Journal of Convergence for Information Technology
Vol. 8. No. 1, pp. 201-206, 2018

ISSN 2586-1816
e-ISSN 2586-4440
DOI : https://doi.org/10.22156/CS4SMB.2018.8.1.201

실시간 데이터 분석의 성능개선을 위한 적응형 학습 모델 연구

구진희
목원대학교 정보통신융합공학부

A Study on Adaptive Learning Model for Performance Improvement of Stream Analytics

Jin-Hee Ku
Division of Information Communication Convergence Engineering, Mokwon University

요 약 최근 인공지능을 구현하기 위한 기술들이 보편화되면서 특히, 기계 학습이 폭넓게 사용되고 있다. 기계 학습은 대량의 데이터를 수집하고 일괄적으로 처리하며 최종 조치를 취할 수 있는 통찰력을 제공하나, 작업의 효과가 즉시 학습 과정에 통합되지는 않는다. 본 연구에서는 비즈니스의 큰 이슈로서 실시간 데이터 분석의 성능을 개선하기 위한 적응형 학습 모델을 제안하였다. 적응형 학습은 데이터세트의 복잡성에 적응하여 앙상블을 생성하고 알고리즘은 샘플링 할 최적의 데이터 포인트를 결정하는데 필요한 데이터를 사용한다. 6개의 표준 데이터세트를 대상으로 한 실험에서 적응형 학습 모델은 학습 시간과 정확도에서 분류를 위한 단순 기계 학습 모델보다 성능이 우수하였다. 특히 서포트 벡터 머신은 모든 앙상블의 후단에서 우수한 성능을 보였다. 적응형 학습 모델은 시간이 지남에 따라 다양한 매개변수들의 변화에 대한 추론을 적응적으로 업데이트가 필요한 문제에 폭넓게 적용될 수 있을 것으로 기대한다.

주제어 : 적응형 학습, 기계 학습, 최근접 이웃 알고리즘, 실시간 분석, 인공지능

Abstract Recently, as technologies for realizing artificial intelligence have become more common, machine learning is widely used. Machine learning provides insight into collecting large amounts of data, batch processing, and taking final action, but the effects of the work are not immediately integrated into the learning process. In this paper proposed an adaptive learning model to improve the performance of real-time stream analysis as a big business issue. Adaptive learning generates the ensemble by adapting to the complexity of the data set, and the algorithm uses the data needed to determine the optimal data point to sample. In an experiment for six standard data sets, the adaptive learning model outperformed the simple machine learning model for classification at the learning time and accuracy. In particular, the support vector machine showed excellent performance at the end of all ensembles. Adaptive learning is expected to be applicable to a wide range of problems that need to be adaptively updated in the inference of changes in various parameters over time.

Key Words : Adaptive Learning, Machine Learning, Nearest Neighbor Algorithm, Stream Analytics, Artificial Intelligence

1. 서론

스마트 홈, 헬스 케어, 스마트 카, 스마트 팩토리 등 사물인터넷 핵심 플랫폼이 발전하면서 인공지능 관련 시장이 확대되고 있다. 최근 인공지능에 대한 관심이 고조되면서 수많은 신생 기업과 인터넷 거대 기업들이 인공지

*Corresponding Author : Jin-Hee Ku(jhku@mokwon.ac.kr)
Received January 25, 2018
Revised February 07, 2018
Accepted February 20, 2018
Published February 28, 2018

능을 선점하기 위해 경쟁하고 있으며 인공지능에 대한 기업들의 투자와 채택이 크게 증가하고 있다. Narrative Science의 설문 조사에 따르면 2016년에 기업의 38%가 이미 인공 지능을 사용 중이며 2018년까지 62%로 성장할 것이라고 전망하였다. Forrester Research의 2017 보고서에서는 2016년에 비해 2017년 인공지능 투자가 300% 이상 증가 할 것으로 예측하였다. 또한 IDC는 인공지능 시장이 2016년 80억 달러에서 2020년에는 470억 달러 이상으로 성장할 것으로 예측하였다[1].

인공지능을 구현하기 위한 방법 중에서 특히 기계 학습은 분류, 예측, 변이 감지, 클러스터링과 같은 분야에 폭넓게 사용되는 기술이다. 일상생활이나 다양한 산업군에서 발생하는 데이터 스트림의 분석을 통해 비즈니스의 가치, 제품의 불량 여부, 기기 고장 예측, 구매 예측 등 일상생활의 많은 관심사항을 예측하는데 활용될 뿐만 아니라 금융 산업에서는 지난 20년 동안 시장 방향 예측[2], 감성 분석, 포트폴리오 최적화, 신용 위험, 파산과 같은 금융의 복잡한 문제를 해결하는데 서포트 벡터 머신, 신경망 네트워크, 의사 결정 트리, 강화 학습, 유전 프로그래밍 등의 기계 학습 모델이 널리 적용되어 왔다[3]. 그러나 이러한 주제들에서 다양한 매개 변수 체제로 인해 잡음이 많은 데이터의 비선형성 문제와 발견된 패턴에 즉각적으로 대응할 수 없다면 실시간 거래에서 예측성과 및 수익성을 기대하기 어렵다.

이와 같이 비즈니스에서 스트림 데이터의 실시간 분석은 매우 중요한 이슈다. 최근에는 다양한 IoT 디바이스로부터 생성되는 센서 스트림을 통해 실시간 모니터링을 제공하고 이상을 예측하며 변화를 학습할 수 있는 다양한 기계 학습 모델이 제안되고 있으나 이는 시간과 공간 다른 차원에서 특성들이 변화하지 않는다는 것을 전제한다. 따라서 시간이 지남에 따라 변화하는 통계 모델에 대해서 추론을 적응적으로 업데이트하기 위한 알고리즘을 설계할 필요가 있다. 본 연구에서는 정상인 상태를 학습하고 편차가 발생하면 새로운 모델을 신속하게 학습하는 적응형 학습 알고리즘을 제안하였다. 적응형 학습에서 알고리즘은 의사 결정을 내리고 기존 데이터 및 설정에서 이미 가지고 있는 정보를 기반으로 학습 프로세스를 조정할 수 있다. 본 논문의 구성은 다음과 같다. 2장에서는 적응형 학습과 관련 있는 기계 학습 모델 및 기저 알고리즘에 대한 관련 연구를 기술하고 3장에서는 스트림 분석을 위한 적응형 학습 모델에 대해서 기술한다. 4장에서는 연구결과를 기술하고 마지막으로 결론을 기술한다.

2. 관련 연구

2.1 스트림 분석과 기계 학습

기계 학습(Machine Learning)은 알고리즘과 통계의 결합을 통해 대량의 데이터로부터 학습한다는 아이디어에서 출발한다[4]. 오늘날 기계 학습은 대량의 데이터를 오프라인에서 수집하고 일괄적으로 처리하며 최종 조치를 취할 수 있는 통찰력을 제공하고 있으나, 작업이 오프라인으로 처리 될 수 있으며 작업의 효과가 즉시 학습 과정에 통합되지는 않는다[5]. 전통적인 단순 기계 학습에서는 실시간을 위한 추적(Tracking) 솔루션에 많은 관심을 기울이지 않았다. 이 경우 비즈니스 솔루션은 단순화되며 시간이나 공간 또는 다른 차원에서 변화하지 않는다는 것을 가정한다. 그러나 실제로 모든 모델은 시간이 지남에 따라 진화하고 관련된 엔티티가 다양하기 때문에 시간이나 공간 그리고 또 다른 차원에서 혼합을 조정하고 적응적인 변경이 필요할 수 있다.

이미 지도 학습(Supervised learning)에 대한 예제를 레이블로 지정했다면 간단한 기계 학습이 빠르다. 통계가 단어나 구문과 같은 기능을 결정할 때 인간의 직관보다 일반적으로 더 낫기 때문에 더 정확하다. 그러나 지도된 기계 학습의 가장 큰 단점은 레이블이 있는 예제가 필요하다는 것이다. 레이블이 너무 적거나 레이블이 전체 데이터세트를 나타내지 않으면 정확도가 낮거나 특정 도메인에만 한정된다[6]. 각 샘플은 매우 다양한 특성으로 이루어 질 수 있고 샘플의 모든 특성을 고려하는 것은 분류기의 학습과 수행 속도를 늦출 뿐만 아니라 예측 능력까지 떨어뜨리게 된다[7]. 최근 비즈니스에서 스트림 분석은 분야를 막론하고 매우 중요한 영역으로 인식되고 있다. 실시간 스트림 분석에서 데이터는 도착하자마자 처리되어야 하며 통찰력은 신속하게 생성되어야 한다. 본 연구에서는 이를 해결하기 위한 접근법으로 적응형 학습(Adaptive Learning) 모델을 제안하고자 한다.

2.2 분류 모델

분류에 사용되는 kNN(k-Nearest Neighbor)모델은 기계 학습 모델 가운데 가장 직관적이고 간단한 지도 학습

모델 중 하나이다. kNN은 학습에 필요한 데이터를 메모리에 기억만 하고 있다가 새로운 데이터의 태스크 요청이 오면 일반화 즉, 분류를 수행하기 때문에 인스턴스 기반 러닝(Instance-Based Learning) 이라고도 한다. 이러한 알고리즘의 일반적인 개념은 반복적으로 프로세스를 실행하여 현재 프로토타입 세트에 대한 특정 기준의 만족도를 확인하고 중지 조건이 충족 될 때까지 프로토타입을 추가하거나 삭제하는 것이다[8]. kNN 분류를 위해 데이터 집합을 축소하려는 목적으로 설계된 알고리즘이 CNN(Condensed Nearest Neighbor)이다 Table 1은 CNN 알고리즘을 나타낸 것이다.

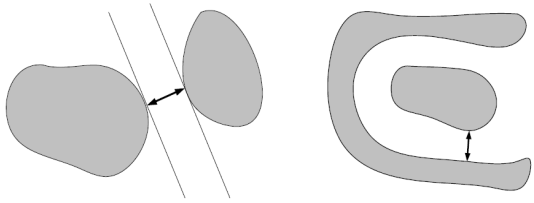
Table 1. Condensed nearest neighbor rule

```
Given
    X_Train Training set, with patterns x1, x2,···,xn
    |X_Train| Number of patterns in the training set
    CNN Condensed Nearest Neighbor set
    Additions A Boolean flag

Do

1       CNN = {x1}
2       Repeat
3           Additions=FALSE
4           For i = 2 to |X_Train|
5               Classify xi with CNN
6               If xi is incorrectly classified
7                   CNN = CNN ∩ { xi }
8                   Additions=TRUE
9       Until Additions = FLASE
```

이 알고리즘은 트레이닝 세트의 모든 패턴이 CNN 및 원래 분류 세트와 동일한 분류를 가지며 새로운 세트가 원본 세트보다 크지 않음을 보장한다. 실제로 우리가 CNN이라고 부르는 분류 세트는 원래 세트보다 훨씬 작다[9]. CNN 규칙은 CNN = x_1으로 시작하여 x_1이 트레이닝 세트 X_{Train}에서 무작위로 선택되면 CNN은 x_n의 모든 구성원을 검색한다. 그런 다음 가장 가까운 프로토타입의 레이블이 x_i의 레이블과 일치하지 않는 x_n의 구성원 x_i를 CNN에 추가한다. 알고리즘은 x_n의 모든 구성원이 흡수 될 때까지 x_n을 필요한 만큼 검색한다[10]. Chou, Kuo & Chang은 CNN이 채택한 약한 흡수 기준보다는 강한 흡수 기준을 사용하는 GCNN을 제안하였다. GCNN(Generalized Condensed Nearest Neighbor)은 SVM보다 몇 가지 유연한 분류 조건을 갖는다[11]. GCNN은 라벨이 다른 두 세트의 데이터가 양의 거리를 갖도록 요구된다. SVM (Support Vector Machine)은 한 쌍의 레이블을 한 번에 생성하지만, GCNN은 모든 라벨에 대한 프로토타입을 동시에 생성한다. CNN의 일관성은 프로토타입이 샘플이므로 항상 서로 일정한 거리를 유지한다는 사실에서 파생된다.



(a) (b)

Fig. 1. (a) A margin exists between two data sets, (b) A positive distance exists between two data sets

SVM은 주로 다루고자 하는 데이터가 2개의 그룹으로 분류될 때 사용하는 지도 학습 모델이다[8]. SVM에서 객체는 레이블이 다른 샘플 간의 마진을 최대화하여 분류하게 된다. Fig. 1의 (a)와 같이 SVM에서 마진은 두 개의 병렬 평면 사이의 간격으로 정의된다. 표본이 한정되어 있고 다른 레이블을 가진 표본 사이의 마진이 유지되면 SVM의 일관성은 보장된다[11]. Fig. 1의 (b)에서 레이블이 같은 두 개의 엔티티 (즉, 샘플 또는 프로토타입)는 편의상 동일하다고 가정한다. 그렇지 않으면 이질적으로 판단하며 따라서 이질적 샘플 사이에 0이 아닌 거리가 필요하다[12,13].

2.3 적응형 학습

적응형 학습은 이전 세대의 규칙 기반, 단순 기계 학습 및 기계 학습에 대한 강화 학습 접근 방식을 결합한다. 인공지능의 4세대 기술인 적응형 학습은 비정상적인 상황에서 분석의 정확도가 매우 높은 것으로 알려져 있다. 감성분석(Sentiment Analysis)에서 인공지능 2세대 기술인 단순 기계 학습과 3세대 기술 강화 학습, 그리고 4세대 기술 적응형 학습의 정확도를 비교한 결과 각각 60%, 69%, 95%로 밝혀져 적응형 학습 기술이 매우 높은 정확도를 나타내었다[6]. 비정상 상태에서 단순 기계 학습 알고리즘은 '비정상'을 나타내는 것을 학습하지만 비정상 패턴이 발생했을 때 적응형 학습은 새로운 모델을 신속하게 학습한다[5].

적응형 학습에서 알고리즘은 의사 결정을 내리고 기

존 데이터 및 설정에서 이미 가지고 있는 정보를 기반으로 학습 프로세스를 조정할 수 있다. 두 가지 유형의 적응 설정이 고려되어야 하는데, 첫 번째는 알고리즘이 데이터 집합의 복잡성에 적응하여 앙상블을 반복적으로 생성함으로써 새로운 가설을 추가하는 것이다. 두 번째 설정에서 알고리즘은 샘플링 할 데이터 포인트를 적응적으로 선택하도록 허용하는 보다 일반적인 접근 방식이다. 이 접근법은 특히 데이터 포인트를 획득하는 데 드는 비용이 매우 많이 드는 분야에서 아주 유용하다[14-17].

3. 스트림 분석을 위한 적응형 학습 모델

적응형 학습 알고리즘은 모든 샘플이 흡수될 때까지 필요한 만큼 샘플을 추가한다는 점에서 CNN과 유사하다[18]. 적응형 학습은 프로토타입 흡수 기준과 모델의 특성 면에서 CNN과 차이가 있다. CNN에서 모든 프로토타입은 샘플인 반면, 적응형 학습에서 프로토타입은 샘플 또는 샘플의 가중 평균이 될 수 있다.

Fig. 2는 적응형 학습 모델의 학습 알고리즘 개념도를 나타낸 것이다.

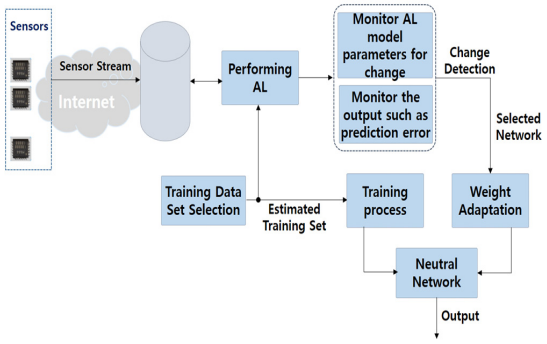


Fig. 2. Concept diagram of adaptive learning model

스트림 데이터는 서버에 수집된 데이터 프레임을 이용하여 모델 개발 데이터세트와 모델 검증 데이터세트로 구분하고 전처리 과정을 거친다. Fig. 1의 AL(Adaptive Learning)메커니즘에서 출력을 기반으로 학습 알고리즘의 목표는 네트워크 가중치를 현재 상태에 적응시키는 것이다. 이것은 이전 및 현재 네트워크 지식을 모두 활용하여 수행된다. 전자는 AL메커니즘에 의해 제공되는 반면, 후자는 트레이닝 세트 선택 모듈에 의해 제공된다. 트레이닝 세트 선택 모듈은 현재 환경에서 가장 대표적인 데이터를 선택하며 의사 결정 메커니즘에 의해 활성화된다. 의사 결정 메커니즘의 목표는 새로운 신경망을 적용해야 할 시기를 결정하는 것이다. 네트워크 성능이 적절하다고 판단되는 경우, 동일한 네트워크 가중치 및 구조가 분류를 수행하는 데 사용된다. 만약, 네트워크 성능이 저하되면 AL메커니즘과 트레이닝 세트 선택 모듈이 모두 활성화된다.

적응형 학습은 모든 단계에서 인간 분석가를 프로세스에 도입한다. 이것은 규칙 기반의 간단한 기계 학습 및 강화 학습 접근법과는 대조적으로, 인간은 프로세스 시작시 규칙 및 레이블 데이터만 작성하면 된다. 실시간 스트림 분석을 위한 적응형 학습 모델은 정상 상태에서는 '정상'을 나타내는 것을 학습하고 편차가 발생하면 새로운 모델을 신속하게 학습하도록 한다. 이것은 변화에 대한 적응형 학습 모델 매개 변수를 정기적으로 모니터링하고 예측 오차와 같은 비 추적 알고리즘의 출력을 모니터함으로써 가능하다.

4. 연구결과

6 가지의 표준 분류 데이터세트에 대한 실험은 기존의 원형 선택 방법과 비교하여 본 방법의 능력을 분석하기 위해 수행되었다. 대규모 데이터세트에 대해서는 엄청난 양의 계산 리소스와 학습 시간이 필요하기 때문에 샘플 수가 비교적 적은 데이터세트에 대해 선정하였다.

Table 2는 Iris, Ionosphere, Heart, Sonar, Wine, Zoo 데이터세트의 prototype, attribute, class의 수를 나타낸 것이다.

Table 2. Datasets used in the experiments

Datasets	Number of Samples	Number of Features	Number of Classes
Iris	150	4	3
Ionosphere	351	34	2
Heart	270	13	2
Sonar	208	60	2
Wine	178	13	3
Zoo	101	17	7

Table 3은 3개의 인스턴스 기반 분류 알고리즘 kNN, CNN, SVM을 6개의 데이터세트에 적용한 결과를 보여

주고 있다. 각 데이터세트에 대해 알고리즘들의 유의미한 정확도 차이는 없었다. 그러나 다양한 조합으로 학습할 경우에는 정확도의 차이가 있었다.

Table 3. Accuracy and training time of datasets

Datasets	Accuracy			Training Time
	CNN	SVM	kNN	
Iris	95.08	97.1	97.95	0.80
Ionosphere	91.72	93.23	93.46	9.45
Heart	89.35	89.43	91.9	3.1
Sonar	88.63	90.8	89.46	1.56
Wine	97.84	98.21	98.05	1
Zoo	93.5	94.43	96.66	0.83

알고리즘간의 확연한 차이는 없었으나 kNN이 가장 적은 학습 시간을 필요로 하였다. Table 1과 Fig. 3은 kNN이 모든 데이터세트에 대해 다른 두 알고리즘보다 더 정확한 정확도를 달성함을 보여주고 있다. kNN과 SVM의 정확도 차이는 평균 0.71% 로 나타났다. Sonar 데이터세트의 경우, SVM이 kNN보다 약간 우위에 있었다.

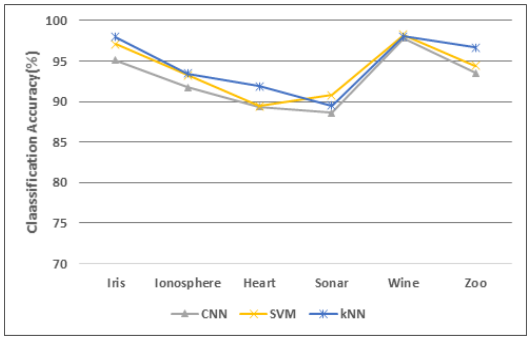


Fig. 3. Results of applying the algorithms to dataset

한편, 적응형 학습을 통한 결합 알고리즘 적용은 데이터세트에 대한 단일 알고리즘 적용의 경우와 유의미하게 정확도 차이가 있었다. CNN과 SVM 결합의 경우 학습을 위한 트레이닝 시간에서는 우위를 보였지만 정확도에서는 SVM을 단일 알고리즘으로 적용했을 때보다 더 낮았다. 그러나 대규모 데이터세트에 대해 SVM을 단일 알고리즘으로 적용하는 경우에는 엄청난 학습 시간을 필요로 하기 때문에 결과적으로 적응형 학습에서 전체 데이터세트에 대한 SVM의 적용은 효율이 낮게 나타났다.

5. 결론

인공지능을 구현하는 기술 중에 대표적인 기계 학습은 복잡한 비즈니스 문제를 해결하는데 서포트 벡터 머신, 신경망 네트워크, 의사 결정 트리, 강화 학습 등의 기계 학습 모델을 널리 적용하여 왔다. 그러나 다양한 매개변수 체제로 인해 시간이 지남에 따라 변화하는 모델과 데이터의 비선형성 문제 등은 실시간 예측에서 매우 중요한 문제로 부각되었다. 본 연구에서는 정상인 상태를 학습하고 편차가 발생하면 새로운 모델을 신속하게 학습하는 적응형 학습 모델을 제안하였다. 적응형 학습에서 알고리즘은 의사 결정을 내리고 기존 데이터 및 설정에서 이미 가지고 있는 정보를 기반으로 학습 프로세스를 조정할 수 있다. 결론적으로 kNN은 전체 데이터세트에 적용했을 경우 가장 적은 학습 시간을 필요로 하였고, CNN은 합리적인 축소율과 전체 데이터 집합의 정확도에 가까운 정확도를 달성하기 때문에 좋은 데이터 감소 알고리즘으로 확인되었다. 또한 SVM을 CNN 프로토타입에 대한 사후 프로세스로 적용하면 정확도가 향상 될 수 있을 것이다. 적응형 학습을 위해 선택된 앙상블의 체계적인 비교는 향후 연구에서 논의될 것이다.

REFERENCES

[1] R. Curran & B. Purcell. (2017). *TechRadar™: Artificial Intelligence Technologies, Q1 2017.* Forrester Research. https://kloudrydermcaasicmforrester.s3.amazonaws.com/mcaas/Reprints/RES129161.pdf

[2] J. H. Ku. (2017). A Study on the Machine Learning Model for Product Faulty Prediction in Internet of Things Environment. *Journal of Convergence for Information Technology, 7(1),* 55-60. DOI : 10.22156/cs4smb.2017.7.1.055

[3] Y. Zeng & D. Klabjan. (2017). Online Adaptive Machine Learning Based Algorithm for Implied Volatility Surface Modeling. arXiv preprint *arXiv : 1706.01833.*

[4] Tom M. Mitchell. (1997). *Machine Learning.* New York : McGraw-Hill Education.

[5] PG. Madhavan. (2016). *ADAPTIVE Machine Learning.* Data Science Central. https://www.datasciencecentral.com/profiles/blogs/adaptive-machine-learning

[6] R. Munro. (2015). *The fourth generation of machine*

learning : Adaptive learning. jungle light speed. http://www.junglelightspeed.com/the-fourth-generation-of-machine-learning-adaptive-learning

[7] Wikipedia. (2017). *AdaBoost.* Wikipedia. https://en.wikipedia.org/wiki/AdaBoost

[8] E. J. Kim. (2016). *Introduction to Artificial Intelligence, Machine Learning, and Deep Learning.* Seoul : Books Wiki.

[9] D. O. Gloria. (2002). *Ship Noise Classification.* Doctoral dissertation. New University of Lisbon, Lisbon.

[10] P. Hart. (1968). The condensed nearest neighbor rule. *IEEE Transactions on Information Theory, 14,* 515-516.

[11] VN. Vapnik. (1995). *The Nature of Statistical Learning Theory.* New York : Springer-Verlag New York, Inc.

[12] C. H. Chou, B. H. Kuo & F. Chang. (2006). The generalized condensed nearest neighbor rule as a data reduction method. *In Pattern Recognition, 2006. ICPR 2006. 18th International Conference on* (2, pp. 556-559). IEEE.
DOI : 10.1109/ICPR.2006.1119

[13] A. Abroudi & F. Farokhi. (2012). Prototype selection for training artificial neural networks based on Fast Condensed Nearest Neighbor rule. *In Open Systems (ICOS), 2012 IEEE Conference on* (pp. 1-4). IEEE..

[14] F. Chang, C. C. Lin & C. J. Lu. (2006). Adaptive Prototype Learning Algorithms : Theoretical and Experimental Studies. *Journal of Machine Learning Research, 7,* 2125-2148.

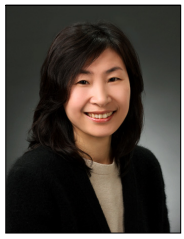
[15] A. Pratap. (2008). *Adaptive Learning Algorithms and Data Cloning.* Doctoral dissertation. California Institute of Technology, California.

[16] G. Castillo. (2008). Adaptive Learning Algorithms for Bayesian Network Classifiers. *AI Communications, 21(1),* 87-88.

[17] S. H. Min. (2016). Improving an Ensemble Model Using Instance Selection Method. *Journal of Society of Korea Industrial and Systems Engineering, 39(1),* 105-115.
DOI : 10.11627/jkise.2016.39.1.105

[18] V. Solutions. (2017). *Improving Predictions with Ensemble Model.* Data Science Central. https://www.datasciencecentral.com/profiles/blogs/improving-predictions-with-ensemble-model

구 진 희(Ku, Jin Hee) [정회원]



- 2001년 2월 : 충남대학교 컴퓨터과학교육학과(교육학석사)
- 2010년 2월 : 충남대학교 공업(컴퓨터)교육학과(교육학박사)
- 2010년 9월 ~ 현재 : 목원대학교 정보통신융합공학부 교수
- 관심분야 : 컴퓨터과학교육, 빅데이터, 기계 학습, 클라우드 컴퓨팅
- E-Mail : jhku@mokwon.ac.kr