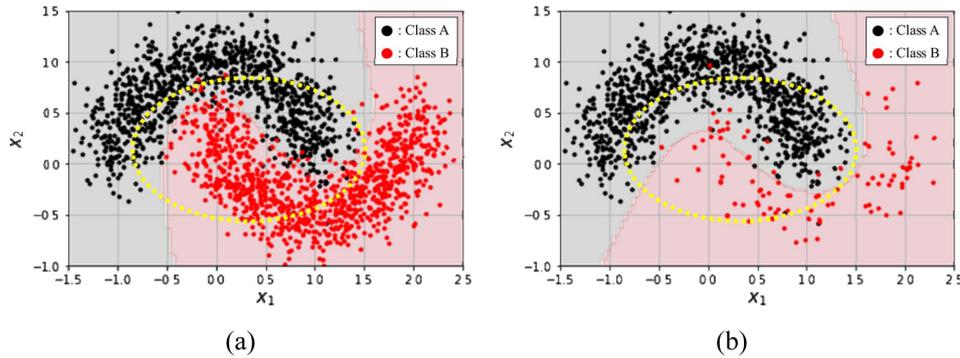


Fig. 2. Comparison of training results for classification models when (a) there is no imbalance between classes and (b) there is an imbalance problem. There is a large difference in the decision boundary in the area outlined with the yellow dotted line.

는 분포 특성이 동일하다고 하더라도 자료 불균형의 존재여부에 따라 부정확한 결정 경계가 형성될 수 있음을 의미한다.

우리는 Two-moon 자료에서 클래스 B의 자료를 1,000개에서 100개까지 점차 줄여감에 따라 학습된 모델의 정확도, 정밀도, 재현율이 어떻게 변화하는지를 관찰하였다(Fig. 3). Fig. 3에서 가로축은 불균형 비율(imbalance ratio, IR)을 나타내며 다음과 같이 정의한다.

$$IR = \frac{\text{The number of majority samples}}{\text{The number of minority samples}}, \quad (8)$$

즉, IR이 증가함은 클래스 간 불균형이 심해짐을 의미한다. Fig. 3에서 재현율의 변화 그래프를 보면 IR이 증가함에 따라 다수 클래스의 재현율이 소폭 상승하고 소수 클래스의 재현율은 크게 감소하는 것을 확인할 수 있다. 이는 IR이 증가할수록 결정경계가 다수 클래스를 많이 예측하도록 이동하게 되며, 결과적으로 실제 소수 클래스 영역이지만 다수 클래스로 오인하는 경우가 늘어났음을 의미한다.

하지만 정밀도 그래프에서는 의미있는 변화를 관찰하기 어렵다. 이는 IR이 증가함에 따라 결정 경계가 잘못 예측되어 소수 클래스로 예측되는 영역이 점차 줄어들지만 그만큼 학습 자료 내 소수 클래스 샘플의 비율 또한 줄어들어 효과가 상쇄되기 때문이다. 따라서 자료 불균형이 심해짐에 따라 결정 경계가 망가지고 다수 클래스로 과대 분류되고 있음에도 불구하고, 오히려 다수 클래스의 정밀도는 소폭 상승하는 경향을 보인다. 이는 정확도에서도 마찬가지로 나타나는 문제이다. IR에 따른 정확도 변화 그래프를 보면 점차 소수 클래스의 영향이 감소하여 자료 불균형이 심해짐에도 정확도는 증가하는 추세를 보인다.

이처럼 학습자료 내 자료 불균형이 존재하는 경우, 정확도와 정밀도는 자료 불균형을 고려하지 않는다는 점에서 평가지표로서 적합하지 않다. 그럼에도 불구하고 예측 모델의 성능 평가 시 정밀도가 갖는 의미가 중요하기 때문에 많은 연구에서 자료 불균형이 존재하는 경우에도 정밀도나 혹은 정밀도와 재

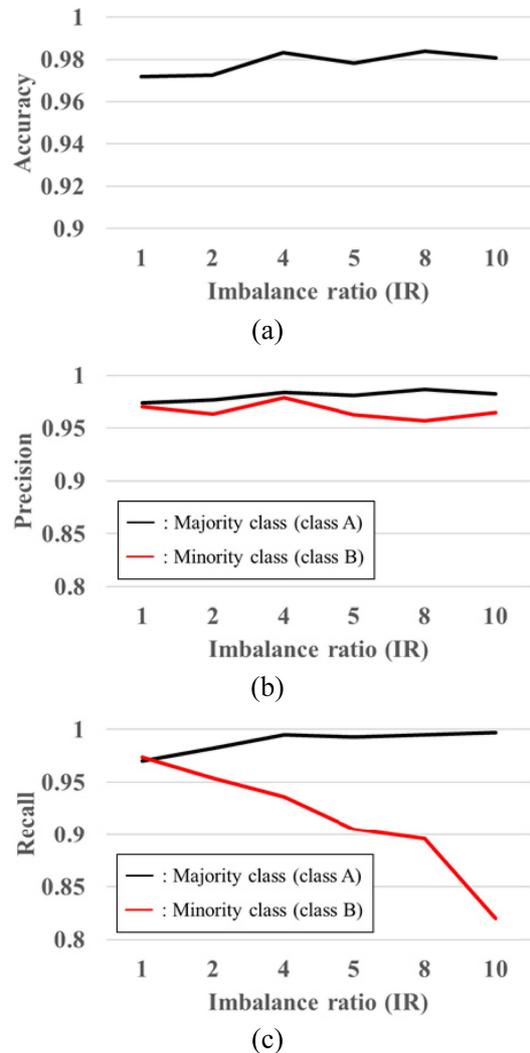


Fig. 3. Variation in (a) accuracy, (b) precision, and (c) recall with changes in the imbalance ratio.

현율의 조화평균을 평가지표로써 사용해왔다(Amin *et al.*, 2015; More, 2016; Suh *et al.*, 2017).

반면 이 연구에서는 기존의 정확도와 정밀도의 구성 식을 개선하여 자료 불균형이 존재하는 경우에도 의미있게 사용할 수 있는 평가지표를 개발하고자 하였다. 자료 불균형이 존재하는 경우에 기존 정확도와 정밀도가 문제가 되는 경우는 특정 클래스의 샘플 개수가 다른 클래스에 비해 현저히 적을 때 발생하게 된다. 즉, 만약 양성 클래스(positive class)가 소수 클래스라고 했을 때, 실제 양성 클래스인 TP와 FN이 실제 음성 클래스(negative class)인 TN, FP와 비교하여 작아진다. 따라서 불균형이 심해질수록 정확도와 정밀도가 TP와 FN보다는 TN과 FP에 의해서 지배적인 영향을 받게 된다. 이에 우리는 TP와 FN에 IR만큼 가중치를 준 가중정확도(weighted accuracy)와 가중정밀도(accuracy precision)를 고안하였다.

$$\text{Weighted accuracy} = \frac{TP * IR + TN}{(TP + FN) * IR + (FP + TN)}, \quad (9)$$

$$\text{Weighted precision} = \frac{TP * IR}{TP * IR + FP}. \quad (10)$$

즉, 자료 불균형이 존재하지 않는 경우 가중정확도와 가중정밀도는 기존의 정확도, 정밀도와 동일하지만 자료 불균형이 존재하는 경우에는 IR이 고려된다. Fig. 4는 Two-moon 자료에서 IR에 따른 가중정확도와 가중정밀도의 변화 그래프이다. IR이 증가함에 따라 결정 경계가 세로축 방향으로 하강하고 있다는 정보를 Fig. 3의 기존 정밀도 그래프에서는 찾아볼 수

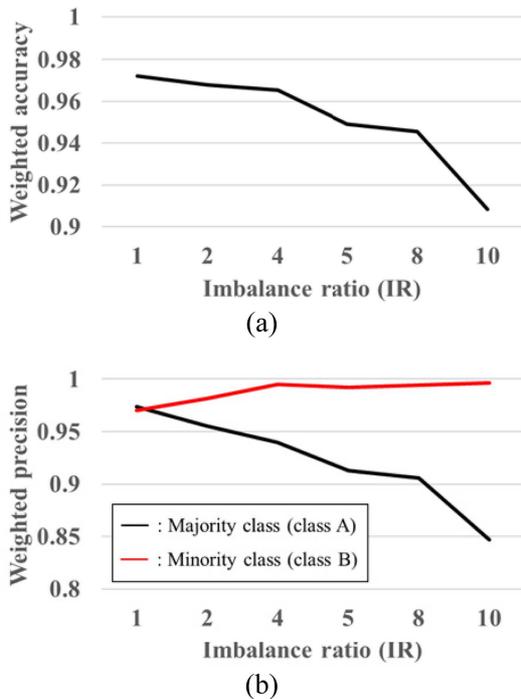


Fig. 4. Variation of (a) weighted accuracy and (b) weighted precision with changes in the imbalance ratio (IR).

없었으나 Fig. 4에서는 다수 클래스의 가중정밀도가 점차 감소하고 소수 클래스의 가중정밀도는 점차 증가한다는 점에서 결정 경계의 변화를 모니터링 할 수 있다. 마찬가지로 가중정확도 또한 IR이 증가함에 따라 감소하는 추세를 보이고 있는데, 이를 통해 자료 불균형이 심해질수록 이상적인 결정 경계로부터 멀어지고 있음을 잘 표현한다.

오버샘플링 적용 시 합성자료의 양 결정

우리는 자료 불균형 문제를 해결하기 위하여 앞서 소개한 Borderline-SMOTE를 적용하였다. 오버샘플링을 적용하여 합성자료를 생성할 때에는 합성자료의 크기를 결정하는 것이 중요하다. 합성자료의 양을 너무 적게 설정하면 자료 불균형 문제를 충분히 해소하지 못하게 되어, 다수 클래스에 대하여 과적합된 모델이 학습될 수 있다. 반면 합성자료를 너무 많이 생성하면 반대로 소수 클래스에 대하여 과적합될 수 있다. 따라서 이 연구에서는 재현율과 고안한 가중정밀도를 이용하여 합성자료의 양을 결정하였다. Fig. 5는 Fig. 2b의 IR이 10인 불균형 자료를 이용하여 IR이 1이 될 때까지 오버샘플링을 수행하였을 때 가중정확도, 가중정밀도, 재현율이 어떻게 변화하는지 나타낸 그래프이다. 그 중 가중정밀도와 재현율 그래프를 보면 두 그래프 모두에서 관계가 역전되는 교차점이 발생하는 것을 확인할 수 있다. 각 그래프에서 교차점은 다수 클래스와 소수 클래스의 가중정밀도와 재현율이 균형을 이루는 지점을 의미한다. 즉, 해당 교차점보다 더 적은 합성 자료를 생성하는 경우 합성자료를 충분히 생성했다고 보기 힘들고, 반대로 더 많은 합성 자료를 생성하는 경우 소수 클래스로의 과적합이 발생했다고 볼 수 있다. 따라서 우리는 가중정밀도와 재현율을 함께 모니터링하여 역전이 발생하기 직전까지 오버샘플링을 적용함으로써 과적합이 발생하지 않는 선에서 최대로 자료증대를 수행하였다. Two-moon 자료의 경우 두 클래스가 대칭의 형태로 분포하기 때문에 Fig. 5에서는 가중정밀도와 재현율의 교차점이 동일하게 IR이 약 3일 때 발생한다. Fig. 6은 IR이 3이 될 때까지 오버샘플링을 수행하였을 때 학습된 결정 경계이다. Fig. 2b와 비교하였을 때 소수클래스에 대하여도 충분히 고려가 이루어진 경계가 형성된 것을 확인할 수 있다.

우리는 오버샘플링을 적용하기 전과 후 학습된 결정 경계들을 정량적으로 비교하기 위해 서로 다른 IR을 갖는 여러 개의 데이터셋에서 학습된 결정 경계들을 Fig. 2a에 도시된 IR=1인 이상적인 데이터셋에 적용하였다(Fig. 7). Figs. 7a, 7c, 7e의 경우 오버샘플링을 적용하기 전 학습된 결정경계를 이상적인 자료에 적용하였을 때의 결과이고 Figs 7b, 7d, 7f는 오버샘플링 후 학습된 결정경계를 이상적인 자료에 적용하였을 때의 결과이다. 오버샘플링은 각각의 경우 정밀도와 재현율의 역전을 모니터링하여 각기 최대로 수행하였다. 오버샘플링을 적용하기 전에는 IR이 증가함에 따라 다수클래스의 정밀도와 소수

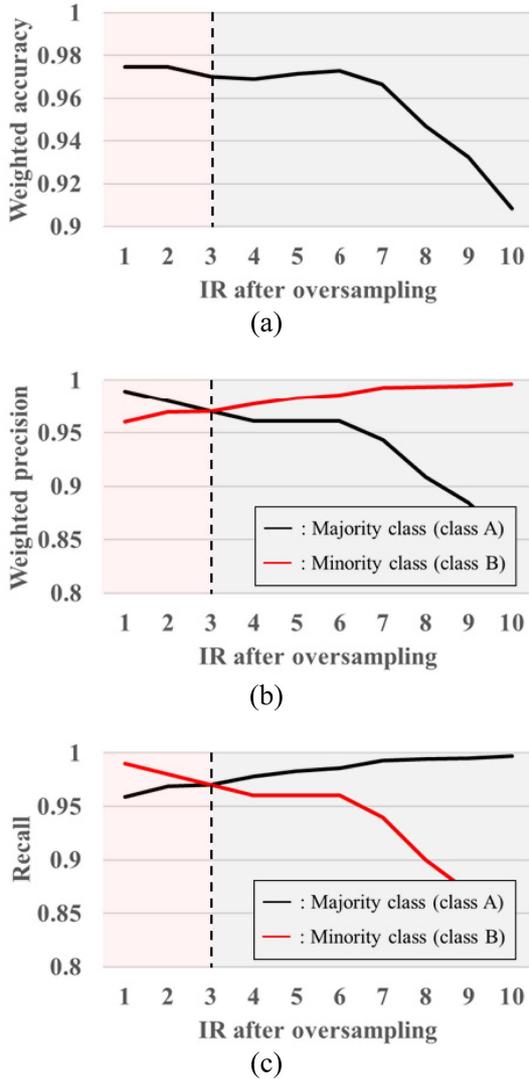


Fig. 5. The changes in (a) weighted accuracy, (b) weighted precision, and (c) recall according to the IR values. The IR is changed from 1 to 10 with oversampling. The dotted lines indicate the IR value at which the weighted precision and recall of the majority class and minority class are equal.

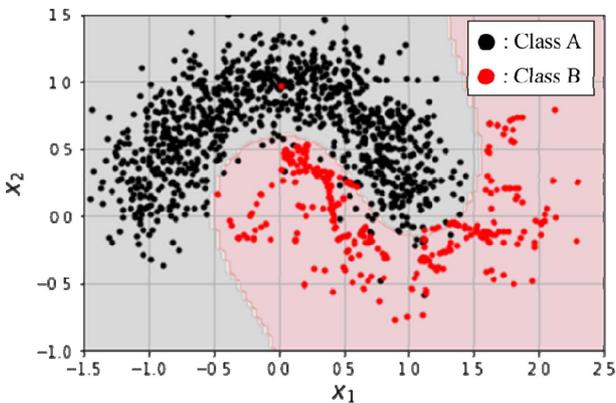


Fig. 6. Decision boundary trained with the data used in Figure 2b after oversampling until the IR becomes 3.

클래스의 재현율이 급격하게 낮아지는 등 전형적인 클래스 사이의 자료 불균형 문제들이 나타난다. 반면 오버샘플링을 적용한 후에는 두 클래스의 정밀도와 재현율이 균형을 이루도록 결정경계가 학습되었고, 정확도 또한 IR이 1보다 큰 학습자료를 사용한 경우에도 이상적인 학습자료를 사용한 경우와 비슷한 수준이 유지되는 것을 확인할 수 있다.

다중 클래스 분류 모델에서 오버샘플링 적용의 자동화

Two-moon 자료와 같이 이진 분류를 수행하는 경우 다수 클래스와 소수 클래스가 명확하게 구분된다. 하지만 암석 및 공극유체의 분류의 경우 암석의 종류 및 공극유체에 따라 다양한 클래스가 만들어질 수 있다. 이와 같이 다중 클래스 분류 모델을 학습시키는 경우 다수 클래스와 소수 클래스는 상대적이게 된다. 예를 들어 오일샌드(oil sand)는 워터샌드(water sand)와 비교하였을 때 소수 클래스이지만 가스샌드(gas sand)와 비교하였을 때에는 다수 클래스일 수 있다. 다중 클래스의 경우 IR은 다음과 같이 정의된다.

$$IR_i = \frac{\max(\text{The number of samples in } j^{\text{th}} \text{ class})_{j=1 \sim n}}{\text{The number of samples in } i^{\text{th}} \text{ class}}, \quad (11)$$

$$i = 1 \sim n$$

여기서 n 은 클래스의 개수이다. 식(8)을 이용하면 다중 클래스에서의 가중정확도와 가중정밀도는 각각 다음과 같이 표현할 수 있다.

$$\text{Weighted accuracy} = \frac{\sum_i^n TP_i * IR_i}{\sum_i^n (TP_i + FN_i) * IR_i}, \quad (12)$$

$$\text{Weighted precision}_i = \frac{TP_i * IR_i}{TP_i * IR_i + \sum_{j, j \neq i}^n FP_{ji} * IR_j}, \quad (13)$$

여기서 TP_i 와 FN_i 는 각각 i 번째 클래스의 TP와 FN을 의미하며, FP_{ji} 는 i 번째 클래스로 예측한 결과 중 실제 클래스가 j 인 경우를 의미한다.

이 연구에서는 Fig. 8과 같이 반복적으로 오버샘플링 비율을 조절하여 클래스 간 가중정밀도와 재현율의 균형을 맞추어주는 알고리즘을 개발하였다. 먼저 자료 불균형이 심할수록 소수 클래스의 재현율이 낮아지고 가중정밀도는 증가한다는 점에 착안하여, 각 클래스의 재현율과 가중정밀도의 차이를 계산하여 가장 작은 값을 갖는 클래스를 소수 클래스, 가장 높은 값을 갖는 클래스를 다수 클래스로 정의하였다. 이후 정의된 다수 클래스와 소수 클래스의 경계부에 합성자료를 소량씩 생성하여 분류 모델을 재학습시킨다. 해당 작업은 중단 조건을 만족할 때까지 반복적으로 수행된다. 중단 조건은 소수 클래스로

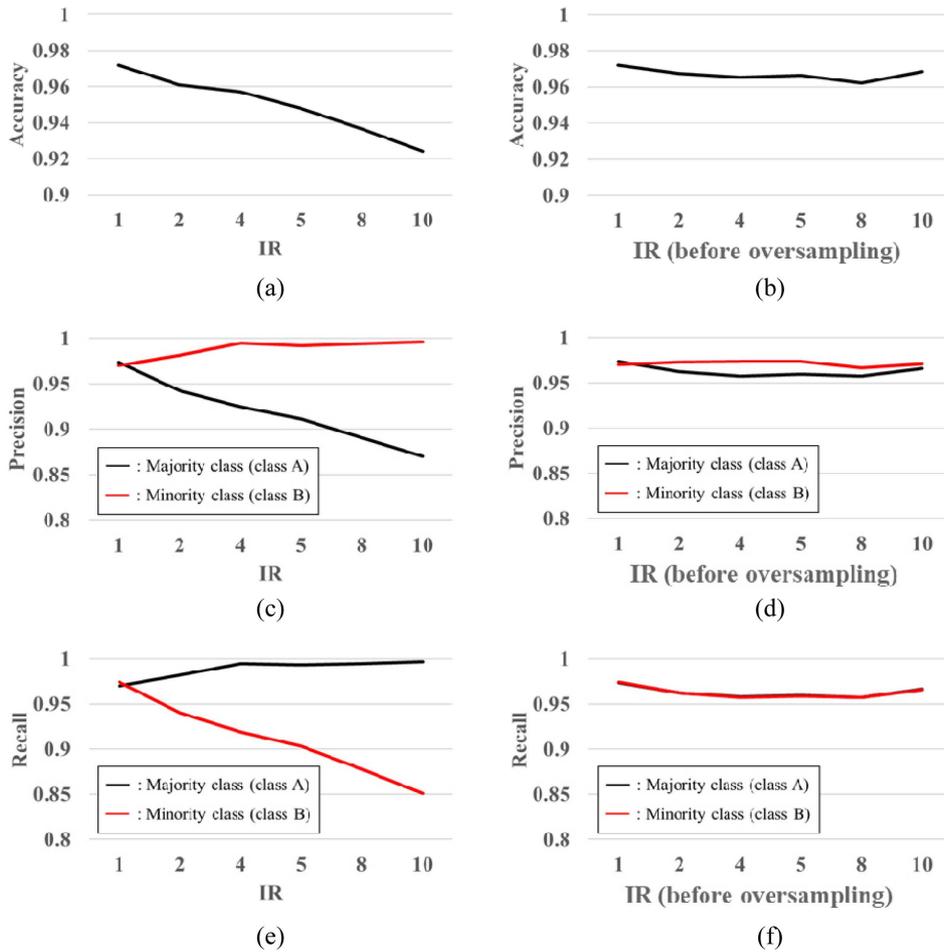


Fig. 7. Comparison of trained decision boundaries before and after oversampling. The result of applying the decision boundaries to an ideal dataset is IR = 1. Panels (a), (c), and (e) illustrate the accuracy, precision, and recall graphs, respectively, of the results classified by the trained model before oversampling. Panels (b), (d), and (f) are the accuracy, precision, and recall graphs, respectively, when oversampling is applied.

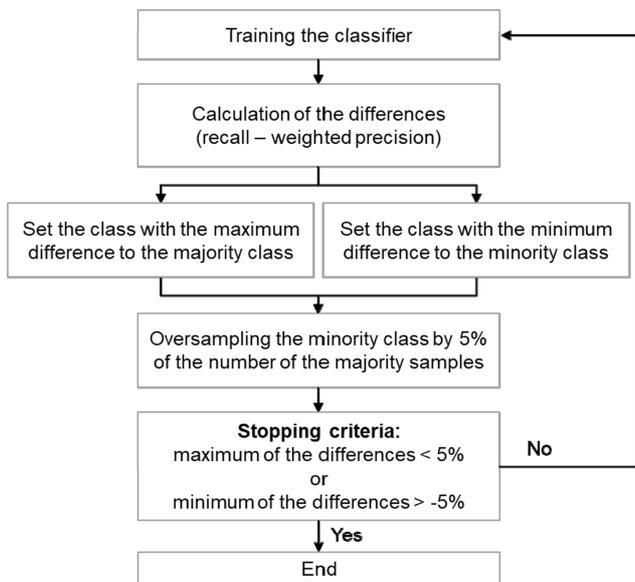


Fig. 8. Automatic oversampling algorithm designed for application to multi-class classification models.

의 과적합을 방지하기 위하여 존재하며, 이 연구에서는 재현율과 가중정밀도 차이의 최대값이 5%보다 작거나 최소값이 -5%보다 커지는 경우 중단하였다.

빈센트 유전 현장 자료를 이용한 다중 분류 실험

우리는 개발된 알고리즘의 적용을 위하여 서호주 내 위치한 빈센트 유전(Vincent oil field)에서 취득된 탄성과 탐사 자료를 이용하여 암상 및 공극유체 분류를 수행하였다. 학습자료로는 빈센트 지역 내 존재하는 세 개의 시추공(Vincent 1, Vincent 2, Vincent 3)에서 취득된 검층자료를 사용하였다. 학습에 사용된 입력 자료로는 음향 임피던스(AI), 탄성 임피던스(SI), 람다로($\lambda\rho$), 뮤로($\mu\rho$), 포아송비(ν), 이로($E\rho$), 속도비(Vp/Vs) 총 일곱 개의 물성을 사용하였다. 여기서 λ 는 라메 상수, ρ 는 밀도, μ 는 전단 탄성계수, E는 영률, Vp 와 Vs 는 각각 P파와 S파의 속도를 의미한다. 또한 해당 지역에서 취득된 탄성과 탐사자료도 임피던스 역산(Hampson *et al.*, 2005)을 수행

Table 2. The calculated evaluation metrics when training a facies classifier using imbalanced training data.

	Class			
	Shale	Water sand	Oil sand	Gas sand
Weighted precision	0.990	0.820	0.998	1.000
Recall	0.835	0.995	0.942	1.000
Weighted accuracy	0.943			

Table 3. The calculated evaluation metrics when training a facies classifier after oversampling.

	Class			
	Shale	Water sand	Oil sand	Gas sand
Weighted precision	0.957	0.897	0.988	1.000
Recall	0.919	0.956	0.961	1.000
Weighted accuracy	0.959			

하여 동일한 물성으로 변환하여 탄성과 탐사 영역의 암상 및 공극유체까지 예측해보고자 하였다.

암상 및 공극유체에 따른 클래스는 물리검층 자료를 해석하여 분류하였다. 먼저 셰일체적 50%를 기준으로 셰일(shale)과 샌드(sand)로 분류하였다. 다음으로 수포화도 70%를 기준으로 샌드를 워터샌드(water sand)와 하이드로카본샌드(hydrocarbon sand)로 구분하였다. 마지막으로 밀도 검층 자료를 해석하여 하이드로카본샌드를 오일샌드(oil sand)와 가스샌드(gas sand)로 구분하였다. 최종적으로는 셰일, 워터샌드, 오일샌드, 가스샌드 총 네 개의 상으로 클래스를 구분하였다. 각각의 클래스는 319개, 2539개, 314개, 44개의 샘플을 가지고 있어 복잡한 자료 불균형을 내포하고 있다.

세 개의 시추공 중 Vincent1와 Vincent3에서 취득된 물리검층 자료를 학습자료로 활용하였다. Vincent2는 학습에 관여하지 않는 자료로 오버샘플링 적용 전과 후의 결과 비교를 위한 검증자료로만 활용하였다. Table 2는 불균형 자료를 그대로 이용하여 분류 모델을 학습시켰을 때, Vincent1와 Vincent3에서의 암상 및 공극유체 분류 결과이다. 분류 모델로는 다섯 개의 은닉층을 지닌 DNN (Deep Neural Network) 모델을 사용하였다. 평가지표를 확인해 보면 자료의 개수가 가장 많았던 워터샌드의 가중정밀도 수치가 현저하게 낮게 나타났으며, 셰일의 경우에는 반대로 매우 낮은 재현율을 기록했다. 이는 두 클래스 간의 자료 불균형으로 인하여 셰일의 중요도가 떨어진 분류 모델이 도출되었음을 의미한다. 비슷하게 오일샌드의 경우에도 다소 재현율이 떨어지는 결과를 획득한 것을 확인할 수 있다. 반면 가장 적은 자료를 가지고 있는 가스샌드의 경우에

는 분류가 완전하게 이루어진 것을 확인할 수 있는데, 이는 자료의 양이 적지만 벡터 공간 내에서 분리성이 좋아 자료 불균형 문제가 발생하지 않은 경우이다.

우리는 암상 및 공극유체 규명에 자료 불균형 문제를 해소하고자 앞서 개발한 오버샘플링 알고리즘을 적용하였다. 총 여덟 번의 반복학습이 수행되었고 최종적으로 셰일 클래스와 오일 샌드에 모두 각각 300개 합성자료를 추가했다. 오버샘플링 적용 후의 학습자료를 이용하여 DNN 모델을 학습시켰을 때 분류 성능은 Table 3과 같다. Table 2와 비교하였을 때 소수 클래스에 해당되는 셰일과 오일샌드의 재현율과 다수 클래스에 해당되는 워터샌드의 가중정밀도가 크게 개선된 것을 확인할 수 있다. 최종적으로는 가중정확도 또한 94.3%에서 95.9%로 개선되었다.

오버샘플링을 후 학습된 모델의 성능 향상을 확인하기 위해 학습에 사용되지 않은 Vincent2의 검층자료를 이용하여 분류를 수행하였다(Table 4). 오버샘플링을 적용한 후 셰일의 재현율과 워터샌드의 가중정밀도가 크게 향상되었다.

Fig. 9은 오버샘플링을 수행하기 전 학습된 모델과 오버샘플링을 수행한 후 학습된 모델을 이용하여 탄성과 탐사 영역에서의 암상 및 공극유체를 예측한 결과 비교이다. 오버샘플링을 적용하기 전에는 자료 불균형으로 인하여 셰일과 오일샌드를 실제보다 적게 예측하고 워터샌드를 실제보다 많게 예측하게 된다. Fig. 9a를 보면 예측에 실패한 영역이 많아져 덮개암과 오일샌드의 영역을 구분하기 어렵다. 반면 오버샘플링을 적용한 경우인 Fig. 9b를 보면 셰일과 오일샌드의 분류 성능이 향상됨에 따라 셰일과 오일샌드의 경계가 보다 명확하게 확인된다.

결 론

이 연구에서 우리는 클래스 간 자료 불균형 문제를 해결하기 위해 자료 불균형을 고려한 평가지표를 개발하고 이를 기준으로 오버샘플링을 수행하였다. 제안한 가중정확도와 가중정밀도는 기존의 정확도, 정밀도와 비교하여 자료 불균형 비율을 고려할 수 있다. 따라서 클래스 간 불균형이 심한 자료일수록 제안된 평가지표들이 보다 효과적임을 확인하였다. 또한 가중정밀도는 재현율과 함께 합성자료의 생성량을 결정하는 기준으로써 사용되었다. 이진 분류 시에는 다수클래스와 소수클래스 사이의 가중정밀도와 재현율의 역전이 발생하지 않도록 오버샘플링을 수행하였다. Two-moon 자료의 경우 다수클래스와 소수클래스의 분포 형태가 대칭을 이루어 가중정밀도 및

Table 4. Comparison of classification performance of models trained before and after oversampling with blind well (Vincent2).

	Class (Before oversampling)			Class (After oversampling)		
	Shale	Water sand	Oil sand	Shale	Water sand	Oil sand
Weighted precision	0.998	0.782	0.983	0.996	0.816	1.000
Recall	0.706	0.997	1.000	0.810	0.997	0.966
Weighted accuracy	0.901			0.924		

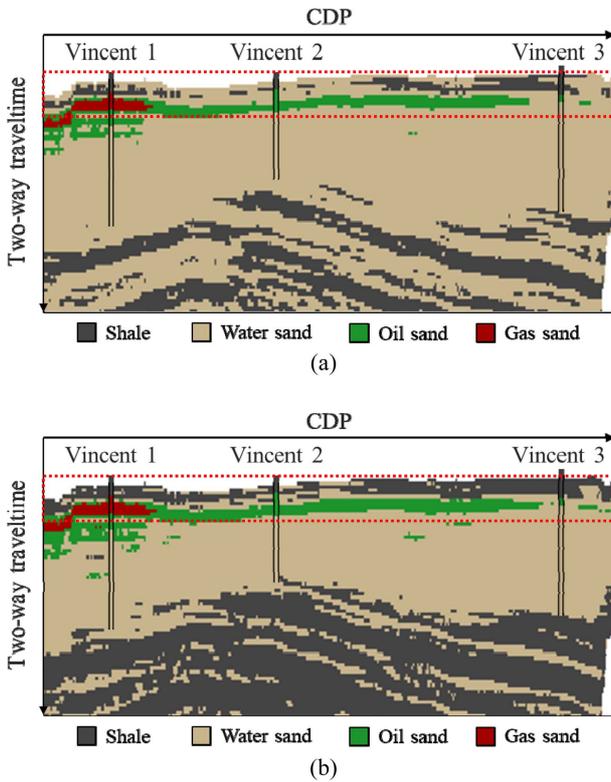


Fig. 9. The results of the facies classification of a seismic exploration area. The facies are predicted using the trained model (a) before and (b) after oversampling. The areas outlined by red dotted lines have been noticeably improved.

재현율의 역전이 동일한 시점에서 발생하였으나, 대칭의 형태가 아닌 경우의 자료에 대해서는 추후 추가적인 연구가 필요할 것으로 보인다.

개발된 평가지표를 이용한 오버샘플링 전략을 다중 클래스 분류 모델에도 적용이 가능하도록 확장하였다. 오버샘플링을 조금씩 반복적으로 수행하여 자동으로 클래스 간 균형을 맞추어주는 알고리즘을 개발하고 이를 이용해 암상 및 공극유체 규명 문제에 적용하였다. 서호주 빈센트 지역에서의 현장자료 검증 결과, 셰일과 오일샌드의 재현율이 알고리즘을 적용하였을 때 각각 8.4%, 1.9% 증가하였으며 워터샌드의 가중정밀도가 7.7% 증가하는 효과를 볼 수 있었다. 최종적으로는 가중정확도 또한 1.6% 상승하였다. 학습에 참여하지 않은 Vincent2를 이용하여 분류 결과를 비교해본 결과에서도 셰일의 재현율, 워터샌드의 가중정밀도 등이 향상되어 가중정확도가 2.3% 향상되었다.

개발된 평가지표와 이를 이용한 오버샘플링 자동화 알고리즘은 이 연구에서 다뤘던 암상 및 공극유체 규명 문제뿐만 아니라 자료불균형이 존재하는 경우, 어떠한 분류 모델 학습에도 적용할 수 있다는 점에서 큰 확장성을 지닐 것으로 기대된다. 반면 제안한 전략의 개선 가능성도 존재할 것으로 보인다. 특

히 오버샘플링 기법으로 취한 Borderline-SMOTE 기법 외 다른 오버샘플링 기법의 적용 가능성이 검토될 수 있다.

감사의 글

본 연구는 산업통상자원부(MOTIE)와 한국에너지기술평가원(KETEP)의 지원을 받아 수행한 연구 과제입니다(No. 20194010201920 & 20182510102471).

References

- Amin, A., Rahim, F., Ali, I., Khan, C., and Anwar, S., 2015, A comparison of two oversampling techniques (SMOTE vs MTFD) for handling class imbalance problem: A case study of customer churn prediction, in Rocha A., Correia A., Costanzo S., Reis L., eds., *New Contributions in Information Systems and Technologies: Advances in Intelligent Systems and Computing*, **353**, 215-225, doi: 10.1007/978-3-319-16486-1_22.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P., 2002, SMOTE: synthetic minority over-sampling technique, *J. Artif. Intell. Res.*, **16**, 321-357, doi: 10.1613/jair.953.
- Chi, X. G., and Han, D. H., 2009, Lithology and fluid differentiation using a rock physics template, *Lead. Edge*, **28(1)**, 60-65, doi: 10.1190/1.3064147.
- Choi, J., Kim, S., Kim, B., and Byun, J., 2019, Probabilistic reservoir characterisation using 3D pdf of stochastic forward modelling results in Vincent oil field, *Explor. Geophys.*, **51(3)**, 341-354, doi: 10.1080/08123985.2019.1696151.
- Chopra, S., and Marfurt K. J., 2005, Seismic attributes — A historical perspective, *Geophysics*, **70(5)**, 3SO-28SO, doi: 10.1190/1.2098670.
- Gao, D., 2003, Volume texture extraction for 3D seismic visualization and interpretation, *Geophysics*, **68(4)**, 1294-1302, doi: 10.1190/1.1598122.
- Goodway, B., Chen, T., and Downton, J., 1997, Improved AVO fluid detection and lithology discrimination using Lamé petrophysical parameters; “ $\lambda\rho$ ”, “ $\mu\rho$ ”, & “ λ/μ fluid stack”, from P and S inversions, *67th Ann. Internat. Mtg., Soc. Expl. Geophys., Expanded Abstracts*, 183-186, doi: 10.1190/1.1885795.
- Hampson, D. P., Russell, B. H., and Bankhead, B., 2005, Simultaneous inversion of prestack seismic data, *75th Ann. Internat. Mtg., Soc. Expl. Geophys., Expanded Abstracts*, 1633-1637, doi: 10.1190/1.2148008.
- Han, H., Wang, W. Y., and Mao, B. H., 2005, Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning, in Huang D. S., Zhang X. P., Huang G. B., eds, *Advances in Intelligent Computing, ICIC 2005, Lecture Notes in Computer Science*, **3644**, 878-887, doi: 10.1007/11538059_91.
- He, H., and Garcia, E. A., 2009, Learning from imbalanced

- data, *IEEE Trans. Knowl. Data Eng.*, **21(9)**, 1263-1284, doi: 10.1109/TKDE.2008.239.
- Jin, L., 2018, Machine learning approaches for seismic-facies prediction and reservoir-property inversion, *88th Ann. Internat. Mtg., Soc. Expl. Geophys., Expanded Abstracts*, 2147-2151, doi: 10.1190/segam2018-2996374.1.
- Kim, D., Choi, J., and Byun, J., 2019, Facies classification using oversampling based machine learning, *2019 Fall Joint Conference of KSMER-KSRM-KSEG* (in Korean).
- More, A., 2016, Survey of resampling techniques for improving classification performance in unbalanced datasets, *arXiv preprint arXiv:1608.06048*.
- Naeini, E. Z., 2019, A machine learning approach to quantitative interpretation, *89th Ann. Internat. Mtg., Soc. Expl. Geophys., Expanded Abstracts*, 3176-3180, doi: 10.1190/segam2019-3216138.1.
- Nieto, J., Batlái, B., and Delbecq, F., 2013, Seismic lithology prediction: a Montney shale gas case study, *CSEG Recorder*, **38(2)**, 34-43.
- Pendrel, J., Schouten, H., and Bornard, R., 2017, Bayesian estimation of petrophysical facies and their applications to reservoir characterization, *87th Ann. Internat. Mtg., Soc. Expl. Geophys., Expanded Abstracts*, 3082-3086, doi: 10.1190/segam2017-17588007.1.
- Saleem, A., Choi, J., Yoon, D., and Byun, J., 2019, Facies classification using semi-supervised deep learning with pseudo-labeling strategy, *89th Ann. Internat. Mtg., Soc. Expl. Geophys., Expanded Abstracts*, 3171-3175, doi: 10.1190/segam2019-3216086.1.
- Suh, Y., Yu, J., Mo, J., Song, L., and Kim, C., 2017, A comparison of oversampling methods on imbalanced topic classification of Korean news articles, *Journal of Cognitive Science*, **18(4)**, 391-437, doi: 10.17791/jcs.2017.18.4.391.
- Yenugu, M., Marfurt, K. J., and Marson, S., 2010, Seismic texture analysis for reservoir prediction and characterization, *Lead. Edge*, **29(9)**, 1116-1121, doi: 10.1190/1.3485772.