J. Soil Groundw. Environ. Vol. 21(3), p. 21~34, 2016
http://dx.doi.org/10.7857/JSGE.2016.21.3.021
ISSN 1598-6438 (Print), ISSN 2287-8831 (Online)

≪ **Research Paper** ≫

다양한 기계학습 기법의 암상예측 적용성 비교 분석

정진아 · 박은규*
경북대학교 지질학과

Comparative Application of Various Machine Learning Techniques for Lithology Predictions

Jina Jeong · Eungyu Park*
Department of Geology, Kyungpook National University, Daegu, Korea

ABSTRACT

In the present study, we applied various machine learning techniques comparatively for prediction of subsurface structures based on multiple secondary information (i.e., well-logging data). The machine learning techniques employed in this study are Naive Bayes classification (NB), artificial neural network (ANN), support vector machine (SVM) and logistic regression classification (LR). As an alternative model, conventional hidden Markov model (HMM) and modified hidden Markov model (mHMM) are used where additional information of transition probability between primary properties is incorporated in the predictions. In the comparisons, 16 boreholes consisted with four different materials are synthesized, which show directional non-stationarity in upward and downward directions. Futhermore, two types of the secondary information that is statistically related to each material are generated. From the comparative analysis with various case studies, the accuracies of the techniques become degenerated with inclusion of additive errors and small amount of the training data. For HMM predictions, the conventional HMM shows the similar accuracies with the models that does not relies on transition probability. However, the mHMM consistently shows the highest prediction accuracy among the test cases, which can be attributed to the consideration of geological nature in the training of the model.

Key words : Secondary information, Well-logging, Subsurface prediction, Machine learning

1. 서 론

지하 구성 물질의 공간적 분포에 대한 불확실성을 적절하게 규명하는 것은 지하 매질을 대상으로 이루어지는 모든 연구 및 사업의 과정에서 가장 중요한 요소 중 하나라고 할 수 있다. 특히, 지하를 구성하는 물질들의 특성을 구분하고 이들의 분포경계를 보다 상세하게 규명함으로써 대수층 내 지하수의 유동 및 오염물질의 거동 그리고 저류층 내 다중 유체의 거동 등을 보다 정확하게 예측할 수 있다(Bwoling et al., 2005, 2007; Thompson and Gelhar, 1990; Guven et al., 1992; Zheng and Gorelick, 2003; Liu et al, 2004). 지하매질의 공간적 분포를 규명하기 위한 가장 효율적인 방법은 시추를 통해 직접적으로 구성 매질을 확인 하고 이를 다양한 통계적 기법에 적용하여 공간적 분포를 예측하는 것이나 일반적으로 시추코어를 통한 직접정보의 획득에는 많은 비용이 소요된다. 또한 시추는 하였으나 코어 회수가 이루어지지 않은 관측정이 존재하는 경우도 다수이다. 이러한 경우 구성매질의 공간적 분포형태를 예측하기 위해 이들의 물리적 특성에 기초한 간접정보(e.g., 지구물리탐사자료)가 주로 활용되며 이것은 직접적인 시추를 통한 방법보다 비용 효율적이고 코어가 존재하지 않은 경우, 구성매질을 예측 할 수 있는 유일한 대안이다(Sandham and Leggett, 2003).

간접정보를 활용하여 지하매질을 규명하기 위한 기존의 연구는 크게 graphical method와 통계적 추론법으로 나눌 수 있다. 그 중 graphical method(Pickett, 1963; Gassaway

*Corresponding author : egpark@knu.ac.kr
Received : 2016. 1. 22 Reviewed : 2016. 3. 15 Accepted : 2016. 3. 29
Discussion until : 2016. 8. 31

et al., 1989)는 간접정보들을 대비도표(crossplot) 상에 도시하여 매질을 분류하는 기법으로 이는 빠르고 직관적 분류를 이행하는 장점이 있다. 통계적 추론법은 보다 다양한 간접자료를 활용한 분석을 위해 이용되는 기법이며 이에는 주성분분석(principle component analysis) 및 군집법(Wolff and Pelissier-Combescure, 1982)과 판별함수분석(Busch et al., 1987; Delfiner et al., 1987) 기법 등이 존재한다. 그러나 이러한 방법들은 분석에 있어 전문가의 주관적 판단이 일부 개입되어야하고 다양한 종류의 간접정보에 대한 객관적인 통합분석에 어려움이 있으며 다량의 정보를 필요로 한다는 단점이 있다(Rogers et al., 1992; Fung et al., 1995; Maiti et al, 2007; Chang et al, 1997).

기계학습은 정보들의 통계적 특성을 자동적이고 객관적으로 학습하여 이들을 군집화하거나 특정 클래스로 분류는 기법이다. 해당 기법은 분류를 위해 이용되는 정보들이 다양하게 연계되어 있고 복잡한 통계적 분포 특성을 보임에 따라 이들 간의 상관성을 직관적으로 분석하고 분류에 고려할 수 없는 경우 유용한 방법이다. 이러한 장점은 다중 간접정보들의 통합적 및 객관적 분석에 제약이 존재하던 기존 구성매질 예측법에 대한 새로운 대안이 되었다. 특히, 기계학습 기법 중에서도 기존 자료로부터 특성정보와 이들의 클래스를 함께 학습한 후 새로운 정보에 대해 클래스를 부여하는 교사학습 기법(supervised learning method)을 적용하여 다양한 간접정보로부터 암상을 예측하는 연구가 활발하게 이루어졌다(Fung et al., 1995; Chang et al., 1997; Schumann, 1997; Benaouda et al., 1999; An, 2000; Sandham and Leggett, 2003; Maiti et al., 2007; Smirnoff et al., 2008). 이들은 최우추정법(maximum likelihood estimation), 인공 신경망(artificial neural network), 지지벡터기반 분류기(support vector machine) 등의 학습기를 적용하여 지구 물리탐사자료를 자동으로 학습하고 새로운 지구물리자료로부터 이들의 암상을 분류하고 있다.

탐사를 이용하여 획득한 간접정보들은 서로 확연히 다른 구성매질로부터 기인하였음에도 불구하고 탐사기기의 오차, 인접한 다른 구성매질로 부터의 간섭 등으로 인하여 이들의 특성 경계가 모호한 경우가 많다(Birch, 1961; Sobolev and Babeyko, 1994; Christensen, 1996; Schon, 1996; Bauer et al., 2003). 따라서 간접정보만을 이용한 매질의 공간적 분포 예측에 제약이 존재한다. 최근 간접정보와 구성매질 간의 통계적 특성뿐 만 아니라 지하 구성매질들의 공간 내 특정 방향으로의 전이특성(transition properties)을 추가적으로 활용하여 보다 정밀한 매질분포를 예측하기 위한 연구들이 진행되고 있다(Schumann, 2002; Eidsvik et al., 2004; Rimstad and Omre, 2013). 이들은 공통적으로 기계 학습기법 중 음성인식을 위해 주로 이용되는 순차 라벨링 알고리듬(sequence labelling algorithm)의 일종인 은닉 마르코프 모델 (hidden Markov model)을 이용하여 시추공 내 지구물리 로깅자료와 주변 시추공으로부터의 암상 간 전이확률을 이용하여 암상정보를 예측하고 있다. 그러나 앞선 연구들은 음성인식 분야에 적용되어오던 은닉 마르코프 모델을 지질학적 예측 분야에 그대로 적용함으로써 지질의 일반적인 분포특성인 지향적 비정규성(directional non-stationarity, Park, 2010)을 제대로 반영하지 않고 있으며 이를 극복하기 위해 Jeong et al.(2014)에서는 지하매질정보의 지향적 비정규성을 반영한 은닉 마르코프 모델을 제안한 바 있다.

앞서 설명한 바와 같이 간접정보만을 이용하여 지하구성 매질 분포를 예측하는 기계학습 기법, 간접정보 및 지하매질들 간의 전이정보를 모두 이용하여 예측을 실시하는 기법, 지질학적 요소들을 고려한 수정된 기계학습 기법 등 다양한 기계학습 기법들을 적용한 연구들이 존재하지만 이들의 예측능에 대한 비교 및 분석 연구는 미흡한 실정이다. 따라서 본 연구에서는 간접정보를 이용한 기계학습 기법과 간접정보 및 전이특성을 이용하는 기계학습 기법의 매질 예측능에 대한 비교를 실시하였으며, 지질학적 요소를 반영할 수 있도록 고안된 학습 기법과의 비교 및 분석 또한 실시하였다.

전반적으로 본 연구에서는 다루어질 내용은 먼저 이용된 다양한 기계학습 기법에 대한 이론적 배경에 대한 설명이 이루어질 것이다. 그 후, 설명된 기법들의 예측능을 검증하기 위해 고안된 가상의 지하구성 매질 자료 및 시추공 내 암상로깅 자료 및 이에 해당하는 지구물리탐사자료와 같은 간접자료에 대한 설명이 이루어질 것이며 이에 기계학습기법을 적용한 예측결과에 대한 토의가 차례로 진행될 것이다.

2. 수학적 배경

본 연구에서는 다양한 기계학습기법을 지질학적으로 활용하여 간접적으로 지하구성매질을 지시하는 정보로부터 직접적인 매질정보를 예측하고자 하였으며 이를 위해 교사학습(supervised learning)에 기반을 둔 기계학습 기법이 적용되었다. 특정 매질과 간접정보자료 간의 통계적 특성, 즉, 방출확률(emission probability)을 이용하여 새로운 간

접정보로부터 구성매질정보를 예측하기 위해 적용된 기계학습기법으로 나이브 베이즈(Naive Bayes), 인공 신경망(Artificial Neural Network), 지지 기반 학습(Support vector machine) 및 로지스틱 회귀 분류기(Logistic regression)가 이용되었으며 방출확률 뿐 만 아니라 매질간의 수직적 전이확률(transition probability)을 추가로 활용하기 위한 방법으로 은닉마르코프 모델(Hidden Markov model) 및 지질학적 요소가 고려된 은닉 마르코프 모델로는 나이브 베이즈 학습 기반의 수정 보완된 은닉마르코프 모델(Jeong et al., 2014)이 이용되었다. 지질학적 적용을 위해 이용된 각 기계학습 모델에 대한 설명은 아래와 같다.

2.1. 나이브 베이즈 분류기(Naive Bayes classification)

일반적으로 분류기들은 특정 간접정보 ($\mathbf{x}$)에 대한 각 매질(c_k) 별 조건부 확률(conditional probability, $p(c_k|\mathbf{x})$)을 계산하고 이를 통해 매질분류를 이행한다. 그러나 나이브 베이즈 분류기는 매질분류에 $p(c_k|\mathbf{x})$를 직접적으로 이용하는 대신 각 매질에서 간접지시정보가 발생할 확률밀도인 우도(likelihood, $p(\mathbf{x}|c_k)$)를 이용하여 매질을 구분하는 생성적 접근법(generative approach)의 일종으로써 베이즈 이론(Bayes theorem)에 기반한 $p(c_k|\mathbf{x})$와 $p(\mathbf{x}|c_k)$의 관계(식 (1))를 이용한다(Kendall and Stuart, 1977).

$$p(c_k|\mathbf{x}) \propto p(\mathbf{x}|c_k)p(c_k),\ k=1,\ldots,C \tag{1}$$

나이브 베이즈 분류기는 분류를 위해 우도밀도의 형태에 대한 가정이 필요하며 최우추정법(maximum likelihood estimation)을 통해 분포형태를 정의하는 파라미터들을 학습한다. 본 연구에서는 각 매질에 대한 간접정보들의 우도밀도 형태가 가우시안분포(Gaussian distribution)임을 가정하였으며 이에 따라 각 매질에 대한 간접정보들의 평균($\mathbf{m}_k$) 및 공분산 (Σ_k)을 학습하고 다음과 같은 결정규칙을 통해 매질분류를 이행하였다(Murphy, 2012).

$$c(\mathbf{x}) = \arg min_k\{(\mathbf{x}-\mathbf{m}_k)^T\Sigma_k(\mathbf{x}-\mathbf{m}_k)+\ln|\Sigma_k|\},\ k=1,\ldots,C \tag{2}$$

2.2. 인공 신경망(Artificial Neural Network)

인공 신경망은 인간 신경계가 시냅스의 연결강도를 스스로 조정하여 학습하는 능력을 모방하여 만든 학습기이다. 인간의 신경계 구조는 정보를 받아들이고 처리하는 뉴런과 두 신경 세포체 간의 연결 강도를 나타내는 시냅스로 구성되고 이들은 계층적 연결 구조를 가지고 있으며, 이 때, 뉴런과 연결 강도는 인공 신경망에서 각 층에서의

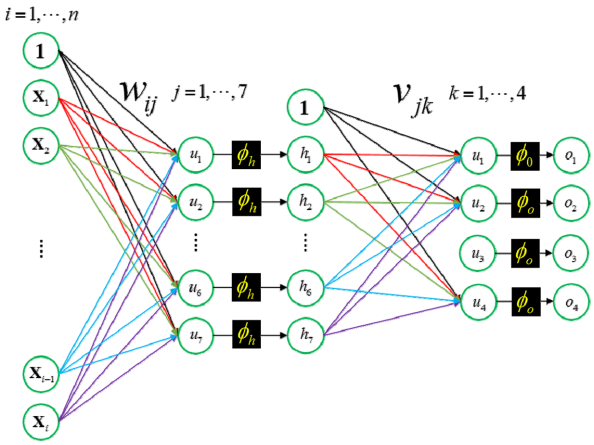

Fig. 1. The schematic diagram for artificial neural network used in this study.

노드($\mathbf{x}_i$, h_j 및 c_k) 및 다음 층 노드와의 가중치(w_{ij} 및 v_{jk})를 나타낸다(Fig. 1). 또한 임계치 이상의 자극에 대해서만 반응을 보이는 신경세포의 특성을 모방하기 위해 인공뉴런은 특정 활성함수(activation function, ϕ_h 및 ϕ_o)를 이용한다. 이와 같은 구조를 통해 입력뉴런으로부터 받아들인 간접지시정보를 이용한 최종 매질 분류 결과(c_k)는 다음과 같은 다층 전방향 신경망(multi-layer feed Forward network) 식을 통해 결정된다(Bishop, 1995).

$$c_k = \phi_o\left(\sum_{j=1}^{M} v_{jk}\phi_h\left(\sum_{i=1}^{N} w_{ij}\mathbf{x}_i + w_{0j}\right) + v_{0k}\right) \tag{3}$$

이 때, M 및 N은 주어진 인공신경망 네트워크에서의 은닉 뉴런 개수 및 입력되는 간접지시정보의 개수를 의미한다. 해당 학습기는 예측된 매질정보(c_k)와 목적암상정보(t_k) 간의 차이에 비례하게 가중치를 단계적으로 조절하는 일반화된 델타학습규칙(generalized delta rule)에 의해 학습이 이루어진다. 본 연구에서는 입력 정보($\mathbf{x}$)가 하나씩 주어질 때 마다 가중치를 조절하는 온라인 학습을 이용하였으며 이에 따라 목적정보와 예측정보 간 차이에 관한 오류함수(E)를 $E = 1/2\,(t_k - c_k)^2$와 같이 정의 하고 이에 경사 하강법(gradient descent method)을 적용하여 가중치 조절을 실시하였다. 경사 하강법을 인공신경망에 적용하여 이루어지는 학습방법을 일반적으로 오류 역전파 알고리듬(error backpropagation algorithm)이라 부르며 이는 식 (4) 및 (5)과 같다.

$$w_{ij}^{t+1} = w_{ij}^{t} - \gamma\frac{\partial E}{\partial w_{ij}} \tag{4}$$

$$v_{jk}^{t+1} = v_{jk}^{t} - \gamma\frac{\partial E}{\partial v_{jk}} \tag{5}$$

여기서 $\frac{\partial E}{\partial w_{ij}}$ 및 $\frac{\partial E}{\partial v_{jk}}$ (식 (6) 및 (7))는 오류함수를 이용한 가중치 수정정도를 나타내며 γ는 가중치의 변화정도를 조절하는 학습률을 의미한다.

$$\frac{\partial E}{\partial w_{ij}} = -\phi_h^{'}\left(\sum_{i=1}^{N} w_{ij}\mathbf{x}_i + w_{0j}\right)\sum_{k=1}^{C} v_{jk}\phi_o^{'}\left(\sum_{j=1}^{M} v_{jk}h_j + v_{0k}\right)(t_k - c_k)\mathbf{x}_i \tag{6}$$

$$\frac{\partial E}{\partial v_{jk}} = -\phi_o^{'}\left(\sum_{j=1}^{M} v_{jk}h_j + v_{0k}\right)(t_k - c_k)h_j \tag{7}$$

2.3. 지지벡터기반 분류기(Support Vector Machine)

지지벡터기반 분류기는 선형의 결정경계를 이용함으로써 다른 학습기를 이용한 분류 시 발생 할 수 있는 과다적합(over fitting) 현상을 피하고 일반화오차를 줄이기 위해 고안된 교사학습 분류기의 일종이다(Cortes and Vapnik, 1995). 이는 q차원의 간접지시정보 공간상에서 각 매질에 대한 간접정보 간의 간격(margin)을 최대화하는 결정경계를 학습한다. 마진을 최대로 하는 경계를 찾기 위해 먼저, 결정경계에 가장 가까운 곳에 위치한 간접지시정보를 지지벡터(support vector)로 정의하고 이들을 통해 결정경계의 파라미터를 학습하여 초평면의 결정경계를 구성한다. 일반적으로 지지벡터기반 분류기는 선형 결정경계를 이용한 분류에 따른 단점을 보완하고 비선형 결정경계를 이용한 분류기의 장점을 획득하기 위해 커널법(kernel method)을 이용하여 학습할 간접지시정보를 고차원의 데이터 공간으로 매핑(mapping)시킨 후 선형의 초평면결정경계를 이용하여 매질분류를 이행한다. 커널(k)을 이용한 지지기반 분류기를 통해 새로운 간접정보($\mathbf{x}$)의 매질정보(c_k)를 결정하는 규칙은 다음 식과 같다(Cortes and Vapnik, 1995).

$$c_k = \text{sign}\left(\sum_{i=1}^{J} \hat{\alpha}_i t_i k(\mathbf{x}_i, \mathbf{x}) + \frac{1}{J}\sum_{i=1}^{J}\left(t_i - \sum_{j=1}^{J} \hat{\alpha}_j t_j k(\mathbf{x}_j, \mathbf{x}_i)\right)\right),$$
$$i = 1,\ldots,J, j = 1,\ldots,J \tag{8}$$

여기서 $\hat{\alpha}$는 이차계획법(quadratic programming)을 이용한 목적함수의 최적화를 통해 산정된 라그랑제 곱수(Lagrange multipliers)를 의미하며 이는 J개 지지벡터에 대해서만 0이 아닌 값을 가진다. t는 지지벡터로 결정된 간접지시정보들에 대한 실제 매질정보를 의미한다.

2.4. 로지스틱 회귀 분류기(Logistic Regression classification)

로지스틱 회귀 분류기는 종속변수(dependent variable)가 범주형 변수(categorical variable)일 경우 이용되는 로지스틱 회귀모델(logistic regression model)을 이용하여 간접지시정보($\mathbf{x}$)들을 학습시킨 후 학습된 회귀모델을 특정 간접정보($\mathbf{x}$)에 대한 각 매질(c_k) 별 조건부 확률($p(c_k|\mathbf{x})$)로 이용하여 새로운 자료에 대한 암상을 분류하는 학습기이다. 이용되는 로지스틱 회귀모델은 식 (9)과 같고 이를 통해 식 (10)과 같은 규칙을 통해 암상을 분류한다(Murphy, 2012).

$$L_\theta(\mathbf{x}) = \frac{1}{1+e^{-\theta^T\mathbf{x}}} \tag{9}$$

$$c_k = \begin{cases} 1, L_\theta \geq 0.5 \\ 0, L_\theta < 0.5 \end{cases} \tag{10}$$

이 때, 로지스틱 함수의 형태를 결정해 주는 파라미터(θ)들은 다음의 로지스틱 회귀에 대한 목적함수(logistic regression cost function)를 경사 하강법을 통해 최적화함으로써 결정된다.

$$\min_\theta J(\theta) = \min_\theta\left[-\frac{1}{N}\left[\sum_{i=1}^{N} t_i \log L_\theta(\mathbf{x}_i) + (1-t_i)\log(1-L_\theta(\mathbf{x}_i))\right]\right] \tag{11}$$

이 때, N은 입력되는 정보의 개수를 의미한다.

2.5. 은닉 마르코프 모델(Hidden Markov Model)

지질학적 적용을 고려한 은닉 마르코프 모델은 앞서 언급된 생성모델의 일종인 나이브베이즈 분류기를 공간적으로 순차에 따라 연결한 형태를 지닌다. 은닉 마르코프 모델을 격자 다이어그램(trellis diagram)으로 표현할 경우 Fig. 2와 같으며 순차적인 공간 위치에서 은닉된 매질로 이루어진 은닉층(hidden layer)과 매질로부터 기원한 관찰가능한 간접정보들로 이루어져 있다. 앞서와 마찬가지로 총 C가지로 이루어진 매질의 확률변수를 $\mathbf{Z}(\{z_1,\ldots z_m\})$, 그리고 매질로부터 유래한 q 종류의 서로 다른 지구물리 탐사결과를 $\mathbf{x}$라고 하였을 때, 지구물리 탐사결과에 기초한 특정 심도 k에서의 매질은 다음의 확률식으로 표현된다.

$$p(z_k|\mathbf{x}) \propto p(z_k|\mathbf{x}) = p(\mathbf{x}_{k+1:n}|z_k)p(z_k, \mathbf{x}_{1:k}) \tag{12}$$

여기서, 확률 $p(z_k, \mathbf{x}_{1:k})$ 은 전향확률 그리고 $p(\mathbf{x}_{k+1:n}|z_k)$ 은 후향확률이라 하며, 전향확률은 다음의 α로 대치될 수 있

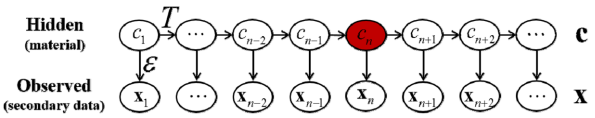

Fig. 2. A schematic diagram for conventional hidden Markov model.

다(Elliott et al., 1995).

$$\alpha_k(z_k) = \sum_{z_{k-1}=1}^{m} p(\mathbf{x}_k|z_k)p(z_k|z_{k-1})\alpha_{k-1}(z_{k-1}) \quad \text{및}$$

$$\alpha_1(z_1) = p(z_1)p(\mathbf{x}_1|z_1) \tag{13}$$

또한 전체 심도가 n개의 구획으로 이루어져 있다고 가정할 때 후향확률은 β로 대치될 수 있으며

$$\beta_k(z_k) = \sum_{z_{k+1}=1}^{m} \beta_{k+1}(z_{k+1})p(\mathbf{x}_{k+1}|z_{k+1})p(z_{k+1}|z_k) \quad \text{및}$$

$$\beta_n(z_n) = 1 \tag{14}$$

와 같다. 따라서 은닉 마르코프 모델을 통한 예측 알고리듬 중의 하나인 전후향 알고리듬(forward-backward algorithm)에 기초하여 수직적 심도 k에서 관찰될 매질이 l일 우도(likelihood)는

$$p(z_k = l|\mathbf{x}) = \alpha_k(z_k = l) \times \beta_k(z_k = l) \tag{15}$$

와 같이 주어진다. 그러나 이러한 방법을 통한 매질의 예측은 지질학적 암층서에서 흔히 나타나는 지향적 비정규성을 효과적으로 설명할 수 없으므로 Jeong et al.(2014)은 은닉 마르코프 모델의 효과적인 지질학적 적용을 위하여 전후향 알고리듬을

$$\alpha_k^{(u,d)}(z_k) = \sum_{z_{k'}=1}^{m} p(\mathbf{x}_k|z_k)p(z_k|z_{k'})\alpha_{k'}(z_{k'}) \quad \text{및}$$

$$\beta_k^{(u,d)}(z_k) = \sum_{z_{k'}=1}^{m} \beta_{k'}(z_{k'})p(\mathbf{x}_{k'}|z_{k'})p(z_{k'}|z_k) \tag{16}$$

와 같이 개선하였으며, 위의 식에서 하향 또는 상향방향을 정방향으로 했을 때 $z_{k'}$은 각각 z_{k-1} 및 z_{k+1}이 되며 $z_{k'}$은 각각 z_{k+1} 및 z_{k-1}이 된다. 따라서, 최종적으로 수직적 심도 k에서 관찰될 매질이 l일 우도는

$$p(z_k = l|\mathbf{x}) = \alpha_k^d(z_k = l) \times \beta_k^d(z_k = l) \times \alpha_k^u(z_k = l) \times \beta_k^u(z_k = l) \tag{17}$$

로 결정된다.

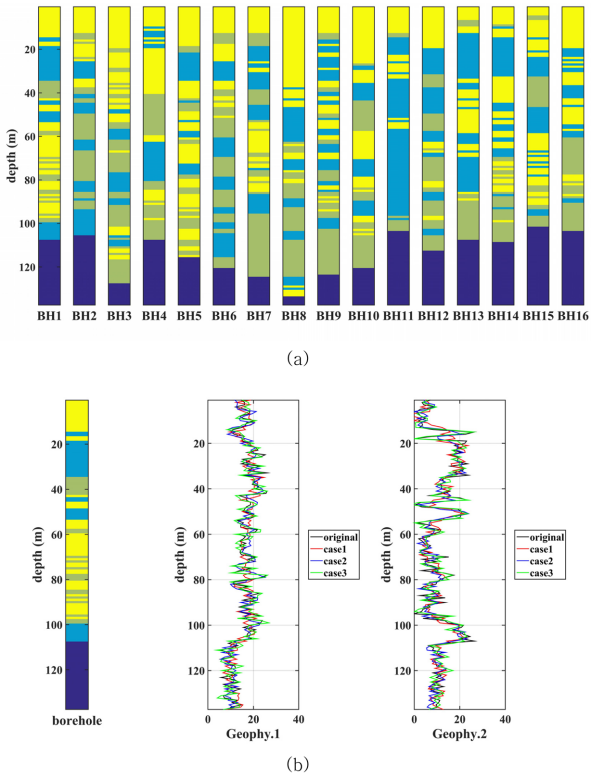

Fig. 3. The hypothetical data used in this study: (a) 16 boreholes composing four different materials with 137 m depth (Jeong et al., 2014) and (b) one example of two type of the secondary information with three difference degree of disturbance and measurement error.

3. 실험 방법

3.1. 가상 자료의 생성

방출확률만을 이용하여 간접정보로부터 매질정보를 예측하는 기계학습기법들과 방출확률 및 전이확률을 모두 활용하는 은닉 마르코프 모델의 성능을 비교하기 위하여 Fig. 3(a)와 같이 총 4개의 매질로 구성된 16개의 시추공을 설정하였으며 각 시추공에 대해 2가지 종류의 지구물리탐사자료가 존재함을 가정하였다(Fig. 3(b)). 이 때, 시추공은 지질분포의 지향적 비정규성을 반영하고 실질적인 매질들의 수직적 분포 특성과 유사한 구조를 모방하기 위하여 매질 1을 기반암으로 가정하고 시추공의 하부에만 존재하도록 설정하였고 다른 3가지의 매질은 랜덤한 순차를 따라 상부에 분포하도록 가정하였다. 지구물리탐사자료는 각 매질에 따라 특정 통계적 분포특성을 보여줌을 가정하여 4개의 암상에 대한 2종 지구물리탐사자료의 평균과 이들 간의 공분산을 Table 1에서 나타나는 바와 같이 가정하고 해당 값을 파라미터로 하는 통계적 분포를

Table 1. The statistical mean, variance, and covariance used to realize two different types of synthetic geophysical data (Jeong et al., 2014)

	Mean		Variance		Covariance
	Second. 1	Second. 2	Second. 1	Second. 2	
C1	9	10	3	2	0
C2	19	20	3	2	−1
C3	22.5	13	2.5	2	1
C4	15	5	2	2.5	−0.5

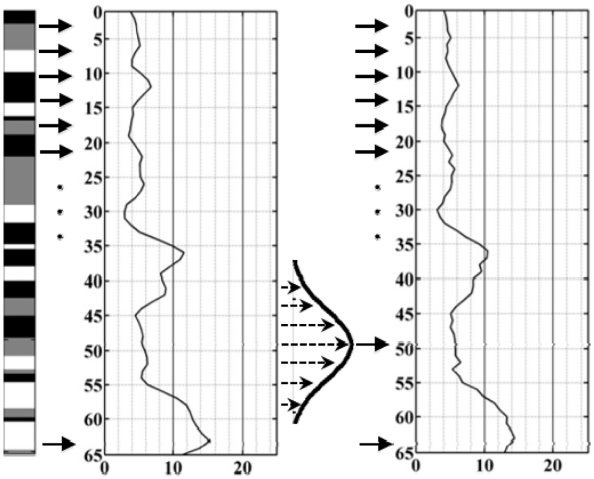

Fig. 4. Concept of degenerating method of secondary information by incorporating higher disturbance of signals.

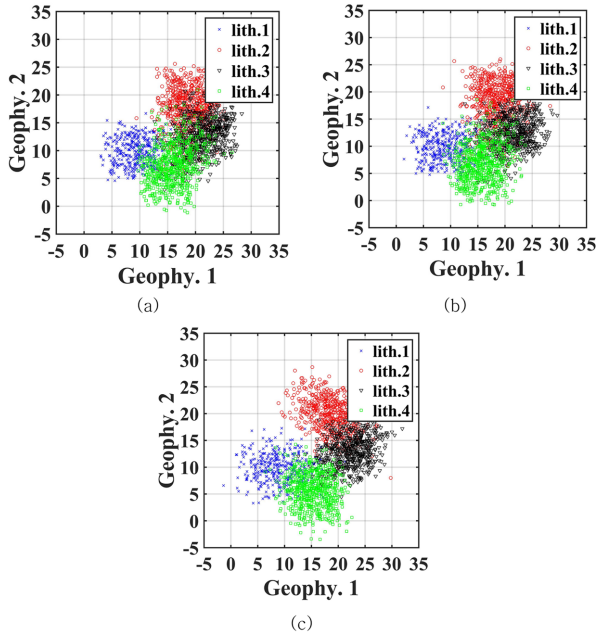

Fig. 5. The distributions of the realized secondary information in the parametric space with (a) the highest, (b) intermediate and (c) the lowest degree of errors.

이용하여 무작위로 발생시켰다. 또한 앞서 설명한 바와 같이 시추공 내 수직적 탐사를 통해 지구물리탐사자료를 획득할 때 인근 매질로부터 기인한 간접정보에 의한 교란이 발생할 수 있으며 각기 다른 매질로부터 기인한 자료 일지라도 이들의 특성 경계가 불분명한 경우가 많다. 이러한 현상을 본 연구에서 설정한 지구물리탐사자료에 반영하기 위해서 가우시안 분포(Gaussian distribution)를 보이는 컨볼루션 함수(convolution function)를 이용하여 필터링(filtering) 함으로써 상부 및 하부에 존재하는 매질들로부터의 영향을 포함시켰다(Fig. 4). 간접정보 획득 시, 이들의 교란정도가 상이할 수 있으며 이러한 양상을 반영하기 위해 본 연구에서는 총 3가지 종류의 컨볼루션 함수를 이용하여 자료를 교란시켰다. 이를 위해 분산이 각각 0.2, 0.5 및 1인 가우시안 분포 형태의 컨볼루션 함수가 이용되었으며 이로써 총 3가지 케이스의 간접정보 분포 형태가 가정되었다. 이 때, 보다 현실적인 자료 생성을 위해 탐사기기의 측정오차 또한 고려되었다. 최종적으로 생성된 가상의 간잡자료의 분표형태는 Fig. 5에서 보는 바와 같다. 그림에서 보는 바와 같이 케이스 1은 간접정보 측정 시 인근 매질로부터의 교란정도가 상당히 큰 경우로써 각 매질에 대한 간접정보 분포들이 서로 오버랩이 많이 되어 있는 경향을 보여준다. 케이스 2는 교란의 정도가 케이스 1보다 작은 경우로 설정하였으며 케이스 3은 교란의 정도가 가장 작도록 설정함으로써 각 매질별 간접정보의 군집이 다른 케이스에 비하여 확연히 분류되는 것을 확인 할 수 있다.

3.2. 예측능의 검증

설정된 시추공 내 매질 및 간접정보를 이용하여 총 2가지 분석을 통해 앞서 설명된 각 분류기의 성능을 비교 및 분석하고자 하였다. 먼저 첫 번째 분석으로 인근 매질로부터 기인하는 간접정보의 교란으로 인한 오류정도에 대한 분류기의 성능을 검증하고자 하였다. 또 다른 분석으로 분류기 학습 시 이용되는 학습 자료의 개수에 따른

각 분류기의 성능을 비교하고자 하였으며 이를 위해 1번부터 16번까지의 시추공을 검증 자료로 이용하여 매질에 대한 예측성능을 검증할 때, 1개, 2개, 4개, 8개 및 15개의 시추공을 무작위로 선택하여 학습 자료로 이용하였으며 이들에 대한 각 분류기의 비교 및 분석이 실시되었다.

3.3. 분류기 설계인자 설정

앞서 설명된 분류기들을 이용하여 간접정보로부터 매질정보를 예측하기 위해 먼저 각 분류기들을 설계하는데 필요한 설계인자들을 설정하였다. 본 연구에서의 목적인 분류기들의 성능 비교에 있어 모든 분류기들이 가장 높은 예측능을 보이는 경우를 기준으로 이들을 비교 및 분석하는 것이 타당하다. 이를 위해 학습 데이터의 일부를 분류기의 성능 검증을 위한 검증 데이터(validation data)로 분류한 후, 다양한 설계인자를 가지는 분류기에 대해 시행착오법(trial and error method)을 적용하여 학습 데이터를 통한 학습 및 검증 데이터를 이용한 오류율 계산을 반복적으로 시행한 후 최적 설계인자를 도출하고 분류기를 구성하였다. 이 때, 인공 신경망과 로지스틱 회귀 분류기 학습 시 고려하여야 할 설계인자 중 하나인 가중치의 학습률(γ)은 다른 설계인자에 비하여 학습기의 성능에 큰 영향을 미치지 않음을 고려하여 충분히 작은 값으로 고정한 후 테스트를 실시하였다.

본 연구에 이용된 기계학습 기법 중 최적 설계가 필요한 분류기는 인공 신경망, 지지벡터기반 분류기 및 로지스틱 회귀 분류기이다. 먼저 인공 신경망은 Fig. 1과 같이 2종류의 가상 간접정보자료를 이용하기 위해 2개의 입력뉴런 및 1 개의 바이어스 뉴런을 이용하였으며 입력된 데이터를 4종류의 매질로 분류하기 위해 4개의 출력뉴런을 설정하였다. 그리고 7개의 은닉뉴런 및 1개의 바이어스 은닉뉴런으로 은닉층을 구성하였다. 또한 본 연구는 4종의 매질을 분류하는 문제로써 다계층 분류(multi-class classification)에 보다 적합한 인공신경망을 구성하기 위해 출력층에서의 활성함수는 소프트맥스 함수(softmax function)을 이용하였으며 이에 따라 크로스 엔트로피(cross entropy)를 인공신경망의 목적함수로 설정하고 이를 공역 구배법(conjugate gradient method)을 이용한 최적화를 통해 학습을 실시하였다(Dunne and Campbell, 1997). 지지벡터기반 분류기는 이분류(binary classification)에 적합하도록 만들어진 학습기이다. 따라서 본 학습기를 4종의 매질을 분류하는 문제에 적용하기 위해서 일대다대응(one-vs.-all, OvA) 방법을 이용하여 학습 데이터의 매질정보를 이분류화(식 (18)) 시킨 후 분류기에 적용하였다. 이를 통해 매질의 종류마다 각 각의 지지벡터기반 분류기를 학습하여 총 4개의 분류기를 생성하고 새로운 간접정보를 이에 대해 모두 테스트 한 후 최종 매질정보를 결정하였다.

$$t_i^* = \begin{cases} 1 \text{ if } \mathbf{x}_i \in c_k \\ 0 \text{ if } \mathbf{x}_i \notin c_k \end{cases}, \quad k = 1, \ldots, C, \tag{18}$$

로지스틱 회귀 분류기는 식 (19)에서 $f(\mathbf{x})$의 형태에 따라 성능이 좌우되며 본 연구에서는 다음과 같은 이차 함수(quadratic function)을 이용하였다.

$$f(\mathbf{x}_i) = \theta_0 x^0 + \theta_1 x_1^1 + \theta_2 x_i^2 + \theta_3 x_i^1 x_i^2 + \theta_4 (x_i^1)^2 + \theta_5 (x_i^2)^2 \tag{19}$$

또한 은닉 마르코프 모델에서 간접정보와 각 매질 간의 상관확률 외에 이용되는 매질 간 전이정보는 학습 자료로 이용되는 시추공 내에서 하향방향으로 매질 간의 전이 양상을 통해 계산된 전이확률을 이용하였으며 수정된 은닉 마르코프 모델에는 하향방향으로의 전이확률 뿐만 아니라 상향방향으로의 전이확률 또한 고려되었다.

4. 결과 및 고찰

먼저 인근 구성매질에 의한 교란과 탐사자료측정 기기의 측정 신뢰도가 낮은 경우의 자료인 케이스 1번에 대한 예측능 검증을 실시하였다. Table 2는 본 연구에서 이용된 모든 분류기에 대하여 16개의 각 시추공에 대한 간접정보자료를 검증자료로 설정하고 학습 자료의 개수를 1개, 2개, 4개, 8개 및 15개로 증가시킴에 따라 총 30번의 검증 테스트를 실시 한 후 종합된 혼동행렬(confusion matrix, CM)들을 보여준다. 혼동행렬은 검증자료에 대해 예측된 구성매질과 실제 구성매질 간의 관계를 보여주는 행렬로써 열방향은 예측된 매질, 행방향으로는 실제 매질을 나타낸다. 본 연구에서는 총 4가지 매질로 구성된 시추공에 대한 검증을 실시하였으므로 총 4×4 크기의 혼동행렬을 구성하였으며 각 행렬에서의 대각요소들은 예측된 매질이 실제매질과 일치하는 개수를 의미하고 비대각 요소들은 각 구성매질에 대한 분류기의 오분류 개수를 의미한다. 또한 혼동행렬을 이용하여 각 매질에 대한 진양성률(True Positive Rate, TP, 식 (20)) 및 위음성률(False Negative Rate, FN, 식 (21))과 정확도(Accuracy, ACC, 식 (22))를 산정하였으며, 이들은 각각 다음 식을 통해 산정되었다.

$$\mathrm{TP}_i = CM_{(i,i)} \Big/ \sum_{j=1}^{C} CM_{(i,j)}, \quad i = 1, \ldots, C \tag{20}$$

Table 2. The confusion matrices and the related rate for each machine learning technique in Case 1

			Predicted Class				
			C1	C2	C3	C4	True Rate
NB	True Class	C1	9608	235	19	1448	84.95%
		C2	173	13390	2312	985	79.42%
		C3	70	2772	11852	2736	68.00%
		C4	1646	963	2466	15085	74.83%
	False Rate		16.43%	22.87%	28.81%	25.52%	
	Accuracy				75.94%		
ANN	True Class	C1	9013	234	34	2029	79.69%
		C2	227	12942	2678	1013	76.76%
		C3	76	2798	11648	2908	66.83%
		C4	1702	843	2634	14981	74.31%
	False Rate		18.20%	23.04%	31.46%	28.43%	
	Accuracy				73.88%		
SVM	True Class	C1	10376	2	0	932	91.74%
		C2	3585	11847	1111	317	70.27%
		C3	5449	2366	7891	1724	45.27%
		C4	7621	327	1829	10383	51.50%
	False Rate		61.61%	18.53%	27.14%	22.26%	
	Accuracy				61.58%		
LR	True Class	C1	8841	228	28	2213	78.17%
		C2	169	12972	2631	1088	76.94%
		C3	62	2590	11522	3256	66.10%
		C4	1258	889	2275	15738	78.07%
	False Rate		14.41%	22.23%	29.98%	29.41%	
	Accuracy				74.62%		
HMM	True Class	C1	11257	0	1	52	99.53%
		C2	1329	13028	1488	1015	77.27%
		C3	3795	1974	9341	2320	53.59%
		C4	936	934	1715	16575	82.22%
	False Rate		34.99%	18.25%	25.54%	16.97%	
	Accuracy				76.34%		
mHMM	True Class	C1	10666	44	223	377	94.31%
		C2	0	14277	1500	1083	84.68%
		C3	20	2470	12062	2878	69.20%
		C4	27	955	1846	17332	85.97%
	False Rate		0.44%	19.55%	22.83%	20.02%	
	Accuracy				82.63%		

$$\mathrm{FN}_j = \sum_{i=2}^{C} CM_{(i,j)} \Big/ \sum_{i=1}^{C} CM_{(i,j)}, \quad j = 1, \ldots, C \tag{21}$$

$$\mathrm{ACC}_i = CM_{(i,i)} \Big/ \sum_{i=1}^{C} CM_{(j,j)}, \quad i = 1, \ldots, C \tag{22}$$

1번 케이스에서 간접정보만을 이용하는 분류기의 경우, 가장 높은 정확도를 보여주는 분류기는 나이브 베이즈 분류기로써 75.94%의 정확도 값을 가진다. 다음으로 로지스틱 회귀분석(74.62%), 인공 신경망(73.88%) 및 지지벡터기반(61.58%) 분류기 순으로 높은 정확도를 보여주고 있다. 지지벡터기반 분류기의 경우 이분류에 적합하도록 고안된 분류기로써 다중 클래스를 분류하기 위해 앞서 설명한 바와 같이 일대다대응법을 이용한 학습 자료의 전처리가 필수적이다. 그러나 일대다대응법을 이용하여 임의

로 학습 자료를 이분화 하고 각 경우에 대하여 결정경계를 따로 학습한 후 최종적인 클래스 분류에 이를 종합·적용하면 여러 개의 클래스로 분류되어 선택이 모호한 경우 및 어떤 클래스로도 분류가 되지 않는 경우와 같이 미결정영역이 존재하게 된다. 이러한 이유로 인해 해당 결과에 대해서도 미결정영역 내 간접자료들이 모두 구성매질 1번으로 분류되어 매질 1번에 대한 진양성률 뿐만 아니라 매질 1번으로의 위음성률 모두 크게 나타났다. 매질 1번에 대한 높은 위음성률은 1번이 아닌 다른 매질로 예측되어야 할 많은 간접자료들이 1번으로 예측되고 있음을 의미한다. 로지스틱 회귀 분류기 또한 지지벡터기반 학습과 마찬가지로 학습 자료를 이분화 시킨 후 학습에 이용하는 학습기법이다. 그러나 지지벡터기반 분류기에서는 검증자료가 서포트벡터들을 이용하여 작성된 결정경계를 기준으로 양의 영역에 포함되는지 또는 음의 영역에 포함되는지를 확인 한 후 결과를 이분화 하여 도출하는 반면 로지스틱 회귀 분류기는 각 매질에 대한 결정경계를 확률에 기반하여 작성하여 4가지 매질에 대한 최종적 분류 결과물 또한 서로 다른 크기의 확률값으로 산출된다. 그 후 이들 값 중 가장 높은 값을 보이는 매질을 최종 매질로 산정함으로써 미결정영역이 나타나지 않고 이로써 높은 분류성능을 보여준다.

간접정보 외 매질 간 전이정보를 추가로 학습하고 분류한 결과에 대한 혼동행렬 또한 Table 2에 제시되어 있다. 먼저 지질학적 요소를 고려하지 않은 기존의 은닉 마르코프 모델의 정확도는 76.34%로 간접정보만을 이용하는 다른 분류기보다 다소 높은 예측능을 보여준다. 그러나 이전 설명된 분류기 중 가장 높은 정확도를 보여주었던 나이브 베이즈 분류기의 정확도(75.94%)와 비교 하였을 때, 은닉 마르코프 모델을 통한 정확도 상승은 굉장히 미미하며 오히려 매질 1번에 대한 위음성률은 두드러지게 상승하고 매질 3번에 대한 진양성률은 현저히 감소하였다. 그러나 3번 외 다른 매질에 대한 진양성률은 증가하고 2번, 3번 및 4번에 대한 위음성률이 상당히 낮아졌음을 통해 추가적 정보 활용이 예측능 향상에 도움이 됨을 알 수 있다. 수정된 은닉 마르코프 모델을 이용한 결과는 정확도가 82.63%으로써 상당히 향상된 수치를 보여주고 있으며 각 매질에 대한 진양성률이 모두 다른 기법들에 비해 높은 값을 보이며 위음성률 또한 현저히 낮아진 경향을 보여준다. 특히, 시추공의 하부에만 존재하는 매질 1번에 대한 위음성률은 거의 0%에 가까운 수치를 보여주고 진양성률 또한 94.31%의 높은 수치를 보여줌을 통해 매질 1번과 다른 매질 간의 분류가 상당히 정확한 것을 알 수 있으며 해당 분류기법이 지질의 지향적 비정규적 특성을 충분히 잘 반영하고 있음을 알 수 있다. 이를 통해 비록 은닉 마르코프 모델이 전이정보를 추가로 이용하여 분류를 이행하지만 지질의 지향적 비정규성에 대한 충분한 고려가 없는 경우 간접정보만을 이용한 결과와 유사하거나 근소하게 높은 예측능을 보여줌을 알 수 있다.

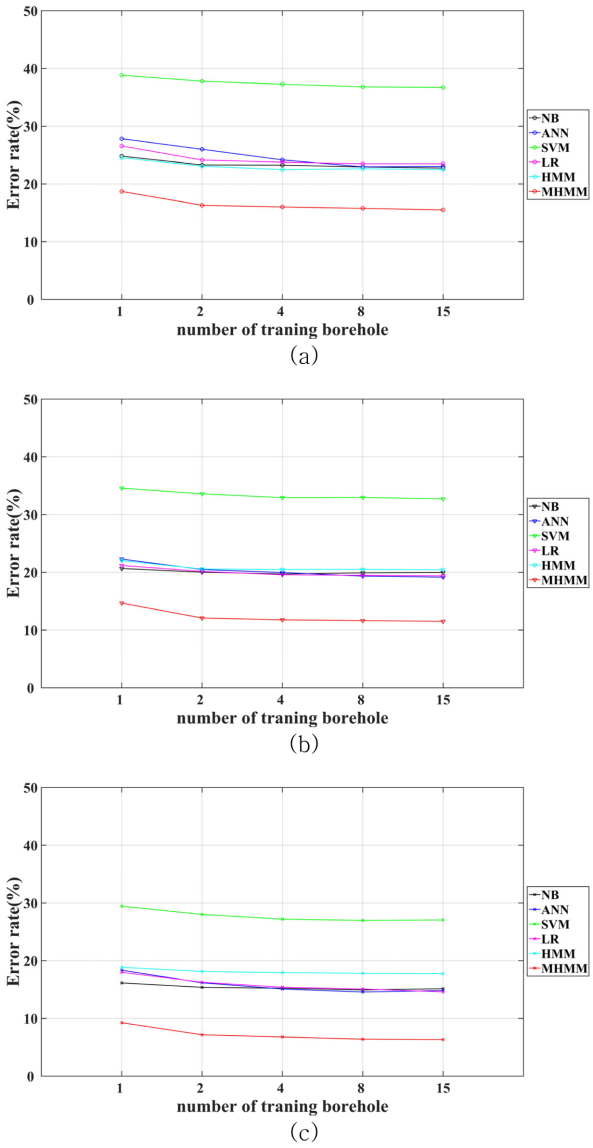

Fig. 6. The changes in the error rate with increment of the training boreholes corresponding to (a) the highest, (b) intermediate and (c) the lowest degree of errors.

Fig. 6(a)는 학습정보의 개수에 따른 각 학습기의 분류성능을 보여준다. X축은 분류기 학습 시 이용된 학습 시추공의 개수를 나타내며 Y축은 오분류율(%)을 나타낸다. 모든 학습기들이 학습정보의 개수가 증가 할수록 낮은 오

Table 3. The confusion matrices and the related rate for each machine learning technique in Case 2

			Predicted Class				
			C1	C2	C3	C4	True Rate
NB	True Class	C1	9534	265	45	1466	84.30%
		C2	115	14289	1695	761	84.75%
		C3	49	2390	12762	2229	73.22%
		C4	1584	552	2163	15861	78.68%
	False Rate		15.49%	18.33%	23.42%	21.93%	
	Accuracy				79.75%		
ANN	True Class	C1	9236	262	42	1770	81.66%
		C2	195	13637	2258	770	80.88%
		C3	63	2130	12819	2418	73.55%
		C4	1582	490	2202	15886	78.80%
	False Rate		16.61%	17.45%	25.99%	23.79%	
	Accuracy				78.43%		
SVM	True Class	C1	10427	31	7	845	92.19%
		C2	3099	12949	664	148	76.80%
		C3	5713	1841	8738	1138	50.13%
		C4	6840	180	1790	11350	56.30%
	False Rate		60.02%	13.68%	21.98%	15.81%	
	Accuracy				66.09%		
LR	True Class	C1	8902	235	60	2113	78.71%
		C2	100	13858	2088	814	82.19%
		C3	51	1901	12932	2546	74.19%
		C4	1043	483	2086	16548	82.08%
	False Rate		11.83%	15.89%	24.67%	24.85%	
	Accuracy				79.44%		
HMM	True Class	C1	11040	2	2	266	97.61%
		C2	1476	13399	1211	774	79.47%
		C3	4163	1235	10475	1557	60.10%
		C4	1296	569	1497	16798	83.32%
	False Rate		38.58%	11.88%	20.55%	13.39%	
	Accuracy				78.64%		
mHMM	True Class	C1	10664	31	137	478	94.29%
		C2	7	14658	1316	879	86.94%
		C3	38	1577	13763	2052	78.96%
		C4	36	577	1673	17874	88.66%
	False Rate		0.75%	12.97%	18.51%	16.02%	
	Accuracy				86.62%		

분류율을 보여주고 있다. 인공 신경망의 경우 학습정보의 양이 많을 경우에는 나이브 베이즈 분류기, 로지스틱 회귀 분류기 및 은닉 마르코프 모델과 유사하거나 낮은 오분류율을 보여주나 학습정보의 양이 적을 경우 이의 오분류율이 다른 학습기에 비해 급격히 증가하고 있음을 확인할 수 있다. 이는 학습 시추공의 개수가 적은 경우에 대해 인공 신경망 분류기 설계 시 설정된 은닉노두의 개수가 다소 많게 작용하여 학습정보에 비해 과도한 비선형 결정경계를 만듦으로써 과다적합(overfitting)의 문제를 발생시켰기 때문이다. 이와 같이 인공신경망 분류기는 학습정보의 양에 대해 다른 학습기 보다 더 큰 민감도를 보이므로 이를 분류를 위해 이용 할 때, 학습정보의 양에 대하여 은닉노두의 수를 비례적으로 조절해 주는 등 주의가 필요하다.

Table 3은 각 분류기에 케이스 2번의 간접정보를 적용하였을 때의 이들의 분류 성능을 나타내는 혼동행렬이다.

Table 4. The confusion matrices and the related rate for each machine learning technique in Case 3

			Predicted Class				
			C1	C2	C3	C4	True Rate
NB	True Class	C1	9557	46	121	1586	84.50%
		C2	42	14728	2085	5	87.35%
		C3	168	1856	14383	1023	82.52%
		C4	2534	2	829	16795	83.31%
	False Rate		22.31%	11.45%	17.42%	13.47%	
	Accuracy				84.34%		
ANN	True Class	C1	8396	103	163	2648	74.24%
		C2	45	14404	2404	7	85.43%
		C3	93	1895	14285	1157	81.96%
		C4	1737	13	828	17582	87.21%
	False Rate		18.26%	12.25%	19.20%	17.82%	
	Accuracy				83.13%		
SVM	True Class	C1	9881	63	19	1347	87.37%
		C2	2032	13640	1184	4	80.90%
		C3	5198	1743	9827	662	56.38%
		C4	5382	0	1127	13651	67.71%
	False Rate		56.07%	11.69%	19.17%	12.85%	
	Accuracy				71.47%		
LR	True Class	C1	8259	83	122	2846	73.02%
		C2	23	14242	2567	28	84.47%
		C3	59	1675	14225	1471	81.61%
		C4	1344	32	775	18009	89.33%
	False Rate		14.72%	11.17%	19.58%	19.44%	
	Accuracy				83.23%		
HMM	True Class	C1	11257	1	2	50	99.53%
		C2	1941	13522	1210	187	80.20%
		C3	4497	1166	11098	669	63.67%
		C4	1606	151	575	17828	88.43%
	False Rate		41.68%	8.88%	13.87%	4.84%	
	Accuracy				81.67%		
mHMM	True Class	C1	10819	44	108	339	95.66%
		C2	0	15281	1340	239	90.63%
		C3	16	1348	15088	978	86.56%
		C4	24	146	696	19294	95.70%
	False Rate		0.37%	9.14%	12.44%	7.46%	
	Accuracy				91.97%		

케이스 1번의 경우와 비교하였을 때 케이스 2번에서 모든 분류기들의 예측능이 전반적으로 향상된 것을 확인 할 수 있다. Fig. 7은 각 케이스에서 분류기들의 정확도를 막대그래프로 보여주며 케이스 2번에서의 전반적인 예측능 향상을 다시 한 번 확인 할 수 있다. Fig. 5(a) 및 Fig. 5(b)에서와 같이 케이스 1번과 케이스 2번에 해당하는 간접정보들을 데이터 공간 내 도시하여 비교하였을 때, 각 매질에 대한 분포 형태가 가시적으로 확연한 차이를 보이지 않음에도 불구하고 기계학습을 통한 분류결과, 케이스 2번에서의 예측능이 뚜렷하게 증가한 것을 확인 할 수 있다. 이를 통해 기계학습을 활용한 자동분류 기법이 보다 정밀한 분류작업이 가능함을 알 수 있다. 케이스 2번 내에서 분류기 간의 성능 패턴이 케이스 1번에서의 결과와 유사하게 나타나고 있으나 은닉 마르코프 모델의 예측능이 케이스 1번에서는 수정된 은닉 마르코프 모델 다음으로 높았던 반면 케이스 2번에서는 지지벡터기반 분류기를

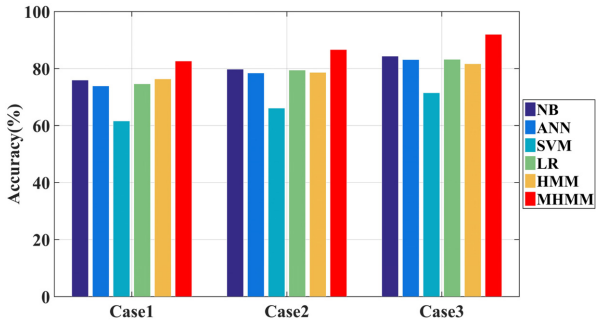

Fig. 7. Comparisons of the accuracy of each machine learning technique at each test cases.

제외한 다른 대부분의 분류기에 비하여 낮아진 경향을 보인다. 이는 간접정보만을 이용하는 분류기들이 케이스 1에 비하여 서로 간의 오버랩이 적은 간접정보를 이용함에 따라 이들의 성능이 향상된 반면, 은닉 마르코프 모델은 예측하고자하는 시추공 내에서 지질학적 비정규성이 뚜렷하게 나타남에도 불구하고 하향방향으로만 학습된 전이확률만을 이용함에 따라 잘못된 정보의 입력이 주어졌기 때문이다. 특히, 매질 1번에 대한 예측능이 오히려 케이스 1번에 비하여 증가하고 있다. 이를 통해, 본 연구에서 가정된 매질 1번과 같이 지질 내 지향적 비정규적 특성이 뚜렷한 지질이 존재할 경우 이에 대한 충분한 고려가 없는 기존 은닉 마르코프 모델의 적용은 해당 매질에 대한 상당한 예측능 저하를 가져옴을 알 수 있다.

케이스 3번의 간접정보들은 이들 간의 분포가 다른 케이스에 비하여 확연히 분류되고 있는 경향을 보여준다 (Fig. 5(c)). 따라서 예측능 또한 향상된 경향을 보여주며 (Fig. 7 및 Table 4) 분류기간의 성능 패턴 또한 케이스 2번과 유사하게 나타나고 있다. 이와 같은 결과를 통해 간접정보들 간의 교란이 심하고 측정의 오차가 감소할 경우 분류기들의 성능 또한 비례하여 증가하는 것을 알 수 있다.

5. 결 론

본 연구에서는 간접정보와 매질 간의 통계적 상관성을 이용하여 간접정보로부터 매질정보를 자동적 및 객관적으로 예측하는 다양한 기계학습기법에 대해 소개하고 이들 간의 성능을 비교 및 분석하였다. 이용된 기계학습 기법으로는 나이브 베이즈 분류기, 인공 신경망, 지지벡터기반 분류기, 로지스틱 회귀 분류기, 은닉 마르코프 모델 및 수정된 은닉 마르코프 모델이 있다. 각 분류기들 간의 비교 및 분석을 위해 4가지 매질로 구성된 가상의 16개 시추공과 각 시추공에 대해 2가지 종류의 간접정보를 가정하였으며 이 때, 간접정보의 경우 자료탐사 시 발생할 수 있는 인근 매질로부터의 교란 및 탐사측정 기기의 오차 정도를 고려하여 3가지 케이스의 간접정보를 설정하였다. 이를 통해 인근 매질에 따른 교란 및 오차 정도에 따른 분류기들의 성능을 비교 및 검증하였다. 또 다른 분석을 위해 분류기 학습 시 학습 자료의 개수를 달리하여 각 분류기들을 학습 시킨 후, 학습자료 개수에 대한 각 분류기의 성능 또한 검증하였다.

3가지 간접정보 분포에 따른 분류기 성능 분석 결과, 교란 및 오차의 정도가 감소할수록 분류기들의 성능이 향상되는 것을 확인 할 수 있었으며 이 때, 각 케이스에 따른 간접정보들을 데이터 공간 내 도시한 그래프에서 각 케이스 간의 분포 차이가 가시적으로 확인하지 못할 정도로 작았음에도 불구하고 예측결과에서의 각 케이스에 대한 확연한 차이를 확인함에 따라 기계학습기법을 통한 매질예측이 정밀하고 효과적임을 알 수 있다. 각 케이스에서 분류기들의 성능을 비교해 본 결과, 간접정보만을 활용하는 분류기들의 예측성능은 지지벡터기반 분류기를 제외하고 유사하게 나타났으며 지지벡터기반 분류기 성능의 저하는 해당 분류기가 이분류 문제에 적합하기 때문이다. 간접정보 외 전이정보를 추가적으로 사용하는 기존 은닉 마르코프 모델의 결과는 다른 간접정보만을 이용하는 학습기에 비하여 좋은 결과를 보였으며 이는 추가적 정보 활용 때문인 것으로 판단된다. 수정된 은닉 마르코프 모델의 경우 지향적 비정규적 특성을 고려하여 하향 및 상향으로의 두 가지 전이정보를 활용함으로써 예측능이 기존의 은닉마르코프 모델을 적용한 결과 보다 확연히 증가한 결과를 나타내었다. 또한 학습 자료의 개수가 증가할수록 분류기의 성능 또한 비례하여 증가하였으며 인공 신경망의 경우 학습 자료의 개수가 줄어들 경우 과다적합의 문제가 발생 할 수 있으므로 학습 자료 개수에 따른 정밀한 분류기 설계가 필요한 것으로 나타났다.

일반적으로 기계학습 기법 중 생성적 접근법을 이용하는 모델이 식별적 접근법의 모델 보다 낮은 성능을 보인다고 알려져 있으나 본 연구에서는 생성적 접근법인 나이브 베이즈 분류기가 모든 케이스 및 테스트에 대하여 전반적으로 우수한 예측능을 보여주고 있다. 이는 본 연구에서 이용한 가상의 간접정보자료를 가우시안 분포를 이용하여 생성하였고 본 연구에서 이용된 나이브 베이즈 분류기의 분포에 대한 가정 역시 가우시안 분포로 설정하였기 때문에 이로부터의 영향이 있을 것으로 판단된다. 따라서 보다 객관적인 검증 및 비교분석을 위해서는 간접정보들을 가우시안 분포가 아닌 다른 형태의 통계적 분포

모델을 적용하여 발생 시킨 후 각각의 분류기에 대해 성능을 검증하여야 할 것으로 판단된다. 또한 본 연구는 기반암이 존재하는 경우의 지질학적 지향적 비정규성만을 고려한 케이스를 활용하여 검증을 실시하였으나 보다 다양한 비정규적 특성의 지질분포를 고려한 검증이 필요할 것으로 판단된다.

사 사

본 연구는 환경부의 이산화탄소 저장 환경관리기술개발 사업에서 지원받았습니다.

References

An, L.H., 2000, Neutral network in lithology determination, *Tap Chi Tin Hoc Va Dieu Khien Hoc*, **16**(2), 59-62.

Bauer, K., Schulze, A., Ryberg, T., Sobolev, S.V., and Weber, M.H., 2003, Classification of lithology from seismic tomography: A case study from the Messum igneous complex, Namibia, *J. Geophys. Res.*, **108**(B3), 2152, doi:10.1029/2001JB001073.

Benaouda, D., Wadge, G., Whitmarsh, R.B., Rothwell, R.G., and MacLeod, C., 1999, Inferring the lithology of borehole rocks by applying neural network classifiers to downhole logs: an example from the Ocean Drilling Program, *Geohphys. J. Int.*, **136**, 477-491.

Birch, F., 1961, The velocity of compressional waves in rocks to 10 kilobars, part 2, *J. Geophys. Res.*, **66**, 2199-2224.

Bishop, C.M., 1995, Neural networks for pattern recognition, Oxford University Press.

Bowling, J.M., Rodriguez, A.B., Harry, D.L., and Zheng, C., 2005, Delineating alluvial aquifer heterogeneity using resistivity and GPR data, 2005, *Ground Water*, **43**(6), 890-903.

Bowling, J.M., Harry, D.L, Rodriguez, A.B., and Zheng, C., 2007, Integrated geophysical and geological investigation of heterogeneous fluvial aquifer in Columbus Mississippi, *Journal of Applied Geophysics*, **62**, 58-73.

Busch, J.M., Fortney, W.G., and Berry, L.M., 1987, Determination of lithology from well logs by statistical analysis, *SPE Formation Evaluation*, **2**, 412-418.

Chang, H.C., Chen, H.C., and Fang, J.H, 1997, Lithology Determination from Well Logs with Fuzzy Associative Memory Neural Network, *IEEE Transactions on Geoscience and Remote Sensing*, **35**(3), 773-780.

Christensen, N.I., 1996, Poisson's ratio in crustal seismology, *J. Geophys. Res.*, **101**, 3139-3156.

Cortes, C. and Vapnik, V., 1995, Support-vector networks, *Machine Learning*, **20**(3), 273, DOI:10.1007/BF00994018.

Delfiner, P., Peyret, O., and Serra, O., 1987, Automatic determination of lithology from well logs, *SPE Formation Evaluation*, **2**, 303-310.

Dunne, R.A. and Campbell, N.A., 1997, On the pairing of the softmax activation and cross-entropy penalty functions and the derivation of the softmax activation function, *Proc. 8th Aust. Conf. on Neural Networks, Melb.*, 181-185.

Eidsvik, J., Mukerji, T., and Switzer, P., 2004, Estimation of geological attributes from a well log: an application of hidden Markov chains, *Mathematical Geology*, **36**(3), 379-397.

Elliott, R.J., Aggoun, L., and Moore, J.B., 1995, Hidden Markov models: estimation and control, Springer, New York.

Fung, C.C., Wong, K.W., Eren, H., and Charlebois, R., 1995, Lithology Classification Using Self-Organising Map, *IEEE international Conference*, 1.

Gassaway, G.R., Miller, D.R., Bennett, L.E., Brown, R.A., Rapp, M., and Nelson, V., 1989, Amplitude variations with Offset: fundamentals and Case Histories, *SEG Continuing Education Course Notes*.

Guven, O., Molz, F.J., Melville, J.G., El Didy, S., and Boman, G.K., 1992, Three-dimensional modeling of a two-well tracer test, *Ground water*, **30**(6), 945-957.

Jeong, J., Park, E., Han, W.S., and Kim, K.Y., 2014, A novel data assimilation methodology for predicting lithology based on sequence labeling algorithms, *Journal of Geophysical Research: Solid Earth*, **119**(10), 7503-7520.

Kendall, M.G. and Stuart, A., 1977, The advanced theory of statistics, Vol. 4, 4th ed., MacMillan, New York.

Liu, G., Zheng, C., Gorelick, S.M., 2004, Limits of applicability of the advection-dispersion model in aquifers containing high-conductivity channels, *Water Resour. Res.*, **40**, W08308, DOI: 10.1029/2003WR002735.

Maiti, S., Tiwari, R.K., and Kumpel, H.-J., 2007, Neural network modelling and classification of lithofacies using well log data: a case study from KTB borehole site, *Geophys. J. Int.*, **169**, 733-746.

Murphy, K.P., 2012, Machine learning: a probabilistic perspective, MIT press, 1067p.

Park, E., 2010, A multidimensional generalized coupled Markov chain model for surface and subsurface characterization, *Water Resour. Res.*, **46**(11), W11509, doi:10.1029/2009WR008355.

Pickett, G.R., 1963, Acoustic character logs and their application in formation evaluation, *J. Petr. Tech.*, **15**, 659-667.

Rimstad, K. and Omre, H., 2013, Approximate posterior distributions for convolutional two-level hidden Markov models, *Computational Statistics and Data Analysis*, **58**, 187-200,

doi:10.1016/j.csda.2012.09.001.

Rogers, S.J., Fang, J.H., Karr, C.L., and Stanley, D.A., 1992, Determination of lithology from well logs using a neural network, *The American Association of Petroleum Geolgists Bulletin*, **76**(5), 731-739.

Sandham, W. and Leggett, M., 2003, Geophysical applications of artificial neural networks and fuzzy logic, Springer-science+business Media, Kluwer Acad., Boston, 348 p.

Schon, J.H., 1996, Physical Properties of Rocks: Fundamentals and Principles of Petrophysics, *Handbook of Geophys. Explor., Seismic Explor.*, 18, Pergamon, New York.

Schumann, A., 1997, Neural networks versus statistics: a comparing study of their classification performance on well log data, Proc. Annu. Conf. *International Association for Mathematical Geology-IAMG'97, 3rd*, Barcelona, Spain, 237-241.

Schumann, A., 2002, Hidden Markov models for lithological well log classification, *In Proc. Annu. Conf. International Association for Mathematical Geology-IAMG'2002, 8th*, Berlin, Germany, 373-378.

Smirnoff, A., Boisvert, E., and Paradis, S.J., 2008, Support vector machine for 3d modelling from sparse geological information of various origins, *Comput. Geosci.*, **34**(2), 127-143.

Sobolev, S.V. and Babeyko, A.Y., 1994, Modeling of mineral composition, density, and elastic wave velocities in anhydrous magmatic rocks, *Surv. Geophys.*, **15**, 515-544.

Thompson, A.F.B. and Gelhar, L.W., 1990, Numerical siulation of solute transport in three-dimensional randomly heterogeneous porous media, *Water Resour. Res.*, **26**, 2541-2562.

Wollff, M. and Pelissier-Combescure, J., 1982, Faciolog: automatic electrofacies determination, *SPWLA Annual Logging Symposium*, 6-9.

Zheng, C. and Gorelick, S.M., 2003, Analysis of the effect of decimeter-scale preferential flow paths on solute transport, *Ground Water*, **41**(2), 142-155.