Journal of KIIT. Vol. 18, No. 11, pp. 17-23, Nov. 30, 2020. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2020.18.11.17

저유소 휘발유 출하량 예측을 위한 머신러닝 예측모델 성능 비교

안성준*[1], 장동식*[2], 박상성**

Comparing Gasoline Shipment Prediction Model Performance of Oil Reservoir using Machine Learning

Sung-Jun Ahn*[1], Dong-Sik Jang*[2], and Sang-Sung Park**

요 약

오일 출하량 예측은 오일 재고 과부족에 큰 영향을 미치기 때문에 정유 이관 비용, 재고 관리 등 물류관리에 있어 중요한 부분이다. 하지만, 기존의 저유소 오일 출하량 예측방법은 업무 전문가들에 의해 정성적으로 수립한 판매계획을 기반으로 수립하기 때문에, 오일 재고 과부족 문제가 발생하고 있다. 이는 재고 부족으로 인한 판매 불가, 재고 과잉으로 인한 이관 비용 손해 등 회사에 큰 손실을 초래한다. 따라서 본 논문에서는 이러한 문제를 해결하기 위하여 머신러닝 기반 저유소 휘발유 출하량 예측 모델을 제안한다. 본 연구는 저유소 휘발유 출하량에 영향을 미치는 변수 18개를 선정하고, 이를 실증적으로 검증하기 위하여 휘발유 출하량 데이터는 정유회사 A의 B 지역 저유소 휘발유 출하량 데이터를 사용한다. 이 데이터들로 GLM, Random Forest, GBM, Deep Learning 그리고 앙상블 학습 모델을 포함하여 7가지 모델을 구축하고 각각의 성능을 비교분석한다.

Abstract

Predicting Oil shipment is an important part of logistics management, such as refining transfer costs and inventory management, as it has a great influence on oil stocks. However, the existing method of predicting shipments at oil reservoir does not predict accurate shipments. The reason is that the existing oil reservoir shipment forecast method is established based on the sales plan qualitatively established by business experts. Accordingly, problems such as inventory stockout/overstock transfer are occurring. This causes huge losses to the company. Therefore, in order to solve this problem, this paper proposes machine learning-based gasoline shipment prediction model of oil reservoir. This paper selects 18 variables that affect gasoline shipments at oil reservoir, and uses data on gasoline shipments from oil company A's area B to verify empirically. With these data, we construct 7 models including GLM, Random Forest, GBM, Deep Learning, and ensemble learning model, and compare and analyze the performance of each.

Keywords

shipment prediction, ensemble learning, gradient boosting machine, random forest, decision tree

* 고려대학교 산업경영공학부
- ORCID[1]: http://orcid.org/0000-0002-9887-9871
- ORCID[2]: http://orcid.org/0000-0002-6180-0664
** 청주대학교 빅데이터통계학과 교수(교신저자)
- ORCID: http://orcid.org/0000-0001-6804-2707

· Received: Oct. 21, 2020, Revised: Nov. 20, 2020, Accepted: Nov. 23, 2020
· Corresponding Author: Sang-Sung Park
Dept. of Big Data Statistics, CheongJu University,
298 Daeseong-ro, Cheongwon-gu, Cheongju-si, Chungcheongbuk-do, Korea.
Tel.: +82-43-229-8198-8203, Email: hanyul@cju.ac.kr

Ⅰ. 서 론

최근 IT 기술과 인공지능 기술이 발전함에 따라, 다양한 산업에서 인공지능 기술을 도입하려는 시도가 활발하게 진행되고 있다.

정유 산업에서도 마찬가지로 인공지능을 활용하여 회사의 손실을 최소화하고 이윤을 최대화할 수 있는 방안을 고안하고 있다. 인공지능을 산업에 적용한 첫 번째 단계로 저유소의 효율적인 재고관리를 통한 비용절감 방안을 고려하고 있다. 효율적인 정유사의 재고관리는 그림 1과 같이 판매수요예측, 생산수급계획, 공장 축하계획, 저유소 출하계획, 수송계획 등 계획기능이 함께 고려되어야 한다[1]. 이 중 판매수요예측은 타 계획 수립 시 기본 자료가 되기 때문에 매우 중요하다.

판매수요예측은 각 저유소에서 판매되는 유종별 출하량을 예측하는 과정이다.

현재는 그림 2와 같은 프로세스에 따라 각 정유회사의 전문가들이 경험을 통해 엑셀 기반의 수작업으로 저유소별 유종별 일별 출하량을 예측하고 있다.

하지만 이런 과정은 판매 계획 중심의 출하량 예측이기 때문에 정확도가 낮고, 시간이 오래 걸린다는 단점이 있다. 전문가 인터뷰를 해본 결과, 이러한 예측은 정확도가 MAPE 약 45.0% 수준이며, 이는 재고 부족(Stock out), 재고 과잉 이관 등의 문제를 초래하고 이에 따른 운영비용으로 손해가 발생하고 있다. 운영비용을 절감하기 위해서는 더 정확하게 출하량을 예측할 필요가 있다.

따라서 본 연구에서는 예측 시간을 단축하고 예측 정확도를 높이기 위해 앙상블 모델로 저유소 휘발유 출하량을 예측하는 방법을 제안하고자 한다.

본 논문은 총 5장으로 구성되어 있다. 1장에서는 본 연구의 배경, 목적과 구성에 대하여 설명하였다. 본 논문의 2절에서는 정유 산업에 대한 선행연구들을 기술하였다. 3절에서는 본 연구의 실험방법을 기술하였고, 4절에서는 이에 대한 결과를 기술하였다. 마지막 5절에서는 본 연구에 대한 결론과 향후 연구 방향에 대하여 제시하였다.

Ⅱ. 관련 연구

정유 산업에 관해 현재까지 진행된 연구들은 다음과 같다.

양순일(2007)은 저유소 유류재고의 특성을 분석하여 경제적 발주량을 제시하는 연구를 수행하였다[1]. 기존의 재고 관리의 문제점을 파악하고 문제점을 해결하기 위하여, 유류재고관리에 대하여 service level 현황을 파악하여 service level로 운영할 때 발생되는 여유물량으로 수익창출비용 및 건설비 절감 효과를 절감할 수 있는 방법론을 제안하였다.

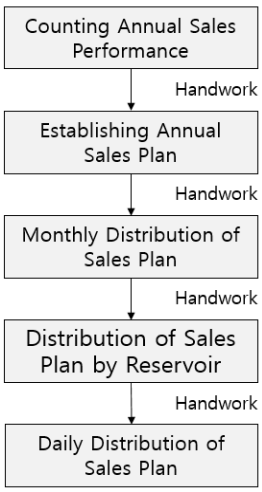

그림 2. 현업 전문가의 판매수요예측 프로세스
Fig. 2. Sales demand forecasting process by business experts

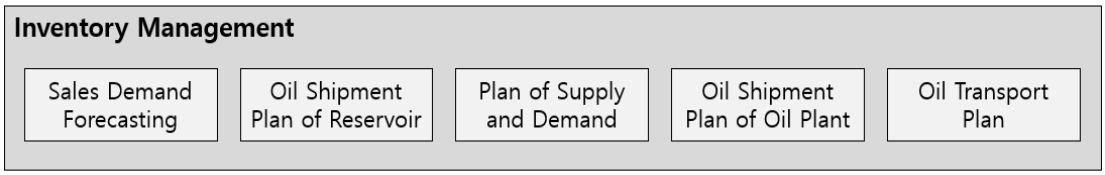

그림 1. 효율적인 재고관리를 위한 계획요소
Fig. 1. Plan factors for efficient inventory management

이인수(1990)는 SK(주)의 1987년도 판매 및 수송 실적자료를 활용하여 정유제품의 적정재고수준 설정에 대한 연구를 수행하였다[2]. 적정재고, 수송중 재고, 재주문점, 시뮬레이션 모형 등 적정재고수준을 설정하기 위한 지표들을 설정하고 경험을 통하여 도출된 SK(주)의 적정재고 휴리스틱 방법론에 일출하량과 수송수단에 관한 정보를 추가로 반영하여 재주문점 휴리스틱을 제안하였다.

Haiying Jia 외 2명(2019)은 선박 경로를 추적하는 AIS(Automatic Identification System)으로 수집한 오일 선박 데이터를 기반으로 기계학습을 적용하여 최적 출하지를 예측하는 모델을 제안하였다[3]. 위 연구에서는 2013년 1월 1일부터 2016년 3월 15일까지 남미 17개 국가의 43개 항구에서 선적된 14,745건의 석유 선적으로 구성된 데이터 세트와 선박데이터(seller's identity, loading port, cargo grade 등), 경제성 데이터(oil prices, crack spread)를 사용하였다. 위 데이터를 Decision Tree, Random Forests 그리고 Boosted Tree Technique 분류 알고리즘을 적용하여 출하지 분류를 통해 최적 출하지를 예측하였다.

기존의 연구들은 유류재고관리를 위한 경제적 발주량, 적정재고수준, 기계학습을 활용한 최적 출하지 예측 등 정유 산업에 대한 다양한 분야의 연구를 수행하였음을 확인할 수 있었다.

하지만 본 연구에서 제안하는 지역 저유소의 유류제품 출하량 예측에 대한 연구는 아직 미미함을 확인할 수 있었다.

III. 저유소 휘발유 출하량 예측을 위한 앙상블 예측모델

3.1 제안된 실험방법

본 연구에서는 지역 저유소의 휘발유 출하량을 예측하기 위해 그림 3과 같은 예측모델을 제안한다. 제안된 모델의 예측성능을 검증하고 비교하기 위하여 국내 정유회사인 A사의 데이터, 기상청, OPINET 데이터를 수집하여 제안된 4개 알고리즘과 앙상블 모델을 구축하고 그 성능을 비교분석하였다. 모델의 예측값을 실증적으로 비교분석하여 성능이 가장 높은 모델을 최종 모델로 선택하였다.

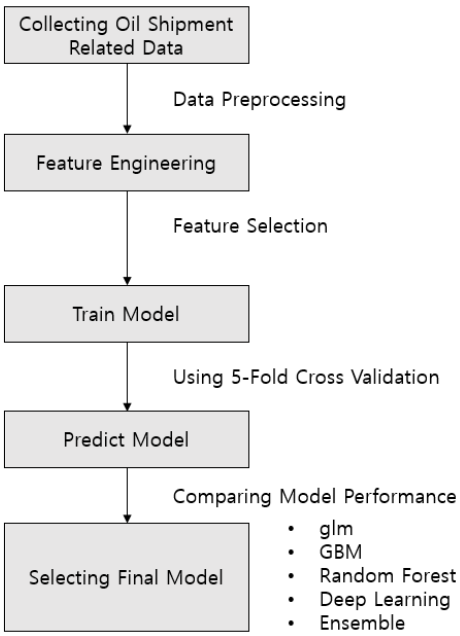

그림 3. 제안된 실험모델
Fig. 3. Proposed experiment model

3.2 데이터 수집

본 연구에서는 국내 정유업체인 A사의 B지역 저유소의 2014년 2월에서 2017년 5월까지의 휘발유 출하 데이터를 사용하였다. 성능평가를 위하여 국 2014년 2월부터 2017년 4월까지의 데이터는 학습데이터로, 2017년 5월 휘발유 출하량 데이터는 검증데이터로 사용하였다.

입력변수로 사용한 온도는 기상청에서 수집하였고, 국제유가는 OPINET에서 제공하는 데이터를 수집하였다. 탐색적 자료 분석(EDA, Exploratory Data Analysis)을 통해 출하량에 상관관계를 미치는 날짜(Date), 휴일 데이터는 네이버에서 제공하는 데이터를 수집하였다.

3.3 학습을 위한 변수 선택

출하량 예측값은 휘발유를 공장에서 각 저유소로 보내기 위한 이관 계획을 위한 데이터다. 이 데이터는 이관 시간, 이관 방법 등 이관 계획 수립 시 필요하기 때문에 그림 4와 같이 적어도 30일 이상을 일별로 예측하여야 한다. 하지만 온도, 국제 유가 등 입력변수의 30일 후 데이터가 없다는 문제점이 있다.

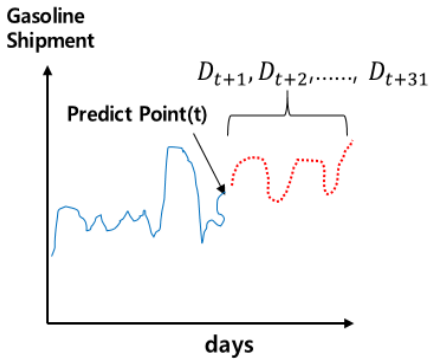

그림 4. 저유소 휘발유 출하량 예측 모델 개념도
Fig. 4. Concept diagram of predicting gasoline shipments

본 연구에서는 이러한 문제를 해결하기 위하여 30일 후 입력변수 데이터를 획득할 수 없는 국제원유가, 국제원유제품가, 일별 거래처수 등 16개 변수는 30일 전 데이터를 변수로 사용하고, 예측 가능한 날짜, 휴일 2개의 변수는 해당일 데이터를 사용하였다. 학습을 위해 앞서 기술한 18개 변수의 데이터를 사용하였다. 다음 표 1과 표 2는 각각 입력변수와 출력변수 목록이다.

표 1. 입력변수 목록
Table 1. List of input variables

Building variable	Variable type	Building variable	Variable type
Temperatures	Continuous	Diesel (0.001%)	Continuous
Heating day	Continuous	Diesel (0.05%)	Continuous
Cooling day	Continuous	Go (180cst/3.5%)	Continuous
Rain	Continuous	Nafta	Continuous
Snow	Continuous	DUBAI	Continuous
Gasoline (95RON)	Continuous	BRENT	Continuous
Gasoline (92RON)	Continuous	WTI	Continuous
Kerosene	Continuous	Holiday	Binary
Num_transaction_store	Continuous	Date	Continuous

표 2. 출력변수 목록
Table 2. List of input variables

Output variable	Output variable type
Gaosline shipment	Continuous

3.4 휘발유 출하량 예측 모델

본 연구는 최종 선택된 변수를 사용하여 앙상블 예측모델을 포함하여 7가지 방법으로 실험하였다.

첫 번째 Generalized Linear Moldels(GLM)은 선형 회귀 모델의 일반화다. 일반적인 선형 회귀 모델은 종속 변수가 독립 변수의 선형 함수이고, 종속 변수가 연속적이고 일정한 분산으로 정규 분포를 따른다고 가정한다. 독립 변수는 연속형, 범주형 혹은 2개의 조합일 수도 있다. GLM의 예로는 다중회귀분석, 분산분석, 공분산분석이 있다[4].

두 번째 Gradient Boosting Machine(GBM)은 X-Y 그리드 형태로 되어있는 데이터에 대한 예측 성능이 좋고, 기계학습 알고리즘 중 가장 예측 성능이 좋은 것으로 알려져 있다. 기본개념은 오분류된 데이터에 가중치를 주어 정확도를 높이는 적응형 Boosting 알고리즘이다. 가중치 계산 시, Gradient Boosting(경사하강법)을 이용하여 최적의 파라미터를 찾는다. 잔차에 대한 오류율을 낮추는 모형으로 과적합(Overfitting)의 문제가 발생할 가능성이 높으므로, Sampling, Penalizing 등의 Regularization 기술을 활용하여 모형을 개발하게 된다. Gradient Boosting의 가장 큰 장점은 복잡한 데이터를 효율적으로 분류한다는 점이다[5].

세 번째 Random Forest(RF)는 Breiman(2001)이 제안한 모델로 배깅(Bagging, Bootstrap aggregation)을 활용하여 추가적인 무작위성 레이어를 추가하는 모델이다. 데이터의 다른 부트스트랩(Bootstrap) 샘플을 사용하여 각 트리를 구성하는 것 외에도 랜덤 포레스트는 분류 또는 회귀 트리가 구성되는 방식을 변경한다. 기본 트리에서 각 노드는 모든 변수 중 최상의 성능을 보이는 노드를 사용한다[6].

네 번째 Deep Learning(DL)은 인공신경망 알고리즘의 히든 레이어를 여러 층을 쌓은 구조로 심층신경망으로 학습하여 예측 성능은 뛰어나다는 장점이 있지만 학습데이터에 대한 과적합과 학습속도가 느리다는 단점이 있다[7].

마지막으로, 앙상블 학습모델은 다양한 알고리즘을 배깅, 부스팅(Boosting) 그리고 스태킹(Stacking)으로 조합하여 우수한 성능을 나타내는 장점이 있다

[8]. 이러한 장점으로 인하여, 최근 예측 관련된 문제들에 다양하게 적용되고 있다.

강윤희 외 5명(2013)은 기후 예측을 위하여 전지구 대기 모델을 위한 앙상블 기후 시뮬레이션 CAM3(Community Atmosphere Model)을 활용하여 100년 이후 대기상태를 예측하였다[9].

박상훈 외 2명(2017)은 의학, 공학 분야에서 동작 상상 뇌파를 분류하는데 사용하는 CSP(Common Spatial Pattern) 방법을 앙상블 모델에 적용하여 기존의 CSP, R-CSP, FBCSP와 비교했을 때 각각 12.34%, 9.00%, 4.95% 더 높은 정확도를 보였다[10].

임명은 외 4명(2017)은 건강검진 데이터를 통해 클러스터 기반으로 예측한 미래건강 예측 데이터를 앙상블 모델의 학습데이터로 사용하여 미래건강을 예측하는 연구를 수행하였고, 앙상블 예측모델 결과와 개별 예측모델 결과 비교를 통해 앙상블 모델이 더 높은 성능을 나타냄을 확인하였다[11].

본 연구에서 사용된 모델의 하이퍼 파라미터는 다음 표 3과 같다.

표 3. 사용된 하이퍼 파라미터 목록
Table 3. List of used hyper parameters

Abbreviation	Method name	Additional attributes and model parameter
GLM	Generalized linear models	Family : Gaussian Regularization : Elastic net Alpha : 0.5 Lamda : 5.8953
GBM	Gradient boosting machine	Number_of_trees : 50 Min_depth : 5 Max_depth : 5 Min_leaves : 8 Max_leaves : 30
RF	Random forest	Number_of_trees : 50 Min_depth : 19 Max_depth : 20 Min_leaves : 637 Max_leaves : 761
DL	Deep learning	Activiation function : ReLU Number of hidden layers : 4 Learning rate : 0.01 Epochs : 1000 Max Iteration : 1000 Alpha : le-05

본 연구에서 제안한 휘발유 출하량 예측 모델에서 사용한 앙상블 기법은 스태킹(Stacking) 기법이다. 스태킹 앙상블 학습은 3단계 절차로 진행된다. 첫 번째 각 base algorithms을 학습데이터로 훈련시키고, 두 번째 각 모델의 예측값을 학습데이터로 만들어 meta learner를 훈련시킨다. 최종적으로 예측은 첫 번째에서 나온 예측값을 훈련데이터로 사용하고, 예측모델은 두 번째에서 학습한 meta learner를 통해 예측한다[12].

본 연구에서는 glm, GBM, Random Forest, Deep Learing을 조합하여 3가지 base algorithms을 설정하였고, meta learner로는 GBM을 사용하였다.

Ⅳ. 실험 결과

본 연구는 일반적으로 성능이 좋은 4가지 학습방법과 이를 앙상블하여 저유소 휘발유 출하량을 예측하는 모델을 구축한다.

다음 표 4는 검증데이터를 통해 실제 정유회사 A의 B 지역 저유소 실제 휘발유 출하량과 예측값의 MAPE를 계산한 결과다. GBM과 Random Forest를 활용한 앙상블 모델이 MAPE 29.39%로 7가지 학습방법 중 가장 좋은 성능을 보인다. 다음으로 Random Forest, GBM, Ensemble1, Ensemble2, GLM, Deep Learning 순으로 높은 성능을 보인다.

하지만 Ensemble1과 Ensemble2의 경우, GBM과 RF를 각각 사용했을 때보다 낮은 성능을 보인다. 이는 DL의 성능이 가장 낮기 때문에, 성능에 큰 영향을 미친 것으로 보인다.

표 4. 휘발유 출하량 예측 모델 성능 비교
Table 4. Comparison of performance of prediction model for gasoline shipment

No	Model	MAPE
1	GLM	35.37%
2	GBM	29.98%
3	RF	29.55%
4	DL	44.98%
5	Ensemble1 (GLM, GBM, RFt, DL)	31.95%
6	Ensemble2 (GBM, RF, DL)	31.95%
7	Ensemble3 (GBM, RF)	29.39%

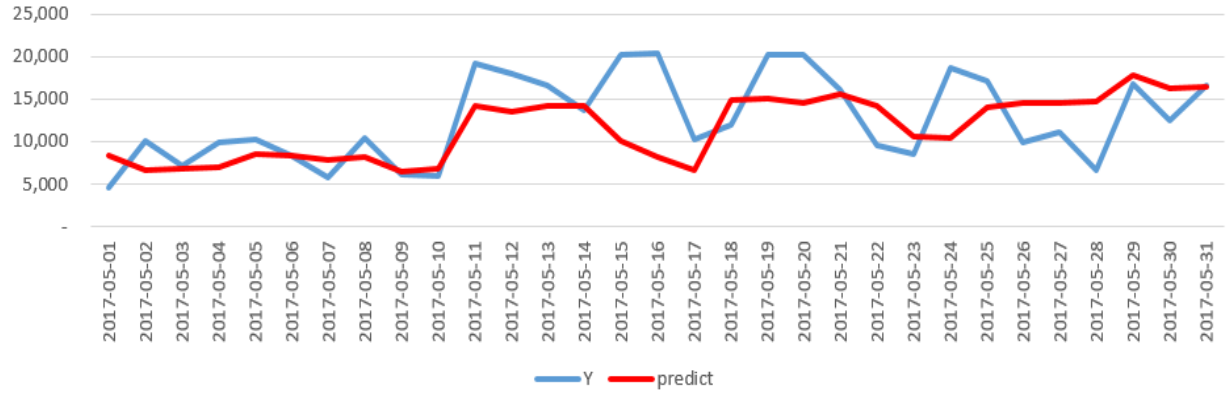

그림 5. 실제 휘발유 출하량과 앙상블 모델을 활용한 예측 출하량 비교 그래프
Fig. 5. Comparison graph of real gasoline shipment and predicting gasoline shipment using ensemble model

Ⅴ. 결론 및 향후 과제

그림 5는 제안된 예측 모델 중 가장 좋은 성능을 보인 Ensemble3의 예측값과 2019년 5월 실제 휘발유 출하량을 비교한 그래프다. 제안된 Ensemble3 모델은 MAPE 29.39% 수준으로 전문가들의 예측수준 약 45.0% 보다 약 15.61% 정도 향상된 수준이며, 다른 알고리즘을 조합하는 것 보다 실험데이터에 대한 성능이 뛰어난 알고리즘을 선택하여 조합하는 것이 예측성능 향상에 도움이 되는 것을 확인할 수 있었다.

본 연구에서는 저유소 출하량 예측에 머신러닝 기법을 활용하는 것이 경험을 토대로 전문가들이 출하량을 예측하는 것 보다 높은 예측성능을 보임을 확인할 수 있었다. 약 15.61% 정도의 정확도 향상은 전체적인 재고관리, 인관 계획 등 출하량 계획을 고려하였을 때, 정유소에 상당한 비용절감 효과를 보일 것으로 판단된다. 향상된 예측 성능은 재고과부족 문제를 줄임에 따라 회사의 경제적 손실이 줄 것으로 보인다.

또한, 본 연구는 기계학습을 적용하여 지역 저유소 유류제품 출하량을 예측하는 최초 연구다.

향후 연구로는 휘발유 뿐만 아니라 등유, 경유 출하량 예측과 함께 출하량에 영향을 미치는 사회적인 이슈 등을 예측지표로 반영한다면 더 높은 예측 성능을 보일 것이라고 판단된다. 또한 시계열 예측에 높은 성능을 보이는 Bi-LSTM(Bidirectional Long Short Term Memory), GRU(Gate Recurrent Unit) 등을 활용하여 예측하는 것도 높은 성능을 보일 것으로 고려된다.

References

[1] Soon-Ill Yang, "An Inventory Operation Model for Petroleum Terminal", Ajou University Graduate School of Engineering, Feb. 2007.

[2] In Soo Lee, "Heuristics for Estimating Desirable Inventory Levels of Refinery Products", Korean Management Review, Vol. 19, No. 2, pp. 1-14, Feb. 1990.

[3] Haiying Jia, Roar Adland, and Yuchen Wang, "Latin American Oil Export Destination Choice: A Machine Learning Approach", 2019 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Macao, Macao, pp. 345-348, Dec. 2019.

[4] Dunteman, George H., and Moon-Ho R. Ho, "An introduction to generalized linear models", Sage Publications, Vol. 145, 2005.

[5] Lu, Haihao and Rahul Mazumder, "Randomized gradient boosting machine", SIAM Journal on Optimization, Vol. 30, No. 4, pp. 2780-2808, 2020.

[6] Liaw, Andy, and Matthew Wiener, "Classification and regression by randomForest", R news Vol. 2, No. 3, pp. 18-22, 2002.

[7] Lauzon, F. Q., “An introduction to deep learning”,

In 2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA), IEEE, pp. 1438-1439, Jun. 2012.

[8] Hong-Mo Je and Sung-Yang Bang, "Improving SVM Classification by Constructing Ensemble", Journal of KISS : Software and Applications, Vol 30, No. 34, pp. 251-258, Apr. 2003.

[9] Yun-Hee Kang, Sang-Hwan Kung, Kyung-Woo Kang, Baek-Min Kim, Haeng-jin Jang, and Sung-yun Yu, "Design of Datafarm System for Suporting an Ensemble based Climate Simulation", The Journal of Korean Institute of Information Technology, Vol. 11, No. 1, pp. 159-167, Jan. 2013.

[10] Sang-Hoon Park, David Lee, and Sang-Goog Lee, "A Study on Regularized Common Spatial Pattern Ensemble Method Based on Filter Bank for Small Sample Setting Motor Imagery EEG", The Journal of Korean Institute of Information Technology, Vol. 15, No. 4, pp. 73-80, Apr. 2017.

[11] Myungeun Lim, Jae-Hun Choi, Ho-Youl Jung, Youngwoon Han, and Hwin-Dol Park, "An ensemble prediction of health records using a deep learningmethod", Proceedings of KIIT Conference, pp. 197-198, Dec. 2017.

[12] Eric C Polley and Mark J van der Laan, "Super Learner In Prediction", U.C. Berkeley Division of Biostatistics Working Paper Series, Paper 266, pp. 1-15, May 2010.

저자소개

안 성 준 (Sung-Jun Ahn)

2014년 : 한국산업기술대학교 컴퓨터공학과(공학사)
2014년 ~ 현재 : 고려대학교 산업경영공학부 석사과정
2019년 ~ 현재 : (주)에이아이더 선임연구원
관심분야 : Data Mining, Management of Technology, Business Analysis

장 동 식 (Dong-Sik Jang)

1979년 : 고려대학교 산업공학과(공학사)
1985년 : 텍사스주립대학교 산업공학과(공학석사)
1988년 : 텍사스 A&M 산업공학과(공학박사)
1989년 ~ 현재 : 고려대학교 산업경영공학부 교수
관심분야 : Project Management, Pattern Recognition, Data mining

박 상 성 (Sang-Sung Park)

2006년 : 고려대학교 산업시스템정보공학부(공학박사)
2006년~2014년 : 고려대학교 산업경영공학부 연구교수
2015년 ~ 2018년 : 고려대학교 기술경영 전문대학원 조교수
2019년 ~ 현재 : 청주대학교 빅데이터통계학과 조교수
관심분야 : Patent analysis, Data mining, Management