지구물리와 물리탐사 Geophysics and Geophysical Exploration
Vol. 23, No. 3, 2020, p. 157-167 https://doi.org/10.7582/GGE.2020.23.3.00157

ISSN 1229-1064 (Print)
ISSN 2384-051X (Online)

베이지안 최적화를 이용한 암상 분류 모델의 하이퍼 파라미터 탐색

최용욱[1] · 윤대웅[1]* · 최준환[2] · 변중무[2]

[1]전남대학교 에너지자원공학과
[2]한양대학교 자원환경공학과

Hyperparameter Search for Facies Classification with Bayesian Optimization

Yonguk Choi[1], Daeung Yoon[1]*, Junhwan Choi[2], and Joongmoo Byun[2]

[1]Dept. Energy & Resources Engineering, Chonnam National University
[2]Dept. of Earth Resources and Environmental Engineering, Hanyang University

요 약: 최근 인공지능 기술의 발전과 함께 물리탐사의 다양한 분야에서도 인공지능의 핵심 기술인 머신러닝의 활용도가 증가하고 있다. 또한 머신러닝 및 딥러닝을 활용한 연구는 이미지, 비디오, 음성, 자연어 등 다양한 태스크의 추론 정확도를 높이기 위해 복잡한 알고리즘들이 개발되고 있고, 더 나아가 자료의 특성, 알고리즘 구조 및 하이퍼 파라미터의 최적화를 위한 자동 머신러닝(AutoML) 분야로 그 폭을 넓혀가고 있다. 본 연구에서는 AutoML 분야 중에서도 하이퍼 파라미터(hyperparameter) 자동 탐색을 위한 베이지안 최적화 기술에 중점을 두었으며, 본 기술을 물리탐사 분야에서도 암상 분류(facies classification) 문제에 적용했다. Vincent field의 현장 물리검층 및 탄성파 자료를 이용하여 암상 및 공극 유체를 분류하는 지도학습 기반 모델에 적용하였고, 랜덤 탐색 기법의 결과와 비교하여 베이지안 최적화 기반 예측 프레임워크의 효율성을 검증하였다.

주요어: 암상 분류, 베이지안 최적화, 랜덤탐색, 자동머신러닝, k겹 교차검증

Abstract: With the recent advancement of computer hardware and the contribution of open source libraries to facilitate access to artificial intelligence technology, the use of machine learning (ML) and deep learning (DL) technologies in various fields of exploration geophysics has increased. In addition, ML researchers have developed complex algorithms to improve the inference accuracy of various tasks such as image, video, voice, and natural language processing, and now they are expanding their interests into the field of automatic machine learning (AutoML). AutoML can be divided into three areas: feature engineering, architecture search, and hyperparameter search. Among them, this paper focuses on hyperparamter search with Bayesian optimization, and applies it to the problem of facies classification using seismic data and well logs. The effectiveness of the Bayesian optimization technique has been demonstrated using Vincent field data by comparing with the results of the random search technique.

Keywords: facies classification, Bayesian optimization, random search, autoML, k-fold cross validation

서 론

머신러닝(machine learning)과 딥러닝(deep learning)은 인공지능의 핵심 기술로, 기계가 일일이 명시하지 않은 동작을 자료로부터 학습하여 실행할 수 있도록 하는 연구 분야이다. 최근 컴퓨터 하드웨어의 발전과 오픈소스 소프트웨어의 기여로 인공지능 기술에의 접근성이 용이해지면서 물리탐사 분야에도 머신러닝 및 딥러닝 기술을 적용하는 사례가 점차 증가하고 있다. 또한 기존의 전통적인 자료처리 및 해석 기법보다 높은 성능을 발휘하는 다양한 머신러닝 기반의 기술들이 소개되면서 머신러닝과 물리탐사 융합 기술 관련 연구가 전 세계적으로 가속화되고 있는 추세이며, 이는 학제의 패러다임 변화에까지 영향을 미치고 있다. 대표적인 연구 분야로는 탄성파 자료 내삽 및 외삽(Park *et al*., 2019; Yoon *et al.*, 2020), 탄성파 자료 잡음 제거(Li *et al.*, 2018), 탄성파 및 전자탐사 역산(Araya-Polo *et al.*, 2018; Oh *et al.*, 2018), 정량적 탄성파 해석(Choi *et al.*, 2019) 등이 있다.

Received: 16 June 2020; Revised: 17 July 2020; Accepted: 30 July 2020
*Corresponding author
E-mail: duyoon@gmail.com
Address: 77, Yongbong-ro, Buk-gu, Gwangju, Republic of Korea

그 중에서도 정량적 탄성파 해석(Quantitative Seismic Interpretation) 분야는 물리검층 및 탄성파 자료를 이용하여 탄성파 영역에서의 암상 및 공극 유체를 규명하는 방법으로, 저류층 해석의 불확실성을 감소시키기 위해 암석물리학 기반 통계 기술 및 머신러닝 기법들이 사용되어 왔고, 최근 딥러닝 기술 또한 활발히 적용되고 있다. 대표적인 예로, Wolf and Pelissier-Combescure (1982)는 비지도학습(unsupervised learning) 기반의 PCA (Principal Component Analysis) 및 군집화(clustering) 알고리즘을 이용하여 자동화된 암상 분류 기술을 소개하였고, 이와 유사한 목적으로 Baldwin *et al.* (1990)과 Delfiner *et al.* (1987)은 지도학습(supervised learning) 기반의 인공신경망(Artificial Neural Network)을, Wrona *et al.* (2018)은 심층신경망(Deep Nerual Networks)을 활용하였다. 또한 Choi *et al.* (2017)는 몬테 카를로 시뮬레이션 기반의 petro-elastic 모델의 활용으로 부족한 검층자료를 보완하여 암상 분류 모델의 정확도를 향상시켰고, Lee *et al.* (2018)은 지도학습에서의 학습자료 부족 문제를 보완하기 위한 방법으로 준지도학습(semi-supervised learning)기반의 암상 분류 모델을 제시하였다. 이처럼 정량적 탄성파 해석은 물리탐사 분야에서 오랫동안 머신러닝 기술이 사용된 분야이고, 현재도 암상 분류 모델의 정확도를 향상시키기 위한 목적으로 활발히 연구되고 있다.

최근 머신러닝, 딥러닝 분야에서 다양한 알고리즘이 개발되고 알고리즘 구조의 복잡성이 증가함에 따라 한정된 시간 및 컴퓨팅 자원에서 최적의 알고리즘을 자동으로 선택하고 빠른 훈련을 보장하기 위한 AutoML (Automated Machine Learning) 기술에 대한 연구가 활발히 진행되고 있다. 일반적으로 AutoML 기술이 연구되는 방향은 다음과 같이 3가지로 나눌 수 있다.

- 특성 공학 자동화(automated feature engineering)
- 신경망 구조 탐색(architecture search)
- 하이퍼파라미터 최적화(hyperparameter optimization)

먼저 특성 공학 자동화란 주어진 원시데이터를 머신러닝 모델에 적합하게 변형하는 기술을 말한다. 기존에는 사용자의 도메인 지식을 활용하거나 다양한 특성들을 실험적으로 생성하여 최적의 특성을 추출하였는데, 이 작업은 많은 시간이 소요되어 이를 자동화하기 위한 방안으로 현재 다양한 연구들이 진행되고 있다. 대표적인 방법으로는 Deep Feature Synthesis가 있다(Kanter and Veeramachaneni, 2015).

신경망 구조 탐색은 AlexNet, VGGNet, ResNet 등과 같이 사람이 경험적, 실험적으로 신경망의 구조를 설계하는 대신, 학습을 통해 최적의 성능을 내는 구조를 자동으로 설계하는 방법을 말한다. 주로 강화학습(Reinforcement Learning), 유전 알고리즘(Evolutionary Algorithm) 등의 방법으로 학습을 진행하고, 대표적인 방법론으로는 NAS (Neural Architecture Search), NASNet, DARTS (Differentiable Architecture Search) 등이 있다(Zoph and Le, 2016).

마지막으로 하이퍼 파라미터 최적화는 머신러닝 및 딥러닝 모델 학습에 필요한 하이퍼 파라미터값을 학습을 통해 추정하는 것을 의미한다. 여기서 하이퍼 파라미터란 모델 학습 프로세스가 시작되기 전에 값이 설정되는 매개 변수를 말하고, 학습률, 훈련 반복횟수, 배치 사이즈, 뉴런 및 은닉층의 개수 등과 같이 모델의 훈련 성능에 직접적인 영향을 미치는 변수이다. 최적의 하이퍼 파라미터를 찾는 과정은 일반적으로 반복적인 시행오차(try and error) 방식을 통해 수행되기 때문에 시간과 노력이 많이 필요로 하는 작업이고, 특히 최근 머신러닝 및 딥러닝 모델이 복잡해지고 하이퍼 파라미터의 수가 증가함에 따라 모델 성능의 극대화를 위한 하이퍼 파라미터의 최적 조합을 탐색하는 기술의 필요성이 증대되고 있다.

본 연구에서는 AutoML의 위 3가지 분야 중 하이퍼 파라미터 최적화 기술을 물리탐사 분야의 문제에 적용해 보고자 한다. 하이퍼 파라미터 최적화 기술은 지도학습 기반의 머신러닝 알고리즘이 사용되는 모든 문제에 활용 가능하지만, 본 논문에서는 물리탐사 분야의 대표적인 머신러닝 활용 문제인 정량적 탄성파 해석을 위한 암상 분류 문제에 적용하여 그 기술의 효율성을 입증하고자 한다.

지도학습 기반 암상 분류 모델

물리검층 자료로부터 시추공 주변 매질에서의 암상 및 공극 유체 정보는 획득할 수 있으나, 시추공으로부터 멀리 떨어진 지역에서의 암상 및 공극 유체를 규명하는 데는 한계가 있다. 하지만 물리검층 자료와 탄성파 자료에서 공통된 탄성 물성을 추출하고, 물리검층 자료를 이용하여 탄성 물성과 암상 및 공극 유체간의 관계를 만족하는 함수를 찾을 수 있다면, 이 함수를 이용하여 탄성파 영역에서의 암상 및 공극 유체를 예측할 수 있다. 이때 탄성 물성과 암상 및 공극 유체간의 연결 함수를 결정하기 위해 머신러닝 지도학습 기법을 사용할 수 있다(Yoon *et al.*, 2018).

머신러닝 지도학습 기반 암상 분류 모델은 물리검층 자료의 탄성 물성을 입력으로, 암상 및 공극 유체 정보를 레이블로 이용하여 학습을 통해 얻을 수 있다. P파 검층(sonic log)과 밀도 검층(density log) 그리고 S파 검층(shear log)을 이용하여 입력 자료인 P-impedance, S-impedance, Vp-Vs, Lambda-rho (LR), Mu-rho (MR) 등의 탄성 물성을 얻을 수 있고, Gamma ray 및 전기비저항 검층의 분석을 통해 shale, wet sand, oil sand, gas sand와 같은 암상 및 공극 유체로 분류하고 이를 레이블로 사용한다.

다음으로 물리 검층 자료를 이용하여 학습된 모델은 탄성파 자료로부터 추출된 탄성 물성에 적용된다. 이때 사용되는 탄성 물성은 학습된 모델의 입력 자료의 특성(feature)과 동일해야 하며, 이는 AVO (Amplitude Versus Offset) 역산을 통해 추출 가능하다. 이렇게 추출된 탄성 물성에 물리검층 자료로 학습된

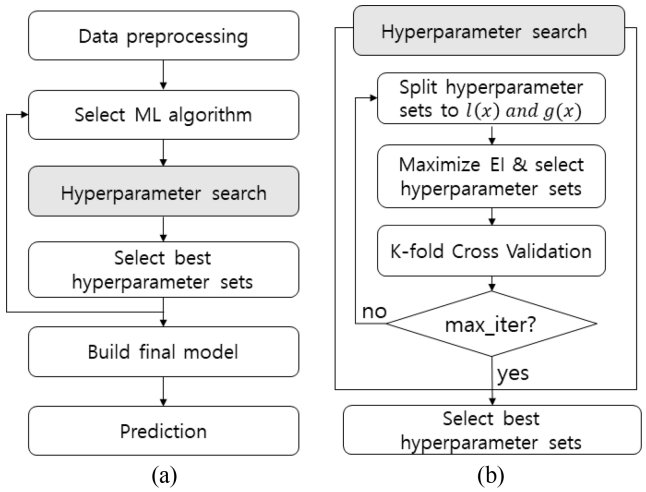

Fig. 1. Flowcharts of (a) procedure of the proposed prediction framework with hyperparameter search and (b) procedure of hyperparameter search using TPE Bayesian optimization and k-fold cross validation.

암상 분류 모델을 적용하면 탄성파 영역에서의 암석 및 유체 특성을 예측할 수 있다. 지도학습을 이용한 암상 분류 모델에 대한 자세한 내용은 Yoon *et al.* (2018)의 9장을 참고 바란다.

하이퍼 파라미터 최적화 기반 예측 모델 결정 프레임워크

본 논문에서는 Fig. 1a와 같이 단계별 예측 모델 결정 프레임워크를 이용하여 지도학습 기반의 최적화된 예측 모델을 결정한다(Nguyen *et al.*, 2020). 제안된 프레임워크는 자료 전처리, 알고리즘 선택, 하이퍼 파라미터 탐색, 최적 하이퍼 파라미터 결정, 최종 예측 모델 결정, 예측 단계로 구성된다.

자료 전처리

자료 전처리 단계는 머신러닝 알고리즘에 적용하기 위한 자료를 준비하는 과정이다. 일반적으로 누락된 자료 처리, 이상치 제거, 특성 추가, 텍스트 또는 범주형 자료 변환, 자료 확장(data augmentation), 특성 스케일링(feature scaling) 등의 프로세스가 적용된다. 이 외에도 PCA (Principle Component Analysis)나 오토인코더(AutoEncoder) 등을 이용한 차원 축소, 또는 생성 모델(generative model)을 이용한 자료 확장 등 최종 모델의 성능을 높이기 위한 고급 전략도 수행된다(Frid-Adar *et al.*, 2018).

알고리즘 선택

지도학습에는 로지스틱 회귀, 서포트 벡터 머신, 랜덤 포레스트, 그래디언트 부스팅, 심층신경망, 합성곱신경망, 순환신경망 등 다양한 알고리즘이 존재한다. 일반적으로 자료의 형태에 따라 선호되는 알고리즘이 있지만 (예를 들어 테이블 형태의 자료는 트리 기반 알고리즘, 영상 자료에는 합성곱 신경망, 시계열 자료에는 순환신경망이 주로 선호됨), NFL (No Free Lunch) 이론에서 말하는 바와 같이 최선의 모델을 찾기 위해서는 모든 모델을 평가해 보는 과정이 필요하다(Wolpert and Macready, 1997). 따라서 Fig. 1a의 알고리즘 선택 단계는 다양한 알고리즘을 이용하여 모델을 평가하는 과정을 반복적으로 수행하는 작업이 필요하다. 하지만 본 연구에서는 하이퍼 파라미터 최적화에 초점을 맞추기 위하여 LightGBM (Light Gradient Boosting Machine) 하나의 알고리즘만을 사용한다.

LightGBM은 Microsoft에서 개발한 오픈소스 알고리즘으로 그래디언트 부스팅 결정 트리(Gradient Boosting Decision, GBDT) 계열에 속한다(Ke *et al.*, 2017). 부스팅은 약한 트리 기반의 학습기(week learner)를 결합하여 강한 학습기(strong learner)를 만드는 앙상블 방법으로, 랜덤 포레스트에서 주로 사용되는 배깅(Bagging: Boostrap Aggregating)과는 다르게 기존의 약한 학습기를 반복적으로 점차 발전시켜서 이를 결합하는 방식으로 작동된다. 대표적인 GBDT로는 Adaptive Boosting, XgBoost, LightGBM 등이 있다. 이 중에서도 LightGBM은 XgBoost의 성능을 개선한 알고리즘으로, 부스팅 기법 적용 시 트리의 깊이(또는 레벨)를 확장하는 Level-wise growth 대신 잎(또는 가지)을 확장시키는 Leaf-wise growth 방식을 선택하여 데이터 피팅(fitting) 능력을 향상 시켰고, 연속된 값을 범주 형식으로 변환하여 특성(feature) 선택의 효율성을 향상하는 히스토그램 기반 알고리즘(histogram based algorithm)과 병렬 컴퓨팅 기술을 활용하여 메모리 및 성능의 효율성을 개선하였다. 본 라이브러리에 관한 자세한 사항은 LightGBM 홈페이지(https://lightgbm.readthedocs.io/en/latest/)에서 확인할 수 있다.

하이퍼 파라미터 탐색

하이퍼 파라미터는 학습률, 학습 반복 횟수 등과 같이 모델 학습 이전에 설정되는 매개 변수를 의미하고, 이는 모델의 성능에 직접적인 영향을 미치기 때문에 최종 모델 결정에 있어서 중요한 역할을 차지한다. 일반적으로 하이퍼 파라미터들은 서로간의 종속성을 가지기 때문에 경험 또는 순차적인 반복을 통한 수동적 탐색(manual search) 방법으로는 최적의 하이퍼 파라미터를 도출하기에는 한계가 있다. 따라서 이를 체계적으로 수행하기 위한 방안으로 다양한 탐색 방법들이 제안되었는데, 그 중에서 대표적인 방법으로는 그리드 탐색(grid search), 랜덤 탐색(random search), 그리고 베이지안 최적화(Bayesian optimization) 방식이 있다.

그리드 탐색과 랜덤 탐색

그리드 탐색은 특정 범위 내에서 일정 간격으로 하이퍼 파라미터값을 선택하여 모델의 성능을 평가하고, 가장 높은 성능을 발휘하는 하이퍼 파라미터값을 최적해로 도출한다. 균등한

탐색이 가능하지만 하이퍼 파라미터의 종류가 많아질수록 탐색 시간이 기하급수적으로 늘어나고, 탐색 간격 따라 최적의 값을 놓칠 가능성이 있다는 단점이 있다. 이를 보완하기 위한 방법으로 랜덤 탐색이 제안되었는데, 본 탐색 기법은 특정 범위 내에서 임의의 값을 무작위로 선정하는 방식으로 진행되며, 불필요한 반복 수행 횟수를 줄일 수 있어 그리드 탐색 보다 빠르게 최적해를 도출할 수 있는 것으로 알려져 있다. 단, 두 방식 모두 하이퍼 파라미터값의 선정 과정에서 사전지식이 반영되지 못해 탐색 과정이 체계적이지 못하다는 단점이 있다.

베이지안 최적화

베이지안 최적화는 임의의 목적함수를 최대로 하는 최적해를 찾는 방법이다. 본 최적화 방법은 관측값을 얻을 수 있는 어떤 함수에도 적용 가능하고, 특히 비용이 많이 들고 형태를 알 수 없는 블랙박스 형태의 목적함수에 사용가능하다(Mockus, 2012). 따라서 이러한 최적화 방법의 특성을 이용하여 머신러닝 모델의 하이퍼 파라미터 최적화 방법으로 주로 사용된다(Bergsta *et al.*, 2011; Li *et al.*, 2018; Klein *et al.*, 2017).

베이지안 최적화의 최대 장점은 그리드 탐색 또는 랜덤 탐색과는 다르게 하이퍼 파라미터값 선정 과정에서 사전지식을 반영한다는 점이다. 이러한 방식을 Sequential Model-based Global Optimization (SMBO)이라 하고, 이는 순차적인 반복 탐색 과정에서 이전 하이퍼 파라미터로부터 얻어지는 목적함수의 추정값을 이용하여 다음 하이퍼 파라미터값을 선정하는 방식을 말한다(Bergsta *et al.*, 2011).

목적함수의 추정값으로 다음 하이퍼 파라미터를 선정하기 위한 기준을 정하기 위해서 선정함수(selection or acquisition function)를 사용하는데 대표적인 선정함수로는 Expected Improvement (EI)가 있다(Jones, 2001).

$$EI_{y^*}(x) := \int_{-\infty}^{\infty} \max(y^* - y, 0) p(y|x) dy \tag{1}$$

여기서 x는 하이퍼 파라미터, y는 목적함수에서 얻어지는 실제 점수(e.g., loss, RMSE 등), y^*는 목적함수의 기준값(일반적으로 상위 15%로 설정), $p(y|x)$는 하이퍼 파라미터를 목적함수에서 얻어진 확률적 스코어로 매핑하는 확률 모델을 의미한다. 이때 EI는 반복적으로 계산되는데, EI를 최대값으로 하는 하이퍼 파라미터 x를 다음 반복에서의 하이퍼 파라미터로 선정된다(Fig. 1b).

$p(y|x)$는 비용이 많이 드는 목적함수를 대체하기 위한 모델로 surrogate 모델이라 부르고, 출력으로 목적함수 점수의 확률값 또는 추정값을 제공한다. 일반적으로 많이 사용되는 Surrogate 모델로는 Gaussian Process, Random Forest, Tree Parzen Estimator (TPE)가 있고, 본 논문에서는 연속형 뿐만 아니라 범주형 및 조건형 하이퍼 파라미터를 사용할 수 있는 TPE에만 중점을 두고 설명하겠다(Bergsta *et al.*, 2011).

TPE 기반의 surrogate 모델, $p(y|x)$를 계산하기 위해 다음과 같이 Bayes'rule을 사용한다.

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \tag{2}$$

여기서 $p(x|y)$는 목적함수의 점수가 주어졌을 때 하이퍼 파라미터값의 확률을 의미하고 다음과 같은 수식으로 표현된다.

$$p(x|y) = \begin{cases} l(x) & \text{if } y < y^* \\ g(x) & \text{if } y \geq y^* \end{cases} \tag{3}$$

여기서 $l(x)$와 $g(x)$는 확률 분포를 의미하고, 목적함수의 점수 y와 기준값 y^*을 비교하여 두 개의 확률 분포로 표현된다. 예를 들어, 총 50번의 반복을 수행하여 min_child_samples에 대한 LightGBM 모델의 점수(로스 값)가 Fig. 2a와 같이 나타난다고 할 때, 기준값 y^*를 상위 15%로 설정 시, 기준값보다 높은 성능(낮은 로스 값)와 낮은 성능(높은 로스 값)에 대해 가우시안 커널을 이용한 확률 밀도 분포 $l(x)$와 $g(x)$는 Fig. 2b와 같이 나타난다. 따라서 다음 반복에서 하이퍼 파라미터 선정 시, $l(x)$는 선호되는 분포로, 반대로 $g(x)$는 피해야 되는 분포로 간주된다.

식 (2)와 (3)을 이용하여 식 (1)을 정리하면 다음과 같다(Bergsta *et al.*, 2011).

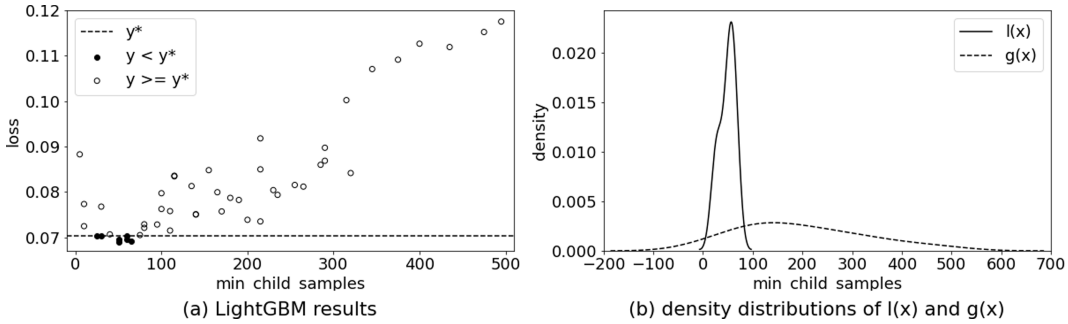

Fig. 2. (a) losses of LightGBM models w.r.t. a hyperparameter of min_child_samples for 50 iterations of Bayesian Optimization. The criteria y^* is set to 15%. (b) The corresponding density distributions of $l(x)$ and $g(x)$.

$$EI_{y^*}(x) = \frac{\gamma y^* l(x) - l(x)\int_{-\infty}^{y^*} p(y)dy}{\gamma l(x) + (1-\gamma)g(x)} \propto \left(\gamma + \frac{g(x)}{l(x)}(1-\gamma)\right)^{-1} \quad (4)$$

여기서 $\gamma = p(y > y^*)$ 이다.

마지막으로 식 (4)의 EI를 최대화 하는 문제를 풀면 다음 반복에서의 하이퍼 파라미터를 선정할 수 있게 되는데, 이때 EI가 $l(x)/g(x)$에 비례하기 때문에, EI의 최솟값 문제는 결국 기준값 y^* 보다 낮은 점수의 분포인 $l(x)$에서 하이퍼 파라미터를 선택하도록 유도되었다고 할 수 있다.

Fig. 1b에서는 TPE 기반 베이지안 최적화의 단계별 과정을 보여주고 있고, 본 논문에서는 hyperopt(http://hyperopt.github.io/hyperopt/) 오픈소스 라이브러리를 사용하여 TPE 기반의 베이지안 최적화를 적용하였다.

k겹 교차검증(k-fold cross validation)

일반적으로 하이퍼 파라미터 선정 후 훈련자료를 훈련세트(training set)와 검증세트(validation set)로 분리한 후 검증세트를 이용하여 모델의 성능을 평가한다. 단 자료의 크기가 작은 경우 검증세트에 대한 성능 평가의 신뢰가 떨어지는 것을 방지하기 위해서 k겹 교차검증(k-fold cross validation)을 실시한다. k겹 교차검증은 Fig. 3와 같이 훈련자료를 k 겹으로 나눈 후 모든 자료를 최소 한 번은 검증세트로 사용하도록 하여 각 겹 당 성능을 평가하는 방법이다. 이때 k의 수는 자료의 양에 따라 달라지고, 주로 3 ~ 10 사이의 값이 사용된다. 또한 검증자료 생성 시 불균형적인 클래스에서 샘플링 편향 현상이 일어나는 것을 방지하기 위해 계층적 샘플링(stratified sampling) 기법을 적용하면 평가의 신뢰도를 높일 수 있다. 마지막으로 모든 겹의 성능 평가 결과를 평균하여 하나의 모델에 대한 성능 평가가 완료된다.

최종 모델 결정

베이지안 최적화와 k겹 교차검증을 반복적으로 수행하여 성능을 평가한 후, 최종 모델을 결정한다. 최종 모델을 결정하는 방법에는 k겹 교차검증을 이용하는 방식과 모든 훈련자료를 이용하는 방식이 있고, 전자의 경우 각 겹에서 얻어진 모델 중 가장 높은 성능의 모델을 선택하는 방법과 각 겹에서 얻어진 모델들을 앙상블하여 단일 모델을 생성하는 방법이 있다. 단, 두 방법 모두 검증세트를 제외한 나머지 훈련 세트로만 모델을 학습해야하는 단점이 있어, 비교적 적은 양의 자료를 사용할 때는 모든 훈련자료를 이용하는 방식을 주로 선택한다. 본 연구에서도 최종 선정된 하이퍼 파라미터만을 추출하여, 모든

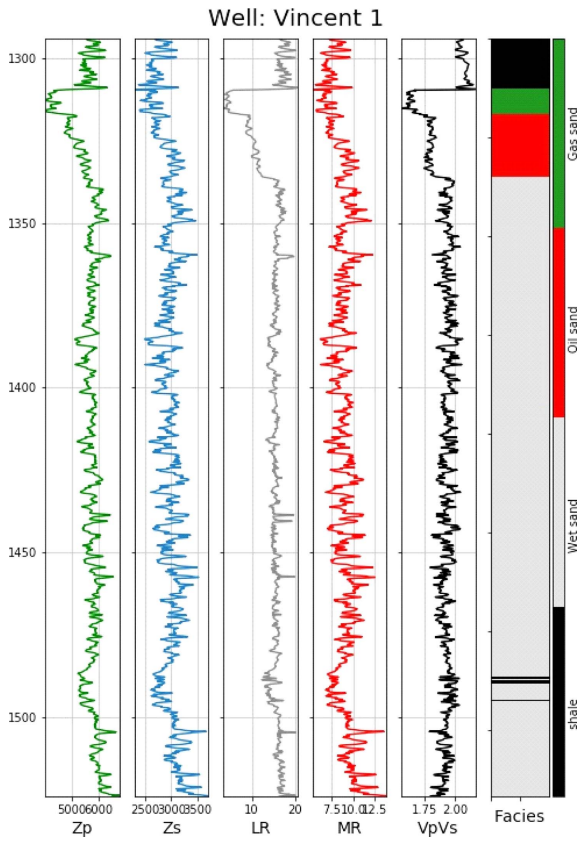

Fig. 4. P-impedance (Zp), S-impedance (Zs), Lambda-Rho (LR), Mu-Rho (MR), Vp/Vs, and Facies from Vincent 1

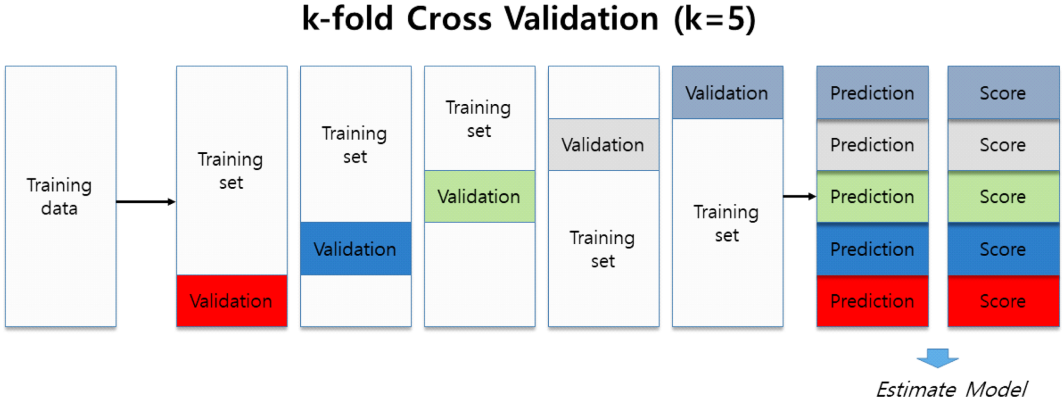

Fig. 3. Schematic of k-fold cross validation

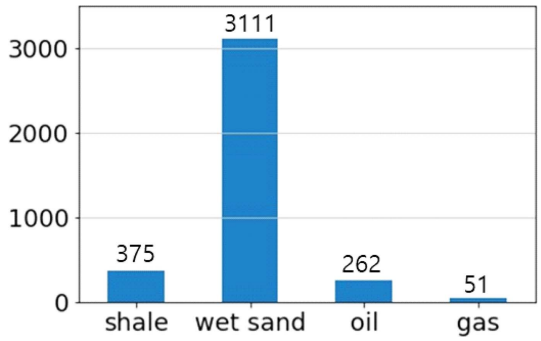

Fig. 5. Imbalanced dataset of well log data.

훈련자료를 재학습 시킨 후 최종 모델을 결정한다.

베이지안 최적화를 활용한 암상 분류

본 논문에서는 서호주의 Exmouth 하부분지에서 획득된 Vincent field 자료를 이용하여 지도학습 기반의 암상 분류 모델을 학습하였고, TPE 기반 베이지안 최적화 기법을 적용하여 제안된 프레임워크의 효율성을 입증하였다. Vincent field는 Muderong shale 트랩과 Sandstone 기반의 가스 및 석유층이 포함된 3방향 경사/단층 폐쇄구조이다. 탐사 자료로는 Vincent field의 세 개의 시추공(Vincent I – III)에서 획득된 감마 검층, 밀도 검층, 중성자 검층, 전기비저항 검층, 음파 검층 및 3차원 탄성파 자료가 존재한다.

암상 분류 모델의 훈련자료로 사용하기 위해 물리검층 자료 분석을 수행하였고, 이를 통해 탄성 물성 값인 P-impedance (Zp), S-impedance (Zs), Lambda-Rho (LR), Mu-Rho (MR), Vp/Vs와 암상 및 공극 유체(Facies)를 추출하였다(Fig. 4). 여기서 탄성 물성 값을 입력으로, 암상 및 공극 유체를 레이블로 사용하여 암상 분류 모델을 학습하였고, 레이블은 shale, wet sand, oil sand, gas sand로 총 4개의 클래스를 사용하였고, 훈련자료로 사용할 물리검층 자료는 불균형적인 클래스를 보인다(Fig. 5).

3차원 탄성파 자료는 near (8° ~ 19°), middle (19° ~ 30°), far (30° ~ 41°), ultra-far (41° ~ 52°) 중합 자료로 구성되어 있고, AVO (Amplitude Versus Offset) 역산을 수행하여 검층에서와 동일한 탄성 물성 값을 추출하였다. 본 논문에서는 Vincent I 과 III을 포함하는 탄성파 역산 결과 섹션을 테스트 데이터로 사용하였다(Fig. 6).

본 연구에서는 LightGBM 만을 이용하여 하이퍼 파라미터 최적화를 수행하였다. 이 때 사용된 파라미터는 범주형 2개, 연속형 5개로 총 7개의 하이퍼 파라미터를 사용하였고, 그 종류와 탐색 범위는 Table 1과 같고 세부 내용은 다음과 같다.

· **boosting_type**: 부스팅 방법을 결정하는 범주형 하이퍼 파라미터로 gbdt (전통적인 GBDT), dart (Dropouts meet Multiple Additive Regression Trees, Rashmi and Gilad-Bachrach, 2015), goss (Gradient-based One-Side Sampling)를 선택 가능하다. 기본값으로 주로 사용되는 gbdt는 가장 안정적인 부스팅 방법이고, dart는 앙상블 과정에서 특정 트리를 제거하여 과적합을

Table 1. Hyperparameters of the LightGBM model

Hyperparameter	Type of distribution	Value set or Range
boosting_type	Categorical	{gbdt, dart, goss}
class_weight	Categorical	{None, Balanced}
num_leaves	Continuous	[7, 1000]
learning_rate	Continuous	[0.005, 0.2]
min_child_samples	Continuous	[5, 500]
reg_alpha	Continuous	[0, 1]
reg_lambda	Continuous	[0, 1]

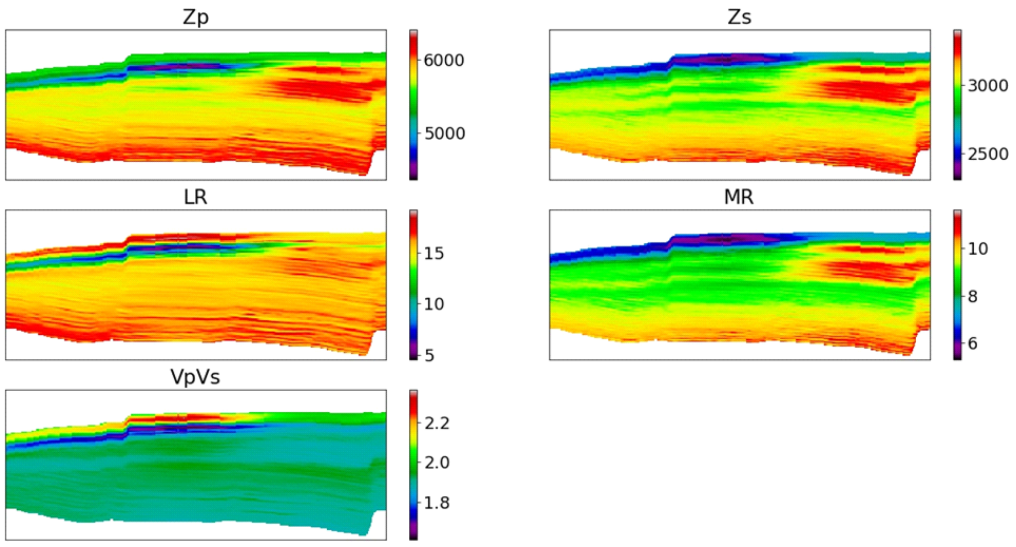

Fig. 6. Results of Seismic AVO inversion: P-impedance (Zp), S-impedance (Zs), Lambda-Rho (LR), Mu-Rho (MR), Vp/Vs.

Table 2. Selected hyperparameter values and mean AUC for models with default, random search, and Bayesian optimization

Hyperparameter	default	Random Search	Bayes optimization
boosting_type	gbdt	gbdt	dart
class_weight	None	None	balanced
num_leaves	31	64	60
learning_rate	0.1	0.024	0.005
min_child_samples	20	40	25
reg_alpha	0	0.857	0.999
reg_lambda	0	0.979	0.329
mean AUC (k=5)	0.971	0.994	0.997

방지하는 드랍아웃 방식을 적용한다. goss는 높은 그래디언트를 가지는 샘플은 유지하고 낮은 그래디언트를 가지는 샘플은 일정 비율만 무작위로 샘플링 하는 방식을 취하여 빠른 속도로 로스(loss)가 수렴하는 장점을 가진다.

- **class_weight**: balanced로 설정 할 경우 클래스가 불균형을 이룰 때 적은 양의 클래스에 보다 큰 가중치를 부여한다.
- **learning_rate**: 반복을 통한 학습에서 각 결정트리에 주어지는 가중치
- **num_leaves:** 결정트리가 가질 수 있는 최대 잎사귀의 수.
- **min_child_samples:** 하나의 잎사귀가 가질 수 있는 최소의 샘플 수.
- **reg_alpha, reg_lambda:** 각각 L1 및 L2 정규화에 사용되는 가중치.

최적의 하이퍼 파라미터를 결정하기 위해 베이지안 최적화와 k겹 교차검증을 실시하였고, 이 때 k=5의 겹수를 사용하였다. 또한 검증 자료 생성 시 불균형적인 클래스에서 샘플링 편향 현상이 일어나는 것을 방지하기 위해 계층적 샘플링(stratified sampling) 기법을 적용하였고, 성능 측정 지표로는 ROC (Receiver Operating Characteristic) 곡선의 AUC (Area Under the ROC Curve)를 사용하였다. 총 300회의 반복을 실시하였으며 Intel Core i9-9820X, 3.30 GHZ, Nvidia RTX super*2개 사양의 컴퓨터로 약 1시간 30분이 소요되었다.

베이지안 최적화의 성능을 확인하기 위해 LightGBM의 기본(default) 하이퍼 파라미터를 사용한 모델 및 랜덤 탐색을 통해 얻어진 모델과 비교하였다. 공정한 비교를 위해 랜덤 탐색 시 하이퍼 파라미터의 탐색 범위와 탐색 횟수를 베이지안 최

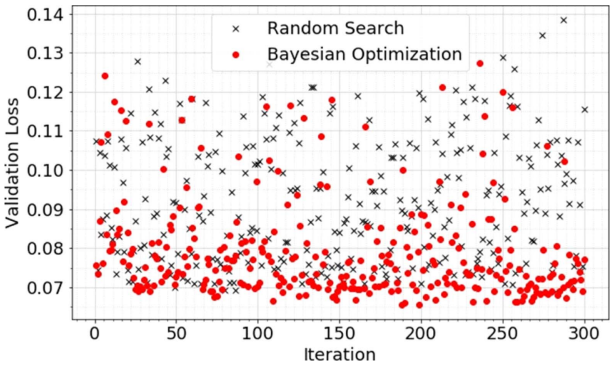

Fig. 7. Validation loss vs. iteration plots for random search and Bayesian optimization.

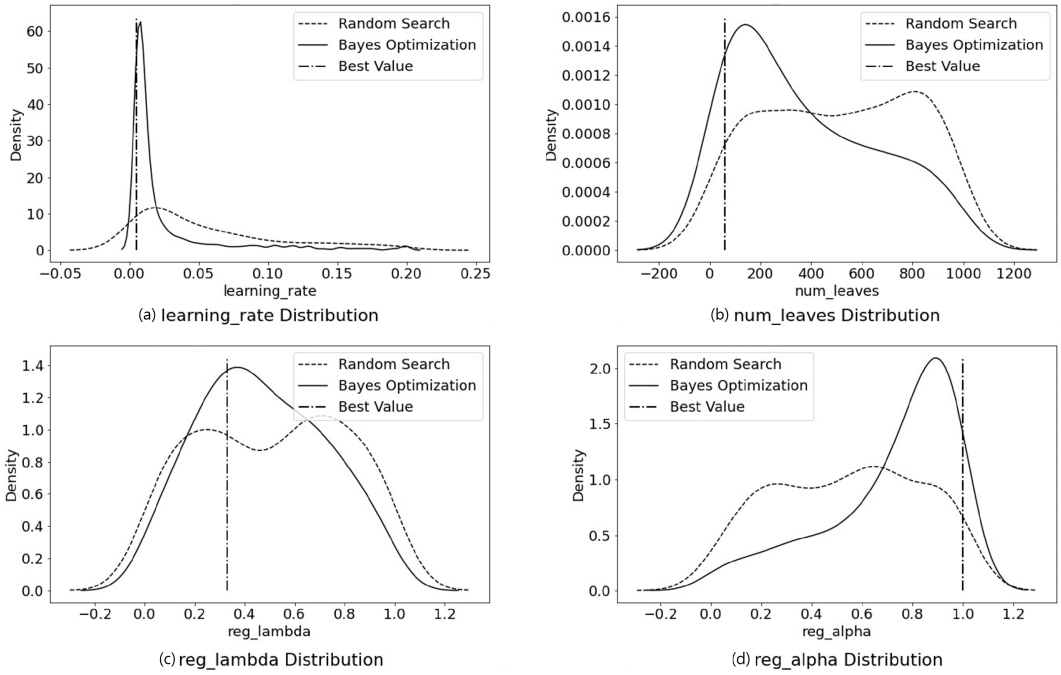

Fig. 8. Categorical hyperparameter distributions of boosting types and class weight selected from 300 iterations of random search and Bayesian optimization.

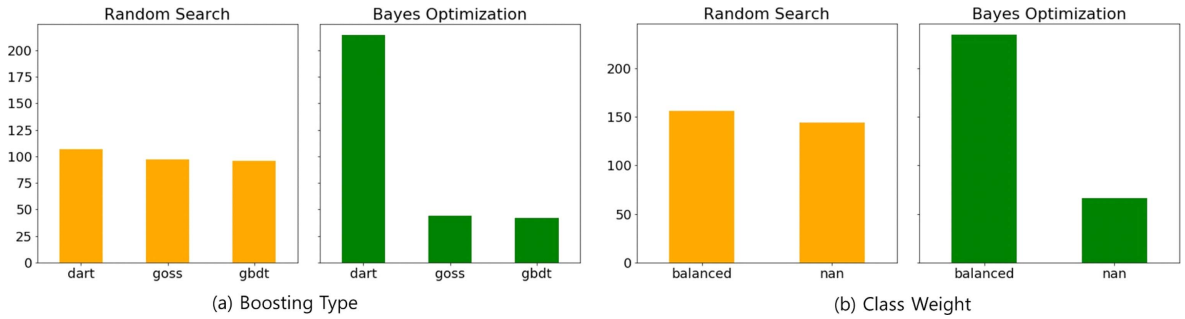

Fig. 9. Continuous hyperparameter distributions of (a) learning_rate, (b) num_leaves, (c) reg_lambda, and (d) reg_alpha selected from 300 iterations of random search and Bayesian optimization. The best hyperparameter values of Bayesian optimization are indicated by dashed lines.

적화와 동일하게 설정하였고, 동일한 k겹 교차검증(k=5)에 대한 AUC 값을 비교하였다.

먼저 5겹의 검증자료에 대한 평균 AUC 성능은 베이지안 최적화, 랜덤 탐색, 기본값 순으로 나타났고, 베이지안 최적화와

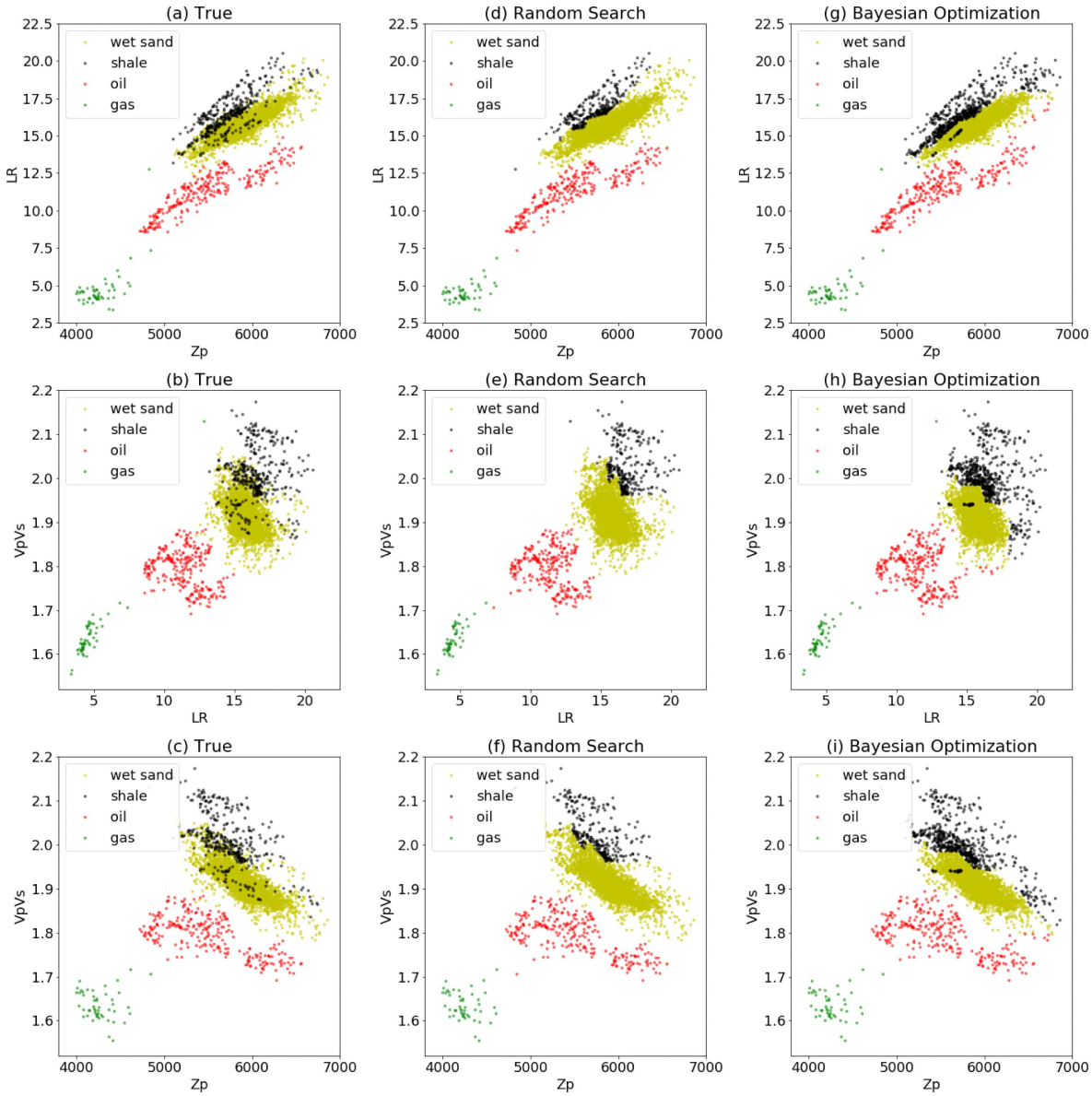

Fig. 10. Comparison of training data and predictions from final models of random search and Bayesian optimization.

회라는 비교적 많은 수의 반복 탐색을 수행할 경우 랜덤 탐색 또한 효율적으로 작동한다는 것을 보여준다. 단, 베이지안 최적화와 랜덤 탐색의 반복횟수에 따른 성능을 비교해 보았을 때, Fig. 7에서 보이는 바와 같이 대부분의 반복횟수에서 베이지안 최적화가 안정적으로 높은 성능을 보이는 것을 확인할 수 있다.

다음으로 총 300번의 반복 탐색 시 선정된 하이퍼 파라미터들의 분포를 확인해 보았다(Fig. 8, 9). 먼저 전체적인 분포의 경향을 살펴보면, 랜덤 탐색은 균일하게 무작위 값을 선택하는 반면, 베이지안 최적화는 특정 값에 치우친 값을 선정하는 경향을 보인다. 이는 베이지안 최적화의 경우 사전 지식을 반영하여 하이퍼 파라미터가 선택되는 원리로 해석될 수 있고, 그 탐색 범위가 베이지안 최적화를 통해 선택된 최적의 하이퍼 파라미터값(Table 2)과 가깝게 분포하는 것을 볼 때, 최적화가 성공적으로 진행된다는 것을 확인할 수 있다. 특히 범주형 하이퍼 파라미터의 경우, 랜덤 탐색은 균일한 분포를 보이는 반면, 베이지안 최적화는 부스팅 방식으로 DART를, 클래스 가중치로 balanced를 주로 선택하는 것을 볼 수 있다(Fig. 9). DART 부스팅 방식은 일반적으로 시간이 많이 소요되지만 대부분의 문제에서 높은 성능을 보인다는 점과, 본 훈련자료가 클래스의 불균형을 보인다는 점을 고려했을 때, 베이지안 최적화가 훈련자료에 맞춰 의미 있는 하이퍼 파라미터를 자동으로 선택한다고 해석할 수 있다.

다음으로 베이지안 최적화와 랜덤 탐색을 통해 얻어진 하이퍼 파라미터(Table 2)를 이용하여 최종 암상 분류 모델을 결정하였다. 본 연구에서는 비교적 적은 양의 데이터를 사용하기 때문에 모든 훈련 데이터를 사용하여 재학습하는 방식을 선택했다. 최종 LightGBM 분류 모델에서 특성 중요도(feature importance)를 측정하였고, 총 5개의 특성 중 중요도는 Vp/Vs, Lambda-Rho (LR), P-impedance (Zp), S-impedance (Zs), Mu-Rho (MR) 순으로 나타났다. 이들 중 특성 중요도가 높은 Vp/Vs, LR, Zp 만을 이용하여 훈련자료의 예측 값을 시각화한 결과는 Fig. 10과 같다. 훈련자료인 Figs. 10a-c에서 볼 수 있듯이 oil sand와 gas sand의 분포에 비해, wet sand와 shale의 분

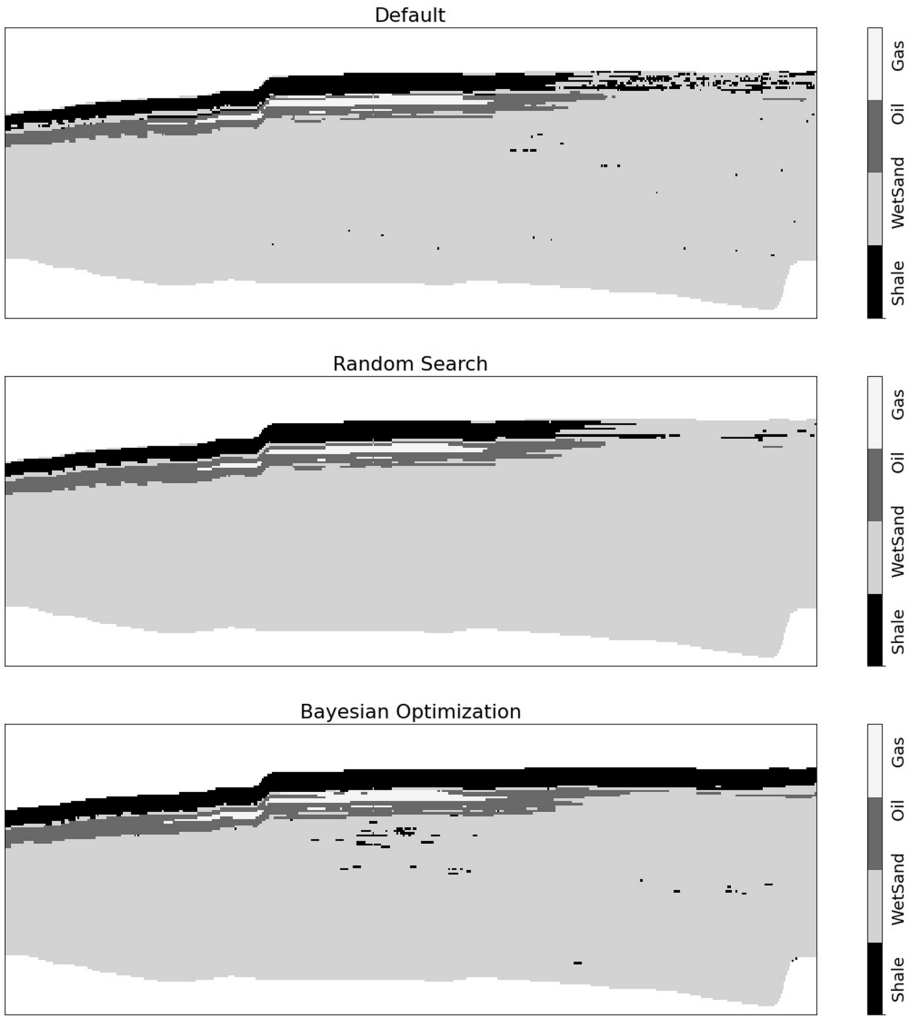

Fig. 11. Facies predictions of trained models with default hyperparameters, random search and Bayesian optimization in the seismic area.

포는 중복되는 부분이 상당량 존재하여 두 분포의 분리가 쉽지 않음을 예상할 수 있다. 이에 대해 랜덤 탐색(Figs. 10d-f)과 베이지안 최적화(Figs. 10g-i)의 예측 결과를 비교했을 때, 베이지안 최적화의 최종 모델이 상대적으로 높은 정확도로 wet sand와 shale의 분포를 분리하는 것을 확인할 수 있다.

마지막으로 기본값, 랜덤 탐색, 베이지안 최적화를 이용한 최종 암상 분류 모델을 탄성파 역산 결과(Fig. 6)에 적용하여 탄성파 영역에서의 암상 및 공극 유체 분포를 예측하였다(Fig. 11). 세 결과 모두 훈련 단계에서의 높은 검증 성능을 보인 만큼, 셰일, 가스샌드, 오일샌드의 분포가 유사하게 예측되었다. 단, 베이지안 최적화의 경우 훈련자료의 예측 결과와 유사하게 상대적으로 많은 양의 셰일 분포를 보였다.

결 론

하이퍼 파라미터 최적화는 자동 머신러닝(AutoML)의 한 분야로 모델 학습 이전에 설정되는 매개 변수인 하이퍼 파라미터의 최적값을 자동으로 결정하는 방법이다. 본 연구에서는 물리탐사 분야의 암상 분류 문제에 Tree Parzen Estimator (TPE) 기반의 베이지안 최적화 기법을 적용하여 최적의 하이퍼 파라미터를 도출하는 실험을 수행하였다. 클래스의 불균형을 보이는 부족한 훈련자료에서 모델 검증에 대한 신뢰도를 높이기 위한 방안으로 베이지안 최적화와 k겹 교차검증을 이용한 프레임워크를 제안하였고, 이를 Vincent field 자료에 적용하여 제안된 프레임워크의 효율성을 검증하였다. 지도학습 기반 암상 분류 모델의 알고리즘으로는 트리 계열 중 하나인 LightGBM을 사용하였고, 랜덤 탐색을 통해 얻어진 최종 모델과 비교하였을 때 베이지안 최적화가 높은 검증 AUC 성능을 보이는 것을 확인하였다. 마지막으로, 물리검층 자료를 통해 학습된 최종 모델을 탄성파 AVO 역산 결과에 적용하였고, 이를 통해 탄성파 영역에서의 shale, wet sand, oil sand, gas sand의 분포를 예측하였다. 비록 본 연구에서는 암상 분류 예제만을 이용해서 베이지안 최적화의 효율성을 검증하였지만, 이 기법은 대부분의 지도학습 기반 모델의 성능을 향상시키는 방법으로 사용될 수 있으며, 하이퍼 파라미터 튜닝 프로세스에 소요되는 시간을 절약 할 수 있다, 따라서 제안된 TPE 기반 베이지안 최적화를 이용한 예측 프레임워크가 물리탐사 분야의 다양한 문제에 활용되기를 기대한다.

감사의 글

본 연구는 한국지질자원연구원의 주요사업 지오빅데이터 구축 및 지질자원 분야 GeoAI 활용 플랫폼 개발(GP2020-031)의 지원으로 수행되었으며, 이에 감사드립니다. 또한 현장자료를 제공해주신 SK innovation에 감사드립니다.

References

Araya-Polo, M., Jennings, J., Adler, A., and Dahlke, T., 2018, Deep-learning tomography, *Lead Edge*, **37(1)**, 58-66, doi: 10.1190/tle37010058.1.

Baldwin, J. L., Bateman, R. M., and Wheatley, C. L., 1990, Application of a neural network to the problem of mineral identification from well logs, *The Log Analyst*, **31(05)**, 279-293.

Bergstra, J. S., Bardenet, R., Bengio, Y., and Kégl, B., 2011, Algorithms for hyper-parameter optimization, *Adv. Neural. Inf. Process. Syst.*, 2546-2554.

Choi, J., Yoon, D., Lee, S., and Byun, J., 2019, Petrofacies characterization using best combination of multiple elastic properties, *J. Pet. Sci. Eng.*, **181**, doi: 10.1016/j.petrol.2019.06.025.

Choi, J., Kim, B., Kim, S., and Byun, J., 2017, Probabilistic facies analysis using 3D crossplot of stochastic forward-modeling results, *87th Ann. Internat. Mtg. Soc. Expl. Geophys., Expanded Abstracts*, 3077-3081, doi: 10.1190/segam2017-17790996.1.

Delfiner, P., Peyret, O., and Serra, O., 1987, Automatic determination of lithology from well logs, *SPE Formation Evaluation*, **2(03)**, 303-310, doi: 10.2118/13290-PA.

Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., and Greenspan, H., 2018, Synthetic data augmentation using GAN for improved liver lesion classification, *2018 IEEE 15th Int. Symp. Biomed. Imaging*, 289-293, doi: 10.1109/ISBI.2018.8363576.

Jones, D. R., 2001, A taxonomy of global optimization methods based on response surfaces, *Journal of Global Optimization*, **21(4)**, 345-383.

Kanter, J. M., and Veeramachaneni, K., 2015, Deep feature synthesis: Towards automating data science endeavors, 2015 *IEEE Int. Conf. Data. Sci. Adv. Anal.*, 1-10, doi: 10.1109/DSAA.2015.7344858.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T. Y., 2017, LightGBM: A Highly Efficient Gradient Boosting Decision Tree, *Adv. Neural Infor. Process. Syst.*, **30**, 3149-3157.

Klein, A., Falkner, S., Bartels, S., Hennig, P., and Hutter, F., 2017, Fast Bayesian optimization of machine learning hyperparameters on large datasets, *International Conference on Artificial Intelligence and Statistics (AISTATS 2017)*, 528-536, doi: 10.1214/17-EJS1335SI.

Lee, S., Choi, J., Yoon, D., and Byun, J., 2018, Automatic labeling strategy in semi-supervised seismic facies classification by integrating well logs and seismic data, *88th Ann. Internat. Mtg. Soc. Expl. Geophys., Expanded Abstracts*, 14-19, doi: 10.1190/segam2018-2998604.1.

Li, H., Yang, W., and Yong, X., 2018, Deep learning for ground-roll noise attenuation, *88th Ann. Internat. Mtg. Soc. Expl. Geophys., Expanded Abstracts*, 14-19, doi: 10.1190/segam2018-2981295.1.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and

Talwalkar, A., 2018, Hyperband: A novel bandit-based approach to hyperparameter optimization, *J. Mach. Learn. Res.*, **18**, 1-52.

Liu, H., Simonyan, K., and Yang, Y., 2018, Darts: Differentiable architecture search, *arXiv preprint arXiv: 1806.09055.*

Mockus, J., 2012, Bayesian approach to global optimization: theory and applications, Springer Science & Business Media, **37**.

Nguyen, H. P., Liu, J., and Zio, E., 2020, A long-term prediction approach based on long short-term memory neural networks with automatic parameter optimization by Tree-structured Parzen Estimator and applied to time-series data of NPP steam generators. *Appl. Soft Comput.*, **89**, 106116, doi: 10.1016/j.asoc.2020.106116.

Oh, S., Noh, K., Yoon, D., Seol, S. J., and Byun, J., 2018, Salt delineation from electromagnetic data using convolutional neural networks, *IEEE Geosci. Remote Sens. Lett.*, **16(4)**, 519-523, doi: 10.1109/LGRS.2018.2877155.

Park, J., Yoon, D., Seol, S. J., and Byun, J., 2019, Reconstruction of seismic field data with convolutional U-Net considering the optimal training input data, *89th Ann. Internat. Mtg. Soc. Expl. Geophys., Expanded Abstracts*, doi: 10.1190/segam2019-3216017.1.

Rashmi, K. V., and Gilad-Bachrach, R., 2015, DART: Dropouts meet Multiple Additive Regression Trees, *Artificial Intelligence and Statistics*, 489-497.

Snoek, J., Larochelle, H., and Adams, R. P., 2012, Practical Bayesian optimization of machine learning algorithms, *Adv. Neural Infor. Process. Syst.*, 2951-2959.

Wolpert, D. H., and Macready, W. G., 1997, No free lunch theorems for optimization, *IEEE Trans. Evol. Comput.*, **1(1)**, 67-82.

Wrona, T., Pan, I., Gawthorpe, R. L., and Fossen, H., 2018, Seismic facies analysis using machine learning, *Geophysics*, **83(5)**, O83-O95.

Yoon, D., Yeeh, Z., and Byun, J., 2020, Seismic Data Reconstruction Using Deep Bidirectional Long Short-Term Memory with Skip Connections, *IEEE Geosci. Remote Sens. Lett.*, 1-5, doi: 10.1109/LGRS.2020.2993847.

Yoon, D., Kim, S., Kim, J., Park, G., Park, H., Byun, J., Suh, J., Lee, C., Jang, I., Jo, S., and Choi, Y., 2018, *Introduction of Resource Engineering with Machine Learning*, CIR press, 377-396 (in Korean).

Zoph, B., and Le, Q. V., 2016, Neural Architecture Search with Reinforcement Learning, *arXiv preprint arXiv:1611.01578.*