http://dx.doi.org/10.14400/JDC.2016.14.2.245

기계학습 응용 및 학습 알고리즘 성능 개선방안 사례연구

이호현*, 정승현*, 최은정**
폴수학학교*, 서울여자대학교 정보보호학과**

A Case Study on Machine Learning Applications and Performance Improvement in Learning Algorithm*

Hohyun Lee*, Seung-Hyun Chung*, Eun-Jung Choi**
Data Science Lab, Paul Math School*
Dept. of Information Security, Seoul Women's University**

요 약 본 논문에서는 기계학습과 관련된 다양한 사례들에 대한 연구를 바탕으로 기계학습 응용 및 학습 알고리즘의 성능 개선 방안을 제시한다. 이를 위해 기계학습 기법을 적용하여 결과를 얻어낸 문헌을 자료로 수집하고 학문 분야로 나누어 각 분야에서 적합한 기계학습 기법을 선택 및 추천하였다. 공학에서는 SVM, 의학에서는 의사결정나무, 그 외 분야에서는 SVM이 빈번한 이용 사례와 분류/예측의 측면에서 그 효용성을 보였다. 기계학습의 적용 사례 분석을 통해 응용 방안의 일반적 특성화를 꾀할 수 있었다. 적용 단계는 크게 3단계로 이루어진다. 첫째, 데이터 수집, 둘째, 알고리즘을 통한 데이터 학습, 셋째, 알고리즘에 대한 유의미성 테스트 이며, 각 단계에서의 알고리즘의 결합을 통해 성능을 향상시킨다. 성능 개선 및 향상의 방법은 다중 기계학습 구조 모델링과 $+\alpha$ 기계학습 구조 모델링 등으로 분류한다.

주제어 : 기계학습 분류, 기계학습 구조, 기계학습 모델링, 성능 개선, 최적화

Abstract This paper aims to present the way to bring about significant results through performance improvement of learning algorithm in the research applying to machine learning. Research papers showing the results from machine learning methods were collected as data for this case study. In addition, suitable machine learning methods for each field were selected and suggested in this paper. As a result, SVM for engineering, decision-making tree algorithm for medical science, and SVM for other fields showed their efficiency in terms of their frequent use cases and classification/prediction. By analyzing cases of machine learning application, general characterization of application plans is drawn. Machine learning application has three steps: (1) data collection; (2) data learning through algorithm; and (3) significance test on algorithm. Performance is improved in each step by combining algorithm. Ways of performance improvement are classified as multiple machine learning structure modeling, $+\alpha$ machine learning structure modeling, and so forth.

Key Words : machine learning classification, machine learning modeling, performance improvement, optimization

* 본 논문은 2015 학년도 서울여자대학교 컴퓨터과학연구소 교내학술연구비에 의하여 지원되었음
Received 24 December 2015, Revised 28 January 2016
Accepted 20 February 2016, Published 28 February 2016
Corresponding Author: Eun-Jung Choi
(Seoul Women's University)
Email: chej@swu.ac.kr

ISSN: 1738-1916

1. 서론

기계학습(Machine learning)은 환경에 따른 데이터의 특징을 추출하고 테스트하여 최적화 또는 자가 발전을 하는 일련의 과정이다. 인간의 학습은 경험에서 비롯된다는 관점에서 기계학습에서는 이러한 경험의 요소로 환경과 데이터를 꼽을 수 있다. 경험의 크기나 질에 따라 인간이 학습하는 양과 질이 달라지는 것과 같이 기계학습에서도 데이터의 크기와 질이 기계학습 결과의 질과 유용성을 결정한다. 경험을 바탕으로 인간에게서 학습이 이루어지면 다음으로 이에 대한 테스트로 피드백이 이어져 인간이 사고과정을 통한 지능을 갖게 되듯이 기계학습에서는 인지사고 알고리즘을 통해 수집한 데이터에 대한 학습이 이루어진 뒤에 특징이 추출되면 그 특징의 정확성을 판단하는 과정이 이루어진다. 그리고 그 알고리즘의 오류와 오차에 따라서 또는 보상(reward)에 따라 이를 피드백 하여 스스로 성능을 개선시켜 문제에 대한 해답에 접근한다. 이러한 일련의 과정 반복은 환경에 따라 변하는 데이터를 주어진 환경과 결부 시켜 주어진 환경 하에서 보다 정확한 분류를 가능케 한다.

기계학습이 각광을 받는 이유는 정보의 양이 방대해짐에 따라 고려해야 할 데이터의 범위가 광범위해지고 정확하면서 저렴하고 빠른 정보처리와 계산을 필요로 하게 되었기 때문이다[1]. 박근혜 대통령은 빅데이터(big data)를 '21세기의 원유'라고 칭하며 현재와 미래에서의 빅데이터의 중요성을 부각시켰다. 기계학습은 이러한 빅데이터에서 발견치 못한 유의미한 정보를 추출하는 기능을 탑재하고 있다. 더 나아가 실시간 분석을 통한 예측 등과 같은 유용한 작업의 수행이 가능하기 때문에 각광을 받아왔다. 아울러 공학, 의학, 사회, 자연, 예술, 체육, 인문과학 등에 활용 및 적용되어 성과를 거두는 등 응용 가능한 범위가 방대해지면서 그 효용성을 인정받게 되었다.

보다 원활한 적용력을 탑재한 기계학습 시스템을 위해서는 데이터 확보 능력, 적합한 알고리즘 선택, 자동화 및 반복 프로세스, 확장성, 앙상블 모델링에서 높은 수준의 기술이 요구된다[1]. 때문에 기계학습을 적용할 때 데이터 준비와 알고리즘, 자동화 및 반복 프로세스에서 실험 환경에 맞는 최적의 방법을 찾는 것을 우선으로 한다. 본 논문에서는 기계학습을 적용해야 하는 연구에서 기계학습 알고리즘의 기능 향상을 통해 유의미한 결과를 도출할 수 있는 방안을 제시하는 것을 목표로 한다.

2. 역사 및 응용 분야

기계학습이라는 용어는 1959년 A. L. Samuel의 논문 "Some Studies in Machine Learning Using the Game of Checkers"[2]에서 처음 모습을 드러내었다. 또한 같은 1950년대에 Rosenblatt은 초기 인공신경망 연구에 큰 영향을 미친 퍼셉트론(Perceptron)을 제안하였다. 이러한 기계학습 태동기의 연구 사례를 바탕으로 발전하여 1980년대 중반에 이르러 하나의 독립적인 연구 분야로 자리매김하기 시작하였다. 그 후로 수많은 연구들이 가지 치듯 진행되었는데 통계학과 인공지능의 결합이 이루어지면서 비약적인 성과를 거두게 되었다. 수학적, 확률통계학적인 내용이 기계학습의 기반으로 작용함으로써 기계학습은 새로운 기술, 학문 분야로 인정받고 정립되기 시작하였다. 1990년대 중반 이후에는 인터넷 상용 서비스가 개시되는 등, 인터넷 통신망이 본격적으로 구축되면서 산업계에서는 데이터마이닝 개념이 태동 되었고 핵심 기술로 기계학습이 사용되었다. 현재 기계학습은 데이터를 분석하고 학습하는 기술로서 인터넷 정보검색, 텍스트 마이닝, 생물정보학, 바이오메트릭스, 자연언어처리, 음성인식, 컴퓨터비젼, 컴퓨터그래픽, 로보틱스, HCI, 통신사업, 서비스업, 제조업 등 거의 모든 융합연구 분야에서 활용되는 핵심 기반 기술이다[3].

3. 기계학습의 분류

기계학습은 학습 데이터를 획득하는 방법에 따라 <Table 1>과 같이 지도학습(Supervised Learning, SL), 자율학습(Unsupervised Learning, UL), 반지도학습(Semi-Supervised Learning, SSL), 강화학습(Reinforcement Learning, RL)으로 구분될 수 있다. 지도학습은 입력값과 그에 상응되는 기대 출력값을 데이터로 갖는다. 입력 데이터를 기대 출력에 최대한 가깝게 하도록 하는 알고리즘이 지도학습 알고리즘이다. 이 알고리즘은 학습을 실행하였을 때 기대 출력값과 학습된 입력

값의 차를 비교하여 오류를 찾아낸다. 이 오류는 모델을 수정하는 데 있어서 근거가 된다. 반면, 자율학습은 입력값만을 데이터로 갖는다. 따라서 입력 데이터에 내재되어 있는 의미를 분석하는 것을 목적으로 한다. 반지도 학습은 지도학습에서 출력값(labeled)이 부족할 경우 입력값과 상응되도록 가공되지 않은 출력값(unlabeled)으로 보충하여 학습 데이터를 구성하고 이를 학습하는 알고리즘이다. 출력값은 높은 비용과 노력이 소요되므로 낮은 비율로 사용할 수밖에 없다는 점으로부터 제안된 학습 알고리즘이다. 강화학습은 입력 데이터가 어떤 특징을 갖고 있느냐에 따라서 행동을 결정하여 실행한다. 각 실행에 따라 이를 평가하고 평가에 따라 보상이 주어진다. 행동은 보상이 극대화되는 방향으로 결정되어 실행된다. 기계학습 방법에서 지도학습이 70%라는 가장 높은 비율로 사용되고, 자율학습이 약 10~20%의 비율로 사용되며, 반지도 및 강화학습이 나머지를 차지하고 있다[1].

〈Table 1〉 Classification based on types of learning

Type	Technique	Learning application
SL	• perception • classification • diagnosis • regression • forecasting	• fraud credit card transaction • forecasting of insurance claim possibility • voice recognition • robot control
UL	• Self Organizing Map (SOM) • grouping • density estimation • dimension reduction • feature extraction • K-means	• segment management • grouping text topics • item recommendation • outlier detection
SSL	• classification • regression • forecasting	• individual facial recognition
RL	• trial & error • reward function • dynamic programming	• robotics • games • navigation

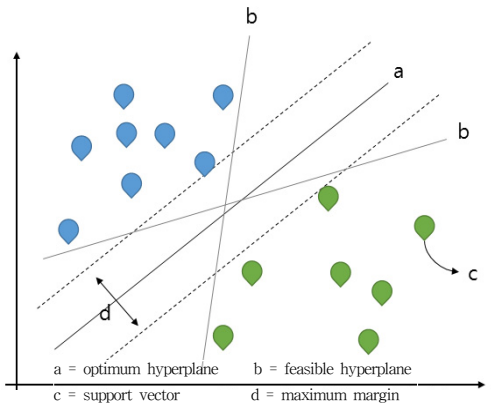
학습법들을 수학적으로 보면, 학습 시스템은 함수 또는 사상 $f(x;M)$을 관측 데이터로부터 구성한다. 여기서 학습 모델 M은 구조 S와 파라미터 벡터 W로 구성된다. 지도학습은 입력벡터 x와 이에 대한 출력벡터 d가 주어져 $D=\{(x,d)\}$에 기반하여 x가 d에 수렴하여 최적화되는 방법을 추론하고 학습한다. 자율학습은 출력벡터 d가 주어지지 않고 오직 입력벡터 x만이 주어져 $D=\{(x)\}$에 기반하여 x들 간의 관계성이나 밀도를 추론하고 학습한다. 강화학습은 지도학습에서의 출력벡터 d 대신에 예측치에 대한 평가치 e를 갖는다. 입력벡터 x는 $D=\{(x,e)\}$를 기반으로 e를 극대화시킨다. 강화학습은 평가치 e로 인하여 지도학습 및 자율학습과는 달리 순차적인 의사결정 과정에 관여한다[3].

특정한 문제에 대한 기계학습의 적용이 이루어질 때 각 학습별 알고리즘이 적용된다. 지도학습 알고리즘에는 인공신경망(Neural Network), Support Vector Machine(SVM), 커널 머신, 의사결정 나무, Naive Bayes, k-최근린(k-NN), 베이지안 망, 은닉마코프 모형(HMM), 잠재변수모델 등의 확률그래프 모델이 있다. 또한, 자율학습 알고리즘에는 k-means, 계층적 군집화, SOM 등이 있고, 강화학습에는 Q-Learning 등이 있다. <Table 2>와 같이 이러한 기계학습 알고리즘의 분류는 모델의 구조 즉, 학습이 이루어지는 방식의 구분에 따른다. 기계학습 성능 분석 프로그램인 WEKA가 이 분류법으로 기계학습 기법을 지원한다[4,16].

〈Table 2〉 Classification of algorithm based on structure

Structure	Algorithm
Bayes	NaiveBayes, BayesNet, etc.
Trees	ID3, J48, RandomForest, LMT, etc.
Rules	Prism, Nnge, etc.
Functions	SVM, SVO, etc.
Lazy	IB1, etc.

본 논문에서는 지도학습 알고리즘에 속한 SVM, 인공신경망, 결정 나무, 베이지안, k-NN과 강화학습 알고리즘을 주로 다룬다. 특히 SVM과 인공신경망은 최근 가장 많이 사용되고 있는 알고리즘들이며 Q-Learning은 강화학습의 대표적인 알고리즘이다.



[Fig. 1] Binary classification of SVM

SVM은 서로 다른 클래스의 벡터들을 그 사이의 거리에 대해 최대의 마진(margin)으로 분류할 수 있는 서포트 벡터들로 이루어진 초평면을 찾는 방법으로 학습하는 이진 분류기이다. SVM은 선형으로 분류 가능한 선형 데이터에 대해서는 2차원 평면상에서 분류를 하지만 선형으로 분류 불가능한 비선형 데이터에 대해서는 커널함수를 이용하여 평면에서 고차원 공간에 사상시킨 뒤 분류한다. 여기에서 커널함수에는 Polynomial, Sigmoid, Radial Basis Function(RBF)이 있는데 적용하는 문제에 적합한 커널을 선택하여 사용할 수 있다. 각각의 커널함수의 식은 다음 <Table 3>과 같다[6].

〈Table 3〉 Types of kernel function

Kernel function	Mathematical expression
Polynomial	$K(x,y)=(x \bullet y)^p$
Sigmoid	$K(x,y)=\tanh(\beta_0 x \bullet y+\beta_1)$
RBF	$\mathrm{P}K(x,y)=e^{(-\frac{\|x-y\|^2}{2\sigma^2})}$

또한, 초평면은 공간상에서 위치설정에 무한한 가능성을 갖는다. 초평면의 목적은 공간상의 각 클래스 데이터 벡터들 간의 사이의 거리를 가장 길게 하여 마진을 최대화하는 것이다. 마진 d는 다음과 같이 나타낼 수 있다.

$$d=2/\|w\|$$

또한, 초평면을 다음과 같이 나타낼 수가 있고

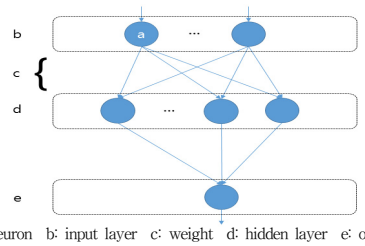
$$w \bullet x+b=0$$

모든 데이터가 초평면에서의 거리가 최소한 1이므로

$$w \bullet x+b \geq 1$$

$$w \bullet x+b \leq 1$$

의 두 가지 클래스로 나눌 수 있다. 목적은 마진을 최대화하는 것이므로 이는 곧, $\|w\|$ 를 최소로 하는 b와 w를 찾는 것을 목표로 한다. 이러한 SVM의 장점은 명확하여 일반적으로 정확한 분류를 보여준다는 것이다.



a: neuron b: input layer c: weight d: hidden layer e: output layer

[Fig. 2] Structure of artificial neural network

인공신경망은 기계에게 인간이나 동물과 같이 학습이 가능케 하고자 뇌의 구조를 수학적으로 모델화한 정보처리 시스템으로 다층의 뉴런과 각 뉴런을 연결하는 가중치로 구성되어 있다[7,9]. 인공신경망은 데이터의 흐름 방향, 뉴런의 계층 구조, 뉴런의 배치 및 연결 형태 세 가지 기준으로 나뉠 수 있다. 데이터의 흐름이 한 쪽으로만 흐르는 순방향 인공신경망과 출력 데이터가 다시 귀환되어 양방향으로 흐르는 순환 인공신경망으로 다시 나뉜다. 뉴런의 계층 구조는 입력층과 출력층 사이에 하나의 뉴런만을 갖는 단층 퍼셉트론과 하나 이상의 은닉층을 갖는 다층 퍼셉트론으로 다시 나뉜다. 단층 퍼셉트론보다는 다층 퍼셉트론을 이용하는 것이 일반적이다[10]. 뉴런의 배치 및 연결 형태에 따른 신경망의 구조에 따른 분류에 의하면 다층 퍼셉트론, 홉필드 넷, 해밍 넷, ART 등으로 분류될 수 있다. 인공신경망의 학습 과정은 오류 역전파 (Back Propagation) 방식에 기반을 둔다. 입력층의 뉴런에서 은닉층의 뉴런, 출력층의 뉴런으로의 이동은 모두 입력 데이터값에 가중치를 곱함으로써 이루어진다. 은닉층에서는 입력층으로부터 받은 데이터들을 모두 더하여 비선형 활성 함수 처리를 하고 이후 출력층에서는 은닉층으로부터 받은 데이터들을 모두 더하여 활성함수 처리를 하여 결과를 내보내게 된다[10]. 이렇게 출력된 결과값과 목표로 하는 출력값의 오차가 최소가 되도록 출력층의 에러를 다시 출력층에서 입력층으로 돌려보내어 가중치를 갱신한다. 여기에서 가중치 갱신은 다음의 식을 통해 이루어진다[10].

$$w_{ji}^{(l)}(n+1)=w_{ji}^{(l)}+\alpha[w_{ji}^{(l)}(n-1)]+\mu\delta_{ji}^{(l)}(n)y_{ji}^{(l)}(n) \quad (1)$$

$$\delta_{ji}^{(l)}(n)=\left[\begin{matrix} e_{ji}^{(L)}(n)\Phi'_j(v_j^{(L)}(n)) \\ \Phi'_j(v_j^{(L)}(n)\sum_k \delta_k^{(l+1)}(n)w_{kj}^{(l+1)}(n) \end{matrix}\right] \quad (2)$$

$$\Phi_j(v_j(n))=\frac{1}{1+\exp(-av_j(n))}(a>0,-\infty<v_j(n)<\infty) \quad (3)$$

$$\Phi'_j(v_j(n))=ay_j(n)\lfloor 1-y_j(n)\rfloor \quad (4)$$

각 층의 오류 역전파를 통한 가중치 갱신은 수식(1)과 같이 이전 층의 i번째 뉴런과 현재 층 l(중간층) 또는 L(출력층)의 j번째 뉴런에 대한 계산으로 이루어지며, 이렇게 나온 값으로 수식 (2)를 이용하여 에러값 δ를 계산하게 된다. δ값을 구하기 위해서는 활성 함수가 미분 가능해야 하는데, Logistic function 또는 Hyperbolic

tangent function 등이 사용된다. 수식 (3)과 (4)는 전자에 해당하는 sigmoid function의 원형식과 미분식을 보여주고 있다[10]. 역전파 알고리즘에는 Gradient descent method, Scaled conjugate gradient method 등이 있다.

실제 학습에 있어서는 크게 순차적인 방법과 일괄적인 방법이 있다. 순차적인 방법은 개개의 입력값들에 대한 에러값을 계산하여 바로 가중치를 갱신하는 방식으로 구현이 간편하다는 점과 방대하고 복잡한 패턴을 학습하는데 상대적으로 신속한 장점이 있다. 반면 일괄적인 방법은 모든 입력 패턴의 값을 처리한 후 마지막에 한 번 에러값을 통한 갱신을 해주는 방식으로 비교적 적은 양의 데이터를 빠르게 처리할 수 있는 장점이 있다.

Q-Learning은 아래의 식과 같은 업데이트 툴에 의해 가치 함수의 학습이 이루어진다[7,9].

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha[r_{t+1} + \gamma \max_a Q(s_{t+1},a_t) - Q(s_t,a_t)]$$

위의 식에서 $Q(s_t,a_t)$는 Q-value를 의미하고 s는 상태, a는 행동, r은 보상, γ는 할인 계수, α는 학습률을 나타낸다. 위의 식을 통해 학습되는 $Q(s_t,a_t)$는 반복 학습을 통해 최적의 행동값 함수 Q^*로 수렴함이 증명되었다. 이러한 Q-Learning의 분석 및 구현은 매우 간단한 장점이 있으며, 다음과 같은 코드로 실행시킬 수 있다.

```
Initialize Q(s,a) arbitrarily
Repeat(for each episode):
  Initialize s
  Repeat(for each step of episode):
    Choose a' from s' using policy derived from Q
    Take action a, observe r, s'
    Q(s,a) ← Q(s,a) + α[r + γ max_a' Q(s',a') − Q(s,a)]
    s ← s'
  until s is terminal
```

[Fig. 3] Pseudo code for Q-learning algorithm [7]

4. 기계학습의 적용사례

기계학습의 적용 사례별 특징 분석 이전에 기계학습이 활용되고 있는 분야를 학문별로 분류하였다. 또한, 학제 영역 내에서 세부적으로 어떠한 분야에 쓰이는지 살펴보았다. 분야 구분은 연구 목적에 따라 <Table 4>와 같이 분류하였다.

〈Table 4〉 Literature by interdisciplinary

Interdisciplinary	Field of study	Literature
Engineering	natural phenomenon forecasting, mobile robot, bio-signal detection, voice recognition, detection of garbage collection on SSD	[4,5], [7,8,9], [11], [12,13,14], [15], [16], [17], [18]
Medical science	disease prediction, medical image analysis, clinical causal relationship extraction	[6], [19], [20], [22], [35]
Social&Natural science	sentiment analysis, stock price prediction	[23], [24,25]
Agriculture	precipitation & flood prediction	[26,27,28], [29,30]
Art	musical genre grouping, box-office prediction	[10], [31]
Humanities	parsing system, competitive relation recognition	[32,33], [34]

<Table 4>는 기계학습에 대한 연구의 빈도가 점차로 증가하였고 공학 분야에서 오랜 기간 동안 가장 많은 연구가 진행된 것을 말해준다. 특히, 최근들어 공학과 의학 분야에서 적용 연구가 가장 활발히 진행되고 있으며, 그 외에도 다양한 분야에서 기계학습이 적용되고 있음을 알 수 있다. 본 연구에서는 공학과 의학, 기타 분야로 나누어 각각의 분야에서 구체적으로 기계학습이 어떻게 적용되고 있는가를 분석하였다.

4.1 공학

공학에 대한 기계학습의 적용사례에서는 로봇 기술의 관점에서 시각과 청각에 활용될 수 있는 영상, 음향, 제어 등에서의 기계학습을 이용한 구분, 분류, 판단 등의 연구가 활발히 진행되고 있다. 영상에서는 장애물이나 사람을 인식하는 연구가 진행되고 있으며 청각에서는 사람의 소리 등을 정확히 듣기 위해 소음을 제거하고 구분하는 등의 연구가 진행되고 있다. 또한, 제어에서는 주변 환경에 따라 행동을 결정할 수 있도록 하는 연구가 진행됨을 <Table 5>의 사례를 통해 알 수 있다. 반면, 후각과 촉각에 대한 센서와 관련된 연구가 시작되었으나 적용사례는 아직 미미하다. 기계학습 연구가 발전함에 따라 향후 로봇 개발에 적용될 수 있다는 공통적인 연구사례들로 인하여 로봇 기술의 관점에서 연구 동향을 조사하였다. 연구 문헌에서 적용된 기계학습 기법에는 SVM과 의사결정나무가 가장 많은 것을 알 수 있다. 그 뒤로 신경망 순이다. 사용 빈도수와 같이 정확성 또한 SVM이 매우 높은 정확도를 보여주었다.

〈Table 5〉 Engineering applications in literature

Technique	Application & feature	Case
Bayesian	■ error rate measure: SMO(21.79%), LMT(21.84%), BayesNet(23.01%), RandomForest(25.52%), J48(25.88%) ■ feature extraction method: cepstrum ■ performance analysis: WEKA, 10-fold cross validation	[4]
Regression	■ MLM(LR+Reinforcement+Merging) - time & cost efficient ■ goodness-of-fit test: EvM	[12]
Neural network	■ Artificial neural network+Q-learning ■ Feature selection algorithm - improvement in prediction/classification - reduction in learning time - relationship trace btw input & output	[7]
	■ same as Bayesian	[4]
SVM	■ Randomforest+SVR applied - robust in noise signal - precision improvement via complexity dispersion	[15]
	■ same as Bayesian	[4]
	■ proposed two-phase SVM algorithm - elaborated detection rate • accuracy: 99.32% • sensitivity: 100% • specificity: 99.21% - hard realization - slow computation time	[18]
RL	■ same as neural network	[7]
Decision Tree	■ C4.5 algorithm applied - quick classification - easy to understand learning results	[17]
	■ Multiple Linear Regression(MLR), CHAID, CART algorithms applied - similar performance btw CHAID&CART , prediction error 33% higher in MLR	[11]
	■ same as SVM	[15]

회귀분석 적용 기법을 사용한 사례는 선형회귀분석(Linear Regression, LR)을 적용하여 하나의 기체(A)에 최적화 되어 있는 제어계수를 특성이 다른 기체(B)에 적용한 후 기체 B의 최적 제어 계수를 자동으로 조정할 수 있는 제어 계수 기계 학습 모듈(MLM, Machine Learning Module)을 설계하고 실험을 통해 시스템의 가능성을 검증하였다[12]. 이 논문에서 제안한 MLM은 LR과 보정학습(Reinforcement), Merging으로 이루어져 있다. LR에서 발생한 잡음을 보정학습으로 보정하고 Merging으로 전송하여 이후에 작동될 시스템 제어에 적용될 수 있도록 포맷에 맞추는 작업을 한 후 내보낸다. 또한, EvM(Evaluation Module)이라는 제어 계수의 적합성을 평가하여 이를 제어 계수로서 받아들일 것인지를 판단하는 시스템을 추가하였다. 이러한 MLM은 시간이나 비용적 측면에서 효율적인 결과를 낼 수 있도록 하였다.

신경망 적용 기법을 이용한 사례는 모바일 로봇의 자율 주행 알고리즘을 사람의 장애물 인지과정과 주행 방법을 학습하는 과정을 모방하여 오직 카메라의 영상정보, 인공신경망, Q-Learning을 이용해 구현하고자 하는 시도가 있었다[7]. 이 논문에서는 인공신경망의 인식 성능을 향상시키기 위해 영상을 여러 방법으로 가공하여 입력 후보 feature(변수 또는 자질)를 생성하고 이를 feature 선택 알고리즘을 이용해 가공된 데이터 중에서 최적의 데이터를 찾아 인공신경망의 입력으로 사용하였다. Feature 선택 알고리즘의 구체적인 효과는 예측 성능 또는 분류 성능을 높일 수 있다는 것과, 학습 시간을 단축할 수 있다는 것, 입력과 출력 사이의 관계를 이해하는데 도움을 수 있다는 것 세 가지로 볼 수 있다.

SVM 적용 기법을 사용한 사례로 가속도 데이터를 바탕으로 낙상과 일상생활의 동작을 구분하는 이중 SVM 알고리즘을 개발하여 유효성을 평가하였다[19]. 개발한 이중 SVM 알고리즘은 정교한 감지율을 갖지만 높은 수학적 기술과 대량의 계산으로 인하여 기술을 구현하기 어렵고 메모리와 계산 시간의 소모가 많다는 단점이 있다. 이중 SVM 알고리즘을 이용하여 낙상을 검출한 결과 1차 SVM에서는 98.64%의 정확도, 90.48%의 민감도, 100%의 특이도가 나타났고 2차 SVM에서는 99.32%의 정확도, 100%의 민감도, 99.21%의 특이도가 나타나 매우 정확한 분류가 가능함을 보여주었다.

SVM과 의사결정나무 적용 기법을 사용한 사례는 분류기를 거친 다음 Support Vector Regression(SVR)을 사용하는 이중 기계학습 구조를 제안하여 단채널 생체신호를 이용하여 사용자의 안구이동을 추적하는 시스템과 여기에 추가적인 전극부착 없이 이를 깨무는 동작을 통해 추가적인 조작도 가능한 시스템을 제안하였다[15]. SVR은 수많은 종류의 곡선 맞춤 (nonlinear fitting) 모델 중 하나로 대만국립대학 소속의 CSIE(Computer Science & Information Engineering)에서 개발된 'LibSVM'에 수록되어 있고 SVM(Support Vector Machine)의 이론과 알고리즘을 기반으로 하며 수많은 곡선 맞춤 모델 중에서도 비교적 뛰어난 성능을 보이는 것으로 알려져 있다. 이의 성능에 매개변수의 영향이 많이 미치기 때문에 커

널 함수를 이용하여 최적화 과정이 필요하다. 커널 함수는 1차원 입력 처리를 할 수 있는 SVR을 다차원 입력 처리가 가능한 함수로 변형해주며 Gaussian 커널 함수를 적용하여 성능을 증가시켰다.

의사결정나무 적용 기법을 사용한 사례로는 SSD(Solid State Drive)의 가비지(garbage) 컬렉션 발생을 운영체제 레벨에서 관찰할 수 있는 SSD 상태정보와 C4.5 알고리즘을 통하여 예측하고 그에 따른 대처를 하여 가비지 컬렉션으로 인하여 발생하는 갑작스러운 대역폭 감소 문제를 해결하는 방법에 대하여 서술하였다[18]. 가비지 컬렉션이 발생하는 상황과 그렇지 않은 상황을 분석하기 위하여 C4.5 알고리즘을 사용하였는데 이 알고리즘의 장점은 처리 속도가 빠르고 학습된 결과를 이해하기 쉽다는 것이다. 결과에서는 가비지 컬렉션의 영향으로 생기는 대역폭 감소 현상이 완화되었으며 SSD의 평균 대역폭이 상승했다는 긍정적인 결과를 보였다.

종합적으로 각 기계학습별 성능을 특수 목적(화산 분출시 발생되는 초저음파 분류) 하에서 분석한 사례가 있다[4]. 다양한 기계학습기법을 적용하여 성능을 분석하고 초저음파의 도메인에 대하여 분류성능을 향상시킬 수 있는 기계학습기법을 제안하였다. 선택한 기계학습 기법은 Bayes그룹에서는 베이지안 망, NativeBayes, Functions 그룹에서는 RBF-Network, SMO, Lazy그룹에서는 IB1, Rules그룹에서는 Part, Nnge, Trees그룹에서는 RandomForest, LMT, J48이다. 이 논문에서는 자질 추출 방법으로 cepstrum 방식을 사용하였고 WEKA에서 10-fold cross validation을 사용하여 기법 성능 분석 및 비교를 하였다. 그 결과, 오류율은 SMO(21.79%), LMT(21.84%), 베이즈 망(23.01%), RandomForest (25.52%), J48(25.88%)의 순으로 나타났다.

4.2 의학

의학 분야에서는 기계학습이 환자나 의료기관의 편의를 증진시키고자 하는 연구가 진행됨을 볼 수 있다. 연구자들은 각자 상이한 연구주제를 가지고 있으며 문제해결을 위하여 광범위한 방면에서 적합한 해결책 등을 찾으려 한다. 이것이 최근 연구 동향이자 앞으로 나아가려는 방향이다. 의학 분야에서도 마찬가지로 각 환자에게 맞는 치료법 등이 세세하게 다르기 때문에 체질을 알아보고 그에 맞는 치료를 받아 건강을 유지 또는 되찾기 위해 노력하고 그렇게 해야 한다. 이러한 흐름에 기계학습이 반영 및 적용되어 체질을 알아보고 진단을 받아 진단결과를 받고 치료받기까지 기계학습이 전문가 시스템(expert system)의 특성을 이어받아 그 분석을 수행하고, 의사와 환자 모두에게 정확한 진단과 치료를 가능케 하는 연구가 진행되고 있다. 의학 분야에서는 신경망, SVM, 의사결정나무가 기계학습 기법 등이 가장 빈번하게 적용되고 있다. 성능은 의사결정나무에 기초한 C5.0 알고리즘과 인공신경망이 가장 좋은 것으로 나타났고 반면, 공학에서 높은 성능을 보인 SVM은 비교적 저조한 성능을 보여주었다.

〈Table 6〉 Medical applications in literature

Technique	Application & feature	Case
Regression	■ PCADP (combination interrelation + decision tree) proposed - large data set applicable - prediction accuracy improved compared to stand-alone logistic regression, neural net, decision tree ■ efficiency test: 5-cross validation	[19]
Neural network	■ same as regression	[19]
	■ combination of neural net & k-NN - detection rate improved 5%	[20]
SVM	■ CART, C5.0, SVM algorithm applied - accuracy 74%, 94.9%, 82.6% respectively	[6]
	■ SVM applied - relatively low accuracy of 61.8%	[35]
k-NN	■ same as neural network	[20]
Decision Tree	■ same as SVM	[6]
	■ same as neural network	[19]

회귀분석 적용 기법을 사용한 연구 사례에서는 U-health 환경에서 판별 기법이 갖추어야 할 조건인 유연한 구조, 실시간 처리, 지속적인 개선, 판별과정의 모니터링 등에 부합하는 통계학적 질병예측 방법인 PCADP (Personalized Computer Aided Diagnosis Probability) 기법을 제안하였다[20]. PCADP 배열 값은 사용자 피드백 정보에 따라 유동적으로 변화되어 유연한 구조를 갖고 통계기반의 질병 예측 기법으로 하여 실시간 처리가 가능하게 하였다. 다량의 데이터가 입력으로 들어오면 연산에 오랜 시간이 걸리고 패턴을 찾기가 어려워진다는 한계를 해결하기 위해서 집합 조합(set association)과 의사결정나무를 이용하여 질병 예측시스템에 들어오는 데이터의 수를 줄여서 다량의 데이터에도 적응 가능하며 예측의 정확도를 높일 수 있도록 구성하였다. 더불어

U-health 온톨로지 (ontology) 시스템을 제안하고 온톨로지 관리기 모니터를 활용함으로써 피드백 정보의 축적을 통한 지속적인 개선과 시각적인 표현이 가능케 하였다. PCADP 기법의 성능 및 효율성을 검증하기 위한 장치로 예측 정확도를 측정하는 5-cross validation 방법을 사용하였다. 지수형 회귀분석기법과 신경망기법, 의사결정규칙기법, 규칙기반기법과의 비교에서 PCADP기법은 효율적인 정확성과 예측성을 갖고 있는 것으로 나타났다. 유연한 구조, 실시간 처리, 지속적인 개선에서 의사결정 규칙기법과 함께 지원이 됨을 확인할 수 있었고 특히, 판별과정의 모니터링이라는 유일한 장점을 지니고 있음을 보였다.

SVM 적용기법을 사용한 사례에서는 먼저, 음성 특성간의 관계를 밝히는 것을 최종 목표로 하여 어떤 음성학적 특성이 사람의 체질을 구별할 수 있게 하는지를 찾아봄으로써, 목소리의 여러 음성학적 변수 중 사상체질에서 구분한 체질과 관련이 깊은 변수를 찾아내어 음성특성과 체질간의 상관관계를 규명하는 것을 목표로 연구되었다[6]. 이 논문에서 사용한 CART 분석 알고리즘은 의사결정나무에 기초한 분류 및 예측 방법으로 학습 레코드를 목표 필드와 유사한 집단으로 분리하기 위해 반복 분할을 한다. CART 모델의 수행을 위해서는 하나 이상의 입력필드와 하나의 출력필드가 요구된다. C5.0 분석 알고리즘은 입력필드와 출력필드를 기반으로 의사결정나무 구조를 생성하며, 일반적으로 이 알고리즘을 ID3(Iterative Dichotomizer 3)라고 한다. 체질분류 정확도에서 CART 알고리즘은 74%, C5.0 알고리즘은 94.9%, SVM 알고리즘은 82.6%의 유의미한 결과를 보여주었다. 이외에도 임상 이벤트를 추출한 후, 임상 이벤트들의 관계를 임상 인과관계로 정의하여 시간 순서에 따라 관계를 추출하는 기법을 제안하였다[36]. SVM은 이벤트 인과관계 패턴을 4가지로 분류하는 데 사용되어 61.8%의 정확성을 보여주었다.

k-NN과 인공신경망의 복합 적용이 진단의학연구에서 활발히 이루어지고 있다[20].영상 기술의 발달로 정교한 의료영상의 확보가 가능해졌으나 획득된 데이터의 양이 방대해짐에 따라 효율적인 자동분석법의 개발이 요구되었다. 의료영상 분석을 위해서는 의료영상 분할, 영상정합, 컴퓨터 보조진단 시스템, 내용기반 검색 등이 필수적인데 유방암 종괴 검출, 뇌 CT 분석 등의 의료진단에서 높은 검출력(독립 수행시 각각 84.5%, 89.1%, 복합수행시 91.2%로 향상)과 낮은 오류율(0.161)을 보여주었다.

4.3 기타

기계학습이 공학 분야에서 활발히 연구되고 인공지능과 함께 눈부시게 발전함에 따라 여러 분야에 전파되기 시작하였다. 공학과 의학 분야에서의 응용 사례에서와 같이 이론적 측면의 연구보다는 실용적으로 응용할 수 있는 방안을 도출하고자 연구 및 실험을 진행하고 있다. 본 절에서는 공학과 의학을 제외한 사회·자연과학, 농수해양, 예술·체육, 인문과학 분야에서의 적용 연구 사례를 살펴보고자 한다. 최근에는 소셜네트워크를 기반으로 한 감성 분석을 통해 의미 있는 정보를 발견하려는 연구가 국내외에서 활발히 이루어지고 있는 추세이다. 기타 분야에서 이용되는 기계학습 기법으로는 SVM이 가장 빈번하게 이용되고 있음을 확인할 수 있다. 성능 면에서도 SVM이 가장 높은 정확도를 나타냈다.

먼저, 강우의 시공간적 분포의 불규칙한 변동성을 고려한 홍수예보 의사결정지원시스템 개발을 위해 기존의 인공신경망 모형과 SOM을 전처리 과정으로 이용한 인공신경망 모형의 예측값과 관측치를 비교한 결과, SOM 전처리 과정을 적용하였을 때, 강우-유출관계를 다양한 패턴으로 구분할 수 있었으며, 구분된 패턴들은 관측된 강우현상과 유출 양상을 복합적으로 반영하였다[26]. 또한, 기존의 인공신경망 모형은 예측치가 선행 관측치를 따라가는 persistence 현상이 명확하게 나타났으나 SOM의 적용 후에는 persistence 현상이 제거되었다. 역전파 학습 알고리즘만을 적용한 모델은 persistence로 인하여 적중률 0%를 초래할 수 있기에 SOM과 역전파 학습 알고리즘의 결합으로 75%의 적중률과 주의보 수위 발령 시점의 평가를 100%로 향상시켰다.

다음으로, 주가 예측의 정확도를 높이기 위해 SNS와 뉴스기사 데이터를 동시에 이용한 다수의 주가예측 모형을 생성한 후 정확성을 비교하였다[23]. SVM에서는 SMO 알고리즘과 Polynomial kernel을 사용하였다. 인공신경망은 4개의 입력노드(input nodes)와 3개의 출력노드(output nodes)를 사용하였고 시행착오법(trial and error)을 통해 은닉층(hidden layers)의 수를 결정하였다. 로지스틱 회귀분석, 베이지안 네트워크, 인공신경망, SVM의 4가지 기법을 WEKA 소프트웨어에서 10중 교차

검증(10-fold cross-validation) 방식을 사용하여 비교한 결과, 4가지 기법 모두 약 80%의 정확도를 보여주었지만, SVM의 정확도가 83.8%로 가장 높게 나타났다.

그리고 트위터를 대상으로 한국어로 작성된 트윗의 감성 분류에 효과적인 기계학습을 살펴보고, 형태소와 음절 방식을 사용하여 한국어 트윗의 감정 분류에 적합한 자질추출 방식을 확인하는 것을 목적으로 한 연구 사례가 있다[24]. 이 논문에서는 구조적인 텍스트에 대한 분류는 기계학습으로 정확히 분류할 수 있으나 트위터에 사용되는 신조어와 같은 비구조적인 텍스트에 대한 분류는 어렵다는 점에서 어떤 기계학습이 비구조적인 텍스트의 분류에서 가장 큰 정확성을 보이느냐를 비교하였다. SVM과 NB에서 각각 4가지의 기법을 적용하여 분류 정확성을 비교한 결과, SVM에서 가장 높은 정확도로 84%가 나와 NB보다 더 높은 정확성을 띄는 것으로 나타났다. 원인은 SVM과 NB의 알고리즘의 특징에서 볼 수 있다. SVM은 불필요한 데이터를 버리고 유의미하다고 판단되는 데이터에서만 학습을 하는 특징을 갖고 있다. NB는 각 데이터가 독립적이라는 가정 하에 학습을 진행하기 때문에 종속적인 데이터에 대해서 오류를 범하는 경우가 생기게 되어 정확성에 악영향을 미치게 된다.

마지막 사례로 언어분석 결과를 이용하지 않고, 단순한 자질들만을 사용해서 효과적인 경쟁관계 인식을 하였다. 분류 알고리즘으로서 분류 문제에 많이 사용되는 ME(Maximum Entropy), CRFs(Conditional Random Fields), SSVMs(Structural Support Vector Machines)를 적용하여 비교하고 SSVMs의 학습속도 향상을 위해 Pegasos 알고리즘을 적용하였으며 오류에 대한 필터링 알고리즘도 적용하였다[34]. 이 논문에서 제안한 필터링 알고리즘은 문장별 경쟁유무 분류 기반 필터링, 스팸 분류 기반 필터링, 거리 제약 기반 자질 필터링 3가지를 결합한 방법이다. 각각에 대해서는 6% 미만의 성능 개선 효과라는 미미한 결과를 보여주었지만 제안한 결합 모델은 14.5% 성능 개선되었다.

5. 기계학습 기법의 적용 및 개선법

기계학습의 적용단계는 크게 3가지로 나눌 수 있다. (1) 데이터 수집, (2) 알고리즘을 통한 데이터 학습, (3) 알고리즘에 대한 유의미성 테스트이다. 기계학습 사용을 위해서는 특정한 상황을 잘 설명해줄 수 있는 데이터를 수집하고 수집한 데이터 집합에서 자질을 추출한다. 이후 분류 및 분석 알고리즘을 적용한 뒤, 테스트가 가능한 알고리즘을 적용하여 분류 및 분석 알고리즘의 정확성을 근거로 데이터 분류의 정확성과 유의미성을 판단한다.

기계학습의 적용단계는 앞서 분석한 기계학습 적용 논문들의 구성과 실험절차를 바탕으로 일반화한 것이다. 기계학습 기법의 적용과 개선을 위한 사례로 초저음파 식별에 관한 연구가 있는데 그 접근법을 보면 다음과 같다[4]. 먼저 (1) “시스템 설계 및 구현” 단계에서 초저음파 분류를 위한 시스템을 설계하고 대상 자료 및 구성에 관하여 논하였다. 여기에서 DTRA 데이터베이스에서 얻은 데이터를 구성하고 실험에 유용한 데이터를 선별하는 작업을 통해 연구대상을 선정하였다. 다음으로 초저음파 분류를 위한 과정을 설계하며 자질 및 기계학습 기법의 선택과 분류과정을 (2) “연구방법” 단계에서 수행하였다. 자질 추출 방법은 cepstrum 방식을 이용하였고 기계학습 기법은 베이즈망, SMO, LMT, RandomForest 등 총 10개를 선택하였다. 이후 이를 (3) “실험환경” 단계에서 알고리즘을 적용하고 결과에 대한 알고리즘별 성능분석을 진행하였다. 이를 위해 소프트웨어 WEKA를 사용하여 10-fold cross validation으로 오류율 측정을 한 결과, SMO (21.79%), LMT (21.84%), 베이즈망 (23.01%), RandomForest (25.52%), J48 (25.88%) 등으로 보고되었다.

앞서 논한 사례들로부터 유추할 수 있듯이, 현재 기계학습에 대한 연구들은 주로 학습 기법의 적용 및 성능 비교가 주를 이룬다. 각 분야마다 어떤 기법이 가장 우수한 성능을 보이는지, 실험과 검증을 통해 최적의 기법을 찾아내려는 노력이 이어지고 있다. 이러한 과정 속에서 기존에 있던 기법에서 더 나아가 하나의 기계학습 기법에 다른 알고리즘을 추가하는 방법을 통해 성능을 향상시켜 가며 새로운 기법을 개발하는 움직임이 보인다. 본 절에서는 기계학습 기법의 성능을 개선할 수 있는 방안을 사례별 분석을 통해 일반화하여 제시한다. <Table 7>에서 보는 바와 같이 기계학습 기법의 성능을 개선할 수 있는 방법은 크게 다중 기계학습 구조와 $+\alpha$ 기계학습 구조로 나눌 수 있다.

〈Table 7〉 Performance improvement methods in literature

Method	Literature
Multi-layer structure	[4], [7], [15], [18], [20, [21], [26]
$+\alpha$ structure	[4], [6], [7], [12], [19], [32], [34]

5.1 다중 기계학습 구조

다중 기계학습 구조는 앞서 언급한 기계학습 적용 3단계 중 (2)단계인 학습 클래스에 속하는 기존 2개 이상의 기계학습 알고리즘을 결합하는 방식으로 형성된다. 이는 경우에 따라 앙상블 학습 알고리즘이라 불리기도 한다. 예를 들어, 기존 기계학습 알고리즘 A와 B가 있을 때, A가 B의 단점에 대해 대응되는 장점을 갖고 있어 그 점을 보완하기 위해 A와 B를 결합하는 식의 형성도 가능하고 기존 알고리즘 A에 대해서만 반복 적용을 하여 학습 성능을 향상시킬 수도 있다. 이러한 다중 기계학습 구조의 개발 및 적용은 공학 분야에서 주로 시도되어 왔다[7,15,18].

먼저, 첫 번째 사례인 모바일 로봇의 자율주행 알고리즘에 관한 연구에서는 다른 특징을 갖고 있는 두 알고리즘을 결합하여 성능을 향상시켰다[7]. 제안된 시스템인 Actor-critic method는 인공신경망과 Q-learning을 결합한 것으로 로봇이 인간과 같이 영상 분석을 통해서만 주행을 할 수 있도록 하였다. 인공신경망은 원본 영상에서 가공한 데이터를 입력 데이터로 하여 주행 환경을 인식한다. 이렇게 정의된 주행 환경은 이후 Q-learning에 적용되어 로봇의 주행 방법의 학습이 이루어지게 된다. 이러한 인공신경망과 Q-learning의 학습에 대한 역할 분담은 이 논문에서 목적으로 하였던 영상 분석만으로 주행을 하는 것을 가능케 하였다.

두 번째 사례인 근전도 생체신호를 이용하여 안구이동을 추적하는 연구에서는 분류기와 SVR을 차례로 거치는 이중 기계학습 구조를 적용하여 분류기 단계에서 시선 이동시의 특징 정보만을 미리 선별하도록 하였다[15]. 이러한 경우에 잡음에 대한 강인함 확보라는 장점을 갖는 것도 중요하지만 주목해야 할 부분은 분류기와 SVR이 각각의 역할을 분담하여 일을 수행한다는 것에 있다. SVR이 분석해야할 정보를 시선 이동시에 나타난 안전도 정보를 한정하고 이외의 부분은 분류기가 분담하는 역할을 하기 때문에 결과적으로 안구이동추적의 분석 성능이 향상될 것을 예측하였고 결과도 동일하게 보고되었다.

세 번째 사례인 가속도 데이터로부터 낙상을 감지하는 연구에서는 동일한 알고리즘을 단계별 중복 학습시켜서 성능을 향상시켰다[18]. 제안된 시스템에서는 SVM을 1단계와 2단계로 구성하여 1단계에서는 여러 가지 행동에 대해서 눕기/낙상 대 나머지 행동으로 분류되게 하였고 2단계에서는 눕기 대 낙상으로 분류하게 하여 낙상에 대한 정확도, 민감도, 특이도를 개선하였다. SVM을 한 번만 사용하면 낙상을 검출해내기 위해서 가속도 데이터에서 많은 양의 특성을 추출해 낼 경우, 계산양의 증가와 많은 양의 메모리 공간을 요구하여 성능에 영향을 미치게 된다. 이 논문은 그러한 문제점을 보완하기 위해 SVM을 중복 사용하여 비교적 간단한 특성인 SMV(Signal Magnitude Vector)와 각도, 최대값, 최소값을 이용하여 낙상을 검출하였다. '이중SVM 알고리즘'을 적용한 결과, 정확도(98.64%→99.32%), 민감도(90.48%→100%), 특이도(100%→99.21%) 측면에서 매우 높은 수치를 기록하여 그 유효성을 검증하였다.

5.2 $+\alpha$ 기계학습 구조

$+\alpha$ 기계학습 구조는 하나의 기계학습 알고리즘과 학습 이전과 이후의 데이터를 최적화해 줄 수 있는 다른 알고리즘을 결합하는 방식이다. 즉, 기계학습 적용 3단계 중 (1), (3)단계에서 학습 데이터를 학습 또는 적용 대상에 맞게 최적화하여 2단계에서의 기계학습 알고리즘을 통한 학습의 유효성을 개선 및 향상시키는 것을 목적으로 한다. 그 외에도 분류기에 직접적으로 결합을 하여 분류기 자체의 성능을 높이는 방법이 기계학습 구조에 포함되며 주요 사례는 다음과 같다.

첫 번째 사례는 생성한 feature가 모두 동일한 정보나 의미, 중요도를 갖는 것이 아니라는 것을 감안하여 학습을 시키는 단계 이전에 실험 환경에 맞는 최적의 feature 조합을 찾아내기 위해 feature 선택 알고리즘을 사용한다[7]. Feature 선택 알고리즘은 예측기 또는 패턴 분류기의 입력으로 사용될 많은 입력 후보 feature중에서 출력에 영향력이 큰 입력 feature를 찾아내는 과정이다. 특히, 데이터가 복잡하여 어떤 feature를 이용해야 좋은 성능을 얻을지 모르는 경우에 이를 사용하는 것이 좋다. 결과적으로 feature 선택 알고리즘을 사용함으로써 예측 및 분류 성능을 높이고 학습 시간을 단축하며 입력과 출력 사

이의 관계를 이해하는데 도움을 얻는 것을 기대할 수 있다. 단, feature 선택 알고리즘도 학습 알고리즘과 같이 다양한 종류가 존재하므로 그 중 목적에 적합한 알고리즘을 선택해야 하고, 이를 위해 feature 선택 알고리즘 적용 후 성능 분석을 위한 평가 방법에 대해서 이해해야 한다. 또한 적용 가능한 학습 알고리즘 선택에 제한이 있음을 감안해야 한다. Feature 선택 알고리즘을 사용함으로써 최적 feature 조합으로 학습된 인공신경망의 검증 데이터 출력 비교에서 기대한 출력값에 근접한 출력값을 보여주었으며, 로봇의 성공적인 자율 이동을 통해 그 유효성이 검증되었다.

두 번째 사례는 분류기의 적용 이후에 이를 피드백하고 사용 환경에 맞는 포맷의 조정이 이루어지게 하여 분류기의 성능을 개선시키고 유효성을 증가시켰다[12]. 여기서 핵심적인 기계학습 알고리즘은 선형회귀분석 알고리즘 이다. 이 분류기에 보정학습 알고리즘을 결합하여 선형회귀분석에서의 결과 값에 대한 피드백으로 잡음 보정 등의 역할을 이루어 성능을 향상시켰다. 또한, 평가모듈(EvM)로 학습된 새로운 제어 계수가 얼마나 적합한지를 평가하여 제어 계수의 사용여부를 판단하고 Merging 알고리즘이 추가되어 포맷 조정을 통해 학습된 제어 계수가 시스템 제어에 용이하게 하였다.

세 번째 사례는 분류기에 직접적인 보완을 가하여 성능을 높였다[34]. 분류 알고리즘으로서 Maximum Entropy(ME), Conditional Random Fields(CRFs), Structural Support Vector Machine(SSVMs)을 적용하면서 SSVMs의 학습속도 향상을 위해 Pegasos 알고리즘을 적용하였다. 또한, 오류필터링 방법을 분류기와 결합하여서 경쟁관계 중 긍정관계에 대한 인식 정확도를 높이는 방법론을 제시하였다. 더불어 오류필터링 방법은 문장별 경쟁유무 분류 기반 필터링과 스팸 분류 기반 필터링, 거리 제약 기반 자질 필터링의 결합으로 이루어졌다. 제안된 방법론을 기존의 일반 관계추출 방법론에 적용한 관계추출기와 비교하였고, 일반 관계추출방법에 비해서 긍정 정확도는 11%의 성능향상이 있었고, 전체정확도는 14.5%의 개선이 이루어져 92.2%의 매우 높은 정확도를 보였다.

5.3 다중 기계학습과 +α 기계학습 구조의 결합

이 외에도 이 두 구조를 결합하는 방법이 있다. 다양한 기계학습을 결합하고 각각에 대해 feature 또는 자질의 생성에서 최적화를 한 후 학습 결과값에 대해 상황에 맞게 포맷까지 하는 적용 단계로서 기계학습 적용단계 (1), (2), (3) 모두를 강화하는 방법이다. 앞서 언급한 첫 번째와 두 번째 사례의 기계학습 시스템 결합이 그러한 형태이다. 이 방법은 위의 두 방식에 비하여 최고의 성능을 보여 줄 수 있을 것이다. 그러나 학습 시스템의 영역이 지나치게 커진다는 단점이 있다는 것을 감안해야 한다. 학습 시스템의 부피가 커질 경우, 학습 시스템의 구성이 매우 어려워지고 구성과 구현 및 계산이 이루어지는 시간 즉, 학습이 이루어지는 시간이 지체되는 등의 패널티(penalty)가 작용할 수 있다. 이러한 경우, 정확성과 정밀성만을 요구하는 분야에서는 활용 가능성이 크지만 실시간으로 학습을 해야 하는 분야에서는 활용 가능성이 매우 낮아진다. 요즘 실생활에서 실시간 정보습득 및 제공이 필수 요소로 작용하는 사례가 증가하고 있는데, 이 경우 두 구조를 결합하는 방식을 택하는 것은 피하는 것이 좋다. 만약 그럼에도 불구하고 기계학습 시스템을 축소하거나 시간을 단축시켜 사용할 수 있는 경우에는 이 방법이 매우 효율적이다.

6. 결론

본 논문에서는 분야별 기계학습 적용 사례 분석을 실시하였으며, 이를 근거로 하여 기계학습 적용법과 성능 개선법의 일반화를 통해 기계학습 적용 실험에서 보다 우수한 성능을 보일 수 있도록 하는 방안을 기계학습의 구조적인 측면에서 조명하였다. 현재 기계학습에 대한 연구는 환경에 직접 사용하는 적용 연구가 주를 이룬다. 적용한 학문 영역은 크게 공학, 의학, 기타(사회·자연과학, 농수해양, 예술·체육, 인문과학)로 나누었다. 공학에서 적용한 기계학습 기법에는 SVM과 의사결정나무가 가장 많았다. 사용 빈도수와 같이 정확성 또한 SVM이 매우 높은 정확도를 보여주었다. 의학 분야에서는 신경망, SVM, 의사결정나무가 기계학습 기법으로서 가장 빈번하게 적용되고 있다. 성능은 의사결정나무에 기초한 C5.0 알고리즘과 인공신경망이 가장 좋은 것으로 나타났다. 반면, 공학에서 높은 성능을 보인 SVM은 비교적 저조한 성능을 보여주었다. 기타 분야에서 이용되는 기계

학습 기법으로는 SVM이 가장 빈번하게 이용되고 있음을 확인되었다. 성능에서도 SVM이 가장 높은 정확도를 나타냈다. 따라서, 공학과 기타 분야에서의 적용에 있어서는 SVM이 가장 적합한 기법이라고 할 수 있다. 반면, 의학 분야에서는 SVM의 빈번한 이용에도 불구하고 의사결정나무가 더 나은 성능을 보여주어 적합한 기법을 선정하자면 의사결정나무를 택할 수 있다. 추가로 학습 기법의 성능을 평가 및 비교하는 소프트웨어로 WEKA가 많이 사용되었다는 특징도 발견할 수 있었다. 이러한 결과는 2000년대 초반부터 2015년까지의 연구 사례를 포함한 결과이므로 경향 조사라기보다는 빈번히 사용되는 기계학습 기법 조사라고 하는 것이 더 옳은 표현이다. 기계학습 적용 사례분석을 통하여 두 가지 교집합을 도출할 수 있었는데 첫 번째가 기계학습의 적용 방식이다. 기계학습의 적용단계를 (1) 데이터 수집, (2) 알고리즘을 통한 데이터 학습, (3) 알고리즘에 대한 유의미성 테스트로 나누었다. 문헌에서는 일반적으로 이 세 단계의 적용단계와 그 적용방식에 있어 유사성을 가진다. 두 번째로 성능 개선 방법이다. 성능 개선 방법은 크게 두 가지로 나눌 수 있다. 첫째, 다중 기계학습 구조이다. 이는 기계학습 적용 3단계 중 (2)단계인 학습 클래스에 속하는 기존에 있던 2개 이상의 기계학습 알고리즘을 결합하는 방식으로 형성된다. 둘째, 기계학습 구조이다. 이는 하나의 기계학습 알고리즘과 학습 이전과 이후의 데이터를 최적화해 줄 수 있는 다른 알고리즘을 결합하는 방식이다.

보다 향상된 기계학습 시스템의 구현을 위해서 앞으로 많은 작업이 남아있다. 본고에서 논한 여러 사례들을 종합하여 다중 기계학습 구조와 $+a$ 기계학습 구조를 결합한 시스템의 구현과 함께 실행과정의 성능평가에 관한 연구를 이어가고자 한다.

ACKNOWLEDGMENTS

This work was supported by a research grant from Seoul Women's University(2015).

REFERENCES

[1] SAS Institute, "Machine Learning: What it is & why it matters", http://www.sas.com/en_us/insights/analytics/machine-learning.html (December 1, 2015)

[2] A. L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development", Vol. 3, No. 3, 1959.

[3] B.-T. Zhang, "Next-Generation Machine Learning Technologies. Journal of Computing Science and Engineering", Vol. 25, No. 3, pp. 96-107, 2007.

[4] Jingu Lee, "Comparative Study of Various Machine Learning Techniques for Infrasound Signals Associated with Volcanic Eruptions. Master's thesis". Korea University. 2014.

[5] A. Cannata, P. Montalto, M. Aliotta, C. Cassisi, A. Pulvirenti, E. Privitera and D. Patane, "Clustering and classification of infrasonic events at Mount Etna using pattern recognition technique. Geophysical Journal International", Vol. 185, No. 1, pp. 253-264, 2011.

[6] Sangbeom Kim, "A Study on Constitutional Classification Using Speech Features and Machine Learning Methods. Master's thesis", Daejon University. 2012.

[7] Jeongmin Choi, "Vision based self learning mobile robot based on machine learning algorithm. Master's thesis", Chungnam National University. 2009.

[8] C. Gaskett, L. Fletcher, and A. Zelinsky, "Reinforcement Learning for a Vision Based Mobile Robot. Intelligent Robots and Systems, IEEE International Conference on", pp. 403-409, 2000.

[9] C. V. Regueiro, J. E. Domenech, R. Iglesias, and J. Correa, "Acquiring contour following behavior in robotics through Q-learning and image-based states. PWASET", Vol 15, 2006.

[10] Sangjun Park, "Content-Based Classification of Musical Genre using Machine Learning. Master's thesis", Seoul National University. 2002.

[11] Shin Hwi Yun, "Estimation of Vessel Service Time Base on Machine Learning. Master's thesis", Pusan National University. 2009.

[12] Mi-Sun Moon, Kang Song, and Dong-Ho Song, "Aviation Application: UAS Automatic Control

Parameter Tuning System using Machine Learning Module. Journal of the Korean Institute of Navigation", Vol. 14, No. 6, pp. 874-881, 2010.

[13] Y. Abe, M. Konosho, J. Imai, R. Hasagawa, M. Watanabe and H. Kamiio, "PID Gain Tuning Method for Oil Refining Controller based on Neural Networks. Proc. of the Second Intl. Conf. on Innovative Computing, Information and Control", 2007.

[14] J. Lu, O. Ling and J. Zhang, "Lateral Control LawDesign for Helicopter Using Radial Basis Function Neural Network. Proc of the IEEE Intl. Conf. of Automation and Logistic", August. 2007.

[15] Gyeong-Woo Gang, "Development of bio-signal based eye-tracking system using dual machine learning structure. Master's thesis", Catholic University. 2013.

[16] Hyunsin Park, Sungwoong Kim, Minho Jin, and Chang D. Yoo, "The latest machine learning-based speech recognition technology trends. The Magazine of the IEEK", Vol. 41, No. 3, pp. 18-27, 2014.

[17] Haneol Kim, "Machine learning to detect garbage collecting SSDs and its use to increase performance predictability. Master's thesis", Hongik University. 2015.

[18] Dongjin Jung, "A Study on Effectiveness of Machine Learning Method for Fall Detection based on Extracted Feature from Acceleration Data. Master's thesis", Inje University. 2015.

[19] Byoung-Won Min, Yong-Sun Oh, "Improvement of Personalized Diagnosis Method for U-Health. Journal of Korea Contents Association", Vol. 10, No. 10, pp. 54-67, 2010.

[20] Sang Cheol Park, Myung Eun Lee, Soo Hyung Kim, In Seop Na, and Yanjuan Chen, "Machine Learning for Medical Image Analysis. Journal of Computing Science and Engineering", Vol. 39, No. 3, pp. 163-174, 2012.

[21] S. C. Park, J. Pu, and B. Zheng, "Improving Performance of Computer-Aided Detection Scheme by Combining Results from Two Machine Learning Classifiers. Academic Radiology", Vol. 16, No. 3, pp. 266-274, 2009.

[22] Seunghak Yu, Sangheon Baek, and Seonro Yun, "Survey of Analysis Methods for Understanding Gene Expression Regulation Mechanisms Using Ensemble Learning. Journal of Computing Science and Engineering", Vol. 32, No. 10, pp. 38-43, 2014.

[23] Dongyoung Kim, Jeawon Park, and Jaehyun Choi, "A Comparative Study between Stock Price Prediction Models Using Sentiment Analysis and Machine Learning Based on SNS and News Articles. Journal of Information Technology Services," Vol. 13, No. 3, pp. 221-233, 2014.

[24] Joa-Sang Lim, Jin-Man Kim, "An Empirical Comparison of Machine Learning Models for Classifying Emotions in Korean Twitter. Journal of Korea Multimedia Society", Vol. 17, No. 2, pp. 232-239, 2014.

[25] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, "Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis. Technical Report HPL-2011, HP Laboratories", Vol. 89, 2011.

[26] Choen Lee Jeong, "Application of Artificial Neural Networks Technique for the Improvement of Flood Forecasting and Warning System. Ph. D. dissertation", Dongshin University. 2010.

[27] Coulibaly P, Anctil F, and Bobee B, "Daily Reservoir Inflow Forecasting using Artificial Neural Networks with stopped Training Approach. Journal of Hydrology", Vol. 230, No. 3, pp. 224-257, 2000.

[28] French, M. N., Krajewski, W. F., and Cuykendall R. R., "Rainfall forecasting in space and time using a neural network. Journal of Hydrology", Vol. 137, pp. 1-31, 1992.

[29] Jae-Hyun Seo, Yong-Hyuk Kim, "A Survey on Rainfall Forecast Algorithms Based on Machine Learning Technique. Proceedings of KIIS Fall Conference", Vol. 21, No. 2, pp. 218-221, 2011.

[30] M. N. French, W. F. Krajewski, and R. R. Cuykendall, "Rainfall Forecasting in Space and Time Using a Neural Network. Journal of Hydrology", Vol. 137, No. 1 - 4, pp. 1 - 31, 1992.

[31] Junyeob Yim, Byung-Yeon Hwang, “Predicting Movie Success based on Machine Learning Using Twitter. Journal of Korea Information Processing Society”, Vol. 3, No. 7, pp. 263-270, 2014.

[32] Seong-Jin Kim, Cheol-Young Ock, “Analysis of Korean Language Parsing System and Speed Improvement of Machine Learning using Feature Module. Journal of The Institute of Electronics and Information Engineers”, Vol. 51, No. 8, pp. 66-74, 2014.

[33] Miikkulainen, R. and Dyer, M. G, “Natural Language processing with modular neural networks and distributed lexicon. Cognitive Science”, Vol. 15, No. 3, pp. 343-399, 1991.

[34] ChungHee Lee, YoungHoon Seo, and HyunKi Kim, “Competition Relation Extraction based on Combining Machine Learning and Filtering. Journal of Computing Science and Engineering”, Vol. 42, No. 3, pp. 367-378, 2015.

[35] Jaeuk Seol, “Clinical causal relationship extraction based on machine learning and rule from discharge summaries. Master's thesis”, Chonbuk National University. 2014.

이 호 현(Lee, Ho Hyun)

・2014년 3월 ~ 현재 : 폴수학학교
・관심분야 : statistical learning, cognitive computing, stochastic modeling
・E-Mail : jesusnme@gmail.com

정 승 현(Chung, Seung Hyun)

・2014년 3월 ~ 현재 : 폴수학학교
・관심분야 : machine learning, data mining, artificial intelligence
・E-Mail : shark9206@gmail.com

최 은 정(Choi, Eun Jung)

・1997년 2월 : 서울여자대학교 컴퓨터학과(이학사)
・2000년 2월 : 서울여자대학교 대학원 컴퓨터학(이학석사)
・2005년 8월 : 서울여자대학교 대학원 컴퓨터학(이학박사)
・2006년 3월 ~ 현재 : 서울여자대학교 정보보호학과 교수
・관심분야 : 시스템보안, 암호, 빅데이터
・E-Mail : chej@swu.ac.kr