ISSN 1229-1889(Print)
ISSN 2287-9005(Online)
J Korean Soc Qual Manag Vol. 49, No.2: 201-211, June 2021
https://dx.doi.org/10.7469/JKSQM.2021.49.2.201

대용량 자료에 대한 밀도 적응 격자 기반의 k-NN 회귀 모형

유의기* · 정욱**†

* 동국대학교 경영학부 박사과정생
** 동국대학교 경영학부 교수

Density Adaptive Grid-based k-Nearest Neighbor Regression Model for Large Dataset

Yiqi Liu* · Jung Uk**†

* Ph.D. Student, College of Business Administration, Dongguk University
** Professor, College of Business Administration, Dongguk University

ABSTRACT

Purpose: This paper proposes a density adaptive grid algorithm for the k-NN regression model to reduce the computation time for large datasets without significant prediction accuracy loss.

Methods: The proposed method utilizes the concept of the grid with centroid to reduce the number of reference data points so that the required computation time is much reduced. Since the grid generation process in this paper is based on quantiles of original variables, the proposed method can fully reflect the density information of the original reference data set.

Results: Using five real-life datasets, the proposed k-NN regression model is compared with the original k-NN regression model. The results show that the proposed density adaptive grid-based k-NN regression model is superior to the original k-NN regression in terms of data reduction ratio and time efficiency ratio, and provides a similar prediction error if the appropriate number of grids is selected.

Conclusion: The proposed density adaptive grid algorithm for the k-NN regression model is a simple and effective model which can help avoid a large loss of prediction accuracy with faster execution speed and fewer memory requirements during the testing phase.

Key Words: Regression, k-nearest Neighbor, Grid, Density, Computation Time

● Received 8 June 2021, 1st revised 11 June 2021, accepted 14 June 2021
† Corresponding Author(ukjung@dongguk.edu)
ⓒ 2021, Korean Society for Quality Management
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0) which permits unrestricted non-Commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.※ This work was supported by the Dongguk University Research Fund of 2021
※ This work was supported by the Dongguk University Research Fund of 2021.

1. 서 론

현대 제조산업에서 발생하는 데이터는 그 크기가 기하급수적으로 커지고 있으며, 이를 빠르고 정확하게 분석하여 우수한 예측 성능을 보이면서도 예측 과정에서 요구되는 시간을 줄일 수 있는 방법론에 대한 관심이 높아지고 있다. 예를 들어 반도체 제조 공정에 투입되는 웨이퍼(wafer)의 정확한 품질관리를 위해서는 정확하고 빠른 예측의 성능을 보이는 가상 계측(virtual metrology, VM) 모델의 개발이 요구된다. 이는 기존의 계측(metrology)과 통계적 공정 관리(statistical process control; 이하 SPC) 기법에 기반한 품질 관리 방식은 별도의 계측 공정을 추가함으로써 전체 공정의 시간이 증가하고(Chen et al., 2005; Su et al., 2007), 샘플링 기법을 사용하는 것으로 인해 모든 웨이퍼에 대한 개별적 품질 관리가 불가능하게 된다는 단점(Chen et al., 2006; Lin et al., 2006)이 있다는 것에 기인한다. 이러한 계측 기반 SPC의 단점을 해결하기 위하여 대두된 가상 계측의 개념은 실제로 계측 공정을 수행하지 않고서도 모든 웨이퍼에 대한 계측 결과를 짐작하는 예측 모델(prediction model)을 개발하는 것을 핵심으로 하고 있다. 이때 가상 계측 모델은 추출된 웨이퍼들의 공정 설비 데이터를 설명 변수(explanatory variables)로 하고 해당 웨이퍼들을 실제로 계측함으로써 얻는 품질 지표들을 목표 변수(target variables)로 하는 예측 모델을 구축하게 된다(Khan et al., 2007). 즉, 연속된 제조 공정에서 가상 계측 모델을 통해 실제 계측 공정을 수행하지 않더라도 모든 개별 웨이퍼에 대한 계측 정보들을 빠르게 얻기 위해서는 새로운 웨이퍼에 대한 높은 정확도와 빠른 예측 모델의 구현이 요구된다. 특히 예측 정확도가 수용 가능한 수준에 한해서는 그 예측의 과정을 실시간으로 수행할 필요가 있다.

예측 모델 구현을 위한 여러 기계 학습(machine learning) 방법들 중에서 k-최근접 이웃(k-nearest neighbor, k-NN) 알고리즘은 이미 여러 분야에서 적용되고 있으며 그 성능을 검증 받은 기계 학습 방법 중의 하나이다. k-NN 알고리즘은 이미 알려진 훈련 셋(training set)을 메모리에 기억한 다음 그 중 가장 유사한 k개의 관측치를 선택하여 선택된 관측치의 레이블(Label)에 따라 새로운 대상 관측치의 레이블을 예측하는 방식의 알고리즘이다. k-NN 알고리즘은 이산 클래스 레이블을 예측하기 위해 분류 문제에 집중적으로 적용되었으며 예측될 레이블이 연속형 속성에 해당하는 회귀 문제에서도 많이 이용되고 있다. 예를 들어, Yang and Zhao(2006)은 k-NN 회귀의 몇 가지 일반화된 알고리즘을 개발하여 얼굴 인식 문제에 적용하였다. Eronen and Klapuri(2009)는 k-NN 회귀를 음악에서의 템포 추정을 위한 접근법으로 활용하였다. 또한, k-NN 회귀 분석은 전기 가격 예측(Lora, et al., 2007), 전기 부하 예측(Lora et al., 2003; Bhanu et al., 2008; AI-Qahtani et al., 2013), 환율 예측(Fernandez et al., 1999), 기후 예측(Dimri et al., 2008), 수자원(hydrological) 시계열 예측(Jayawardena et al., 2002; She and Yang, 2010) 등 여러 응용 분야에서 다양하게 사용되었다. 그러나 최근 대용량의 데이터가 보급되면서 이러한 k-NN 회귀 모형은 대용량의 데이터로 구성된 훈련 셋의 처리를 위한 데이터의 기억 장치 용량, 데이터간의 유사도(similarity) 및 데이터 정렬(sorting)을 위한 계산량 등이 급격히 증가하는 문제점을 직면하게 된다(Hastie et al., 2009).

본 연구에서는 기존 k-NN 회귀 모형의 계산량을 감소시키는 반면, 기존의 k-NN 회귀 모형과 유사한 성능을 보이는 간단하면서도 효과적인 새로운 알고리즘인 “밀도 적응 격자 알고리즘(density adaptive grid-based algorithm (DAG) algorithm)”을 제안하고자 한다. 먼저 원형의 데이터 공간에서 각 변수(feature)별로 적정한 수의 격자(grid)를 구성하고 각 격자 안에 존재하는 관측치들을 이용하여 해당 격자를 대표하는 중심점(centroid)를 계산한다. 그리고 k-NN 회귀 모형의 학습 시에 원형의 관측치가 아닌, 새롭게 계산된 중심점들을 이용하여 회귀분석을 수행한다. 특히, 각 변수 별 격자를 구성할 때 분위수(quantile)의 정보를 이용하여 밀집된(dense) 공간에서는 많은 수의 격자를, 반대로 관측치가 드문(sparse) 공간에서는 적은 수의 격자를 구성함으로써 원형의 데이터가 보유한 밀도(density)의 정보를 반영하게 된다. 이를 통해 k-NN 알고리즘에서 최근접 k개의 이웃을 찾는데 소요되는 계산량을 크게 감소시키면서도 원형의 데이터가 지닌 밀도의 특성을 최대한 반영하는 k-NN 회귀 모형을 구현하게 된다. 즉,

계산 시간과 회귀 모형의 정확도 간의 상쇄관계(trade-off)를 활용하여 무시할 수 있는 수준의 정확도 손실을 일부 허용하되 계산 시간의 큰 감소를 기대하는 방법이라고 할 수 있다.

결국 이러한 본 연구에서 제안하는 밀도 적응 격자(DAG) 알고리즘의 장점은 다음과 같다.

- 첫째, 이 알고리즘의 실행 과정이 간단하고 응용 분야의 제약이 크지 않다.
- 둘째, 이 알고리즘을 통하여 얻게 되는 중심점으로 구성된 새로운 데이터 셋의 크기는 원형의 훈련 데이터 셋에 비해 크게 축소되므로 k-NN 회귀 모형에 적용 시 메모리의 감소와 빠른 계산 시간 등의 장점을 제공할 수 있다.
- 셋째, 이 알고리즘은 원래 데이터의 밀도 정보를 보유할 수 있기 때문에 데이터의 크기를 줄임으로 인해 야기될 수 있는 성능의 감소라는 부정적인 영향을 크게 받지 않는다.

본 논문에서는 다섯 개의 실제 데이터 셋을 통해 기존 k-NN 회귀 모형과 본 연구에서 제안하는 밀도 적응 격자 기반의 k-NN(DAG k-NN) 회귀 모형을 구축한 후, 모형의 예측력, 데이터 축소 비율 및 예측 시간의 효율성을 비교하였다. 본 논문의 실험에서 제안하는 DAG k-NN 회귀 모형은 기존 k-NN 회귀 모형에 비해 예측 시간을 많이 축소하지만 예측력에는 큰 차이가 없다는 결과를 얻었다.

본 논문의 구성은 다음과 같다. 제 2장에서는 관련된 이론적 배경에 대해 설명한다. 제 3장에서는 본 연구에서 제안하는 DAG k-NN 회귀 모형을 상세히 설명한다. 제 4장에서는 실험 과정 및 평가 방법에 대해 설명하고, 제 5장에서는 실험 결과를 토의한다. 끝으로 제 6장에서는 결론과 향후 연구방향을 제시한다.

2. 이론적 배경 및 선행 연구

k-NN은 1968년 Cover에 의해 제안된 알고리즘으로 가장 간단한 기계학습 알고리즘이다. k-NN의 기본적인 아이디어는 유사한 값들은 서로 가까이에 위치한다는 것이다. 따라서 유사도는 두 데이터 간의 거리(distance)로 측정된다. k-NN 알고리즘은 훈련 데이터 셋이 주어진 상황에서 새로운 관측치에서 가장 가까이에 위치한 k개의 데이터의 정보를 이용하여 예측 및 분류 문제를 해결한다. 데이터 속성이 수치인 경우 데이터 간에 거리를 측정하는 방법에 활용되는 거리는 유클리드 거리(Euclidean distance), 맨해튼 거리(Manhattan distance), 민코우스키 거리(Minkowski distance) 등이 있다. 그 중 대표적인 유클리드 거리는 N차원 공간의 두 점 $U(u_1, u_2, \ldots, u_N)$, $V(v_1, v_2, \ldots, v_N)$이 주어질 때 식(1)과 같다.

$$D(U, V) = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + \cdots + (u_N - v_N)^2} \tag{1}$$

k-NN 회귀 모형은 주어진 새 관측치와 훈련 셋(기존 관측치)의 입력 변수(좌표 값)를 비교하고 거리를 계산하여 가장 근접한 k개의 기존 관측치를 찾은 후, 그들에 대한 출력 변수 y_i의 값을 평균하여 새 관측치의 출력 변수 y의 예측 값으로 산출한다. 이에 대한 수식은 아래 식(2)와 같다.

$$y = \frac{1}{k} \sum_{i=1}^{k} y_i \tag{2}$$

여기서 y_i는 새 관측치에 가장 가까이에 위치한 i번째 기존 관측치의 결과값이다. k는 최근접 이웃의 개수가 주요한 초매개변수(Hyper-parameter)로 너무 크게 설정하거나 너무 작게 설정할 경우 과소적합(under-fitting)이나 과

대적합(over-fitting)을 야기할 수 있어 적절한 값의 설정이 필요하다. 적절한 k값을 선정하는 문제는 주로 교차검증(cross-validation)을 활용하는 것으로서 아래 Figure 1과 같이 여러가지 k값을 적용하여 모형의 RMSE(root mean squared error) 값을 비교 분석한 후 오차가 가장 작은 k를 선정하게 된다.

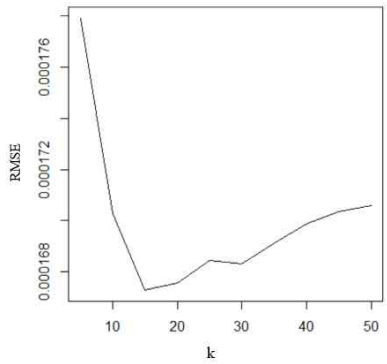

Figure 1. Example of selection of optimal k-value (California dataset)

앞서 언급한 k-NN 회귀 모형의 계산 시간 복잡도 문제점을 개선하기 위해 기존 연구에서는 인스턴스 선택(Instance Selection)이라는 개념의 다양한 알고리즘이 제시되었다(Song et al., 2017). 기존 연구에서 사용하는 인스턴스(Instance)라는 용어는 훈련 데이터 셋의 관측치(observation)를 의미한다. 예를 들면, Guillén et al.(2010)은 시계열 예측에서 상호 정보량(mutual information)을 이용한 새로운 인스턴스 선택 알고리즘을 제안하였다. 이 방법은 인위적으로 생성된 데이터에서는 좋은 성능을 보여주지만 현실의 데이터 셋에서는 그 성능이 검증되지 못하였다(Song et al., 2017). Rodríguez et al.(2013)은 회귀를 위한 클래스 조건부 인스턴스 선택(CCISR) 기법을 제시하였다. 이 알고리즘은 상대적으로 의미 있는 데이터 포인트를 보유하면서 높은 데이터 감소 비율을 보여주지만 계산 시간과 메모리 요구량이 높기 때문에 일반적으로 빠른 결과를 획득해야 하는 실제 상황에서는 그리 활용도가 높지 않다(Arnaiz et al., 2016). 또한, Arnaiz et al.(2016)은 앙상블(Ensemble) 아이디어에 기반한 회귀 분석에서의 인스턴스 선택 방법을 제안하였다. 기존의 인스턴스 선택 알고리즘과 비교하여 볼 때 앙상블 알고리즘은 예측 오류 및 축소된 부분 집합 크기 면에서 최고의 성능을 보였다. 하지만 이러한 기존 방법들은 주로 회귀 모델에 미치는 영향이 적은 관측치를 제거하기 보다는 잡음이 많이 포함된 관측치를 제거하는 것에 중점을 두고 있다. 즉, 잡음이 적은 데이터 셋의 경우에 이러한 방법들이 적용될 시에는 그 성능의 개선을 예측하기 어렵다.

다음 장에서는 기존 k-NN 회귀 모형의 계산 시간을 감소시키는 반면, 수용 가능한 예측 정확도 성능을 보이는 간단하면서도 효과적인 새로운 알고리즘을 제안하고자 한다.

3. 제안 방법론: DAG k-NN 회귀 모형

본 장에서는 제안하는 밀도 적응 격자(DAG) 알고리즘을 설명한 후, 이 알고리즘을 이용하는 밀도 적응 격자 기반의 k-NN(DAG k-NN) 회귀 모형을 상세히 기술한다.

3.1 밀도 적응 격자 알고리즘(Density adaptive grid algorithm)

밀도 적응 격자 알고리즘은 다음과 같은 단계를 거친다. Fig. 2는 이 알고리즘의 수행 과정에 대한 예를 보여준다.

- <Step 1> 데이터 셋 $X \in R^d$ 주어졌을 때 각 차원 $x_i,\ i=\{1,2,3,\cdots,d\}$를 분위수를 이용하여 특정한 개수(number of intervals, NI)의 구간으로 나누고 총 NI^d 개 격자 $g_j,\ j=\{1,2,3,\cdots,NI^d\}$를 생성한다. 관측치 $X_i,\ i=\{1,2,3,\cdots,n\}$ 각각은 스스로가 포함된 격자의 인덱스(grid index), $j=\{1,2,3,\cdots,NI^d\}$가 있고 관측치가 포함되지 않는 격자는 고려되지 않는다. Fig. 2(a), (b)는 2차원 공간에 관측치 1,000개가 있고 NI값이 17로 선정 되었을 때 17^2개 격자를 생성하는 그림이다.
- <Step 2> 생성된 각 격자(단, 관측치를 한 개 이상 포함하는)의 중심점 c_j를 찾는다. 이렇게 생성된 격자별 중심점 c_j의 집합 $c=\{c_j\}=\{c_{(1)},\cdots,c_{(n')}\}$를 형성한다. 이때 n'는 관측치를 한 개 이상 포함하는 격자들의 개수이다(Fig. 2(c)).
- <Step 3> 중심점으로 구성된 새 데이터 셋 C를 새로운 훈련 데이터 셋으로 사용한다(Fig. 2(d)).

이와 같이 분위수를 이용해 격자를 생성하고 중심점을 찾는 알고리즘은 원형의 데이터 셋이 지닌 밀도의 정보를 그대로반영하면서 데이터 크기를 축소하게 된다.

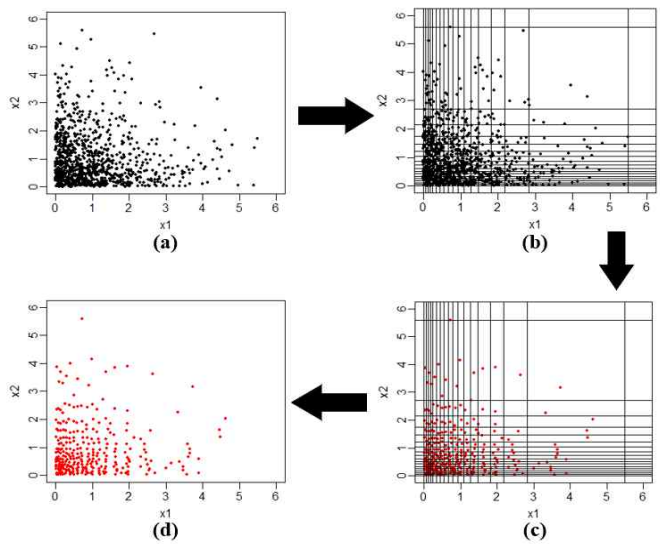

Figure 2. Example of the proposed density adaptive grid algorithm

3.2 밀도 적응 격자 기반의 k-NN 회귀 (DAG k-NN regression)

DAG k-NN 회귀 모형을 설계할 시에는 초매개변수 k와 NI의 설정에 따라 성능이 차이가 나타날 수 있어 적절한 값의 설정이 필요하다. 이러한 문제를 해결하는 방안으로 교차검증을 이용하여 초매개변수를 적절히 선택할 수 있다. 이 논문에서는 10-겹 교차검증을 사용하였다. Fig. 3은 제안하는 DAG 알고리즘을 이용하는 DAG k-NN 회귀 모형의 구조를 나타내며, 그 내용은 다음과 같다.

① 먼저 데이터를 수집하고 전처리(pre-processing)를 수행한다.
② 다양한 NI값의 후보들을 선정하고 각 NI 값에 대해 DAG 알고리즘을 이용하여 중심점으로 구성된 새 데이터 셋 C를 생성한다.

③ 다양한 초매개변수 k의 후보들을 선정하고 특정한 k의 값에 대한 k-NN 회귀분석을 수행한다. 주어진 새 관측치의 입력 변수와 훈련 셋 C를 비교하고 거리를 계산하여 가장 근접한 k개의 중심점 $c_{(i)}$를 찾은 후, 그들에 대한 출력 변수 z_i의 값을 평균하여 새 관측치의 출력 변수 y의 예측 값으로 사용한다. 이에 대한 수식은 아래 식(3)와 같다.

$$y = \frac{1}{k} \sum_{i=1}^{k} z_i \tag{3}$$

여기서 z_i는 새 관측치에 가장 가까이에 위치한 i번째 중심점 $c_{(i)}$의 출력 변수이다. 이는 해당 중심점을 가지는 격자 내의 원형 관측치들의 출력 변수들을 평균한 값이다. 이렇게 예측된 y를 기반으로 평균 제곱근 오차(root mean squared error, RMSE)를 이용하여 회귀모형의 성능을 측정하는 방식으로 교차검증을 사용하여 최적 k값을 선정한다.

④ 각 NI 값에 따라 계산된 예측 정확도와 계산 시간의 효율을 참고하여 최적의 초매개변수 NI 값을 선정하여 최적 DAG k-NN 회귀 모형을 구축한다.

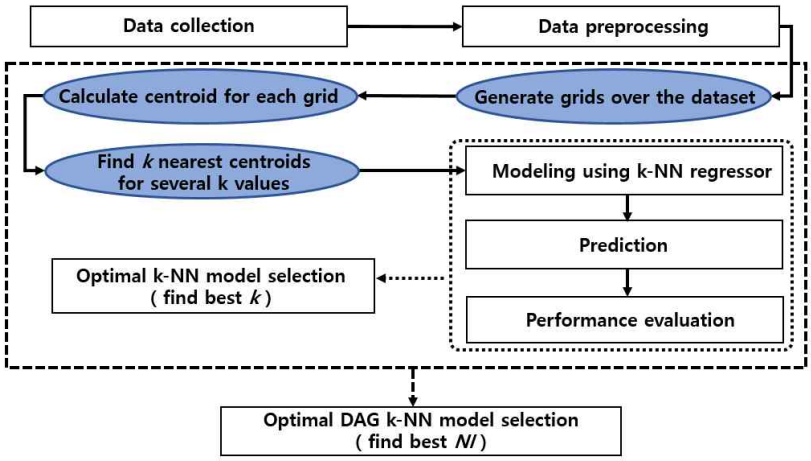

Figure 3. Flow chart of density adaptive grid-based k-NN regression model

4. 실험의 설계 및 평가

4.1 실제 데이터 셋의 활용

본 연구에서 제안하는 DAG k-NN 회귀 모형을 평가하기 위해 다양한 다섯 개의 실제 데이터 셋에 제안하는 모형과 기존 모형을 적용하여 비교 평가하였다. 사용한 데이터 셋은 KEEL Data set repository(http://www.keel.es/data-set.php) (Alcalá-Fdez et al., 2011)에 공개된 데이터 셋 다섯 개를 선정하였다. 해당 데이터 셋의 특성은 Table 1에 요약되어 있다. 선택된 데이터 셋이 훈련과 테스트 셋으로 구성되지 않은 경우 주어진 데이터 셋의 20%를 시험 데이터 셋(test data set)으로 추출하였다. 본 논문에서는 최근 R programming language 및 관련 패키지를 이용하

여 각 모형의 성능을 비교하였다.

Table 1. Characteristics of the Real Datasets

NO	Dataset	Number of samples (n)	Number of features (d)
# 1	California	20,640	8
# 2	Delta Ailerons	7,129	5
# 3	Delta Elevators	9,517	6
# 4	House-16H	22,784	16
# 5	Pole Telecommunications	14,998	26

4.2 예측 모형의 성능 평가

본 연구에서 제안된 DAG k-NN 회귀 모형의 성능을 평가하기 위하여 예측 정확도, 데이터의 축소 비율 및 예측 시간 효율성을 기준으로 기존 k-NN 회귀 모형의 성능과 비교하였다. 본 연구에서는 훈련 데이터 셋에 대한 10-fold 교차검증으로 최적 초매개변수 k와 NI를 선정하고 최종 구축된 모형을 따로 마련한 시험 데이터 셋에 적용하여 성능을 평가하였다. 예측 정확도 평가지표는 평균 제곱근 오차(root mean squared error, RMSE)이다. RMSE는 모형이 예측한 값과 실제 값의 차이를 다룰 때 흔히 사용하는 측도이며, 잔차(residual)의 표준편차로서 산출식은 식(3)과 같다. 여기에서는 모형이 잘 훈련 되어있는지 판단을 위해 사용하였으며, RMSE 값은 작을수록 성능이 좋다.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y_i}\right)^2} \tag{3}$$

여기서 N은 시험 데이터 셋의 관측치 수, y_i는 실제 레이블 값, $\hat{y_i}$는 예측값으로 나타낸다.

모형의 예측 시간은 훈련 데이터를 축소하는 방법에 따라서 차이가 많이 나타날 수 있기 때문에 모형의 성능을 정확하게 평가하기 위해 데이터 축소 비율(Data Reduction Ratio)과 시간 감소 비율(Reduced Time Ratio)을 동시에 측정하였다. 데이터 축소 비율과 시간 감소 비율은 식(4), 식(5)와 같이 계산된다. 식(4)에 표기된 데이터 축소 비율의 분자 값은 축소 전과 후의 데이터 셋 크기의 차이를 의미한다. 데이터 축소 비율은 클수록 우수하고, 시간 감소 비율 은 기본 모형의 효율성을 100%로 볼 때 작을수록 우수한 것으로 본다.

$$Data\ \ Reduction\ \ Ratio(\%)\ \ of\ \ \mathrm{model}(p) = \frac{Nmber\ \ of\ \ reduced\ \ data\ \ points\ \ \ of\ \ \mathrm{model}(p)}{Nmber\ \ of\ \ \mathrm{origianl}\ \ data\ \ points} \times 100\% \tag{4}$$

$$Reduced\ \ Time\ \ Ratio(\%)\ \ of\ \ \mathrm{model}(p) = \frac{t_p}{t_o} \times 100\% \tag{5}$$

여기서 t_o는 기존 모형의 예측 시간이며, t_p는 제안된 모형의 예측 시간이다.

5. 실험 결과

Table 2-6 에서는 여러 NI값이 주어질 때 모형의 데이터 축소 비율과 최적 k값을 이용할 때의 시간 감소 비율과

RMSE를 기록하였다. California, Delta Ailerons와 Delta Elevators 데이터 셋에 대해 NI값을 2, 3, 5, 7, 10으로 증가시키며 성능을 측정하였다. Table 2-4를 살펴보면 DAG k-NN 회귀 모형의 데이터 축소 비율과 시간 감소 비율이 NI값이 작아질수록 k-NN 회귀 모형보다 우세하고 RMSE는 다소 증가하기는 하지만 기존의 모형과 큰 차이는 없다. 이는 새롭게 생성된 훈련 데이터 셋(즉, 격자들의 중심점)이 시험 데이터에 대한 충분한 예측 정보를 가지기 때문이다. House-16H와 Pole Telecommunications 데이터 셋은 변수가 상대적으로 많아서 NI값을 크게 선정할 경우 중심점으로 구성된 새로운 데이터 셋의 크기가 크게 줄어들지 않음으로 NI값을 2, 3, 5만으로 선택하여 성능을 측정하였다. 두 데이터 셋의 결과는 Table 5, 6과 같다. NI값이 낮을수록 데이터 축소 비율과 시간 감소 비율이 비교적 우수하며 RMSE의 차이도 크지 않은 것을 확인할 수 있다.

Table 2. Results of Model Performance (California Dataset)

DAG k-NN regression model						Original k-NN regression
NI	2	3	5	7	10	
Data reduction ratio (%)	98.87	92.01	60.38	33.16	16.59	0.00
Reduced Time ratio (%)	13.8	19.36	46.95	69.71	86.01	100
Optimal k	4	5	10	10	10	10
RMSE	83,068.04	72,298.91	65,347.89	63,197.33	62,754.65	62,067.52

Table 3. Results of Model Performance (Delta Ailerons Dataset)

DAG k-NN regression model						Original k-NN regression
NI	2	3	5	7	10	
Data reduction ratio (%)	99.44	96.39	71.74	40.46	18.62	0
Reduced Time ratio (%)	20.77	35.5	35.93	61.57	83.38	100
Optimal k	2	5	10	20	15	20
RMSE	0.000192	0.000177	0.000168	0.000166	0.000167	0.000165

Table 4. Results of Model Performance (Delta Elevators Dataset)

DAG k-NN regression model						Original k-NN regression
NI	2	3	5	7	10	
Data reduction ratio (%)	99.16	94.51	69.77	36.63	19.29	0
Reduced Time ratio (%)	20.62	23.53	41.66	65.47	80.88	100
Optimal k	3	10	20	25	40	50
RMSE	0.001561	0.001471	0.001463	0.001434	0.001435	0.001430

Table 5. Results of Model Performance (House-16H Dataset)

DAG k-NN regression model				Original k-NN regression
NI	2	3	5	
Data reduction ratio (%)	67.44	20.56	10.51	0
Reduced Time ratio (%)	41.13	80.42	88.85	100
Optimal k	10	10	10	10
RMSE	41,038.73	38,117.81	37,837.50	38,070.58

Table 6. Results of Model Performance (Pole Telecommunications Dataset)

DAG k-NN regression model				Original k-NN regression
NI	2	3	5	
Data reduction ratio (%)	66.07	47.00	30.51	0
Reduced Time ratio (%)	44.86	61.8	72.76	100
Optimal k	3	5	3	5
RMSE	14.88	11.82	10.63	9.69

Fig. 4는 RMSE와 예측 시간 효율성(time efficiency ratio) 결과를 함께 표현한 그림이다. 예측 시간 효율성은 $\frac{t_o - t_p}{t_o} * 100\%$를 구한 결과로서 그 값이 클수록 시간 감소의 효과가 크다는 것을 의미한다. 앞에서 살펴본 Table의 결과와 같이 DAG 알고리즘의 훈련 데이터의 축소는 기존의 k-NN 회귀 모형의 예측 시간을 크게 줄일 수 있음과 동시에 비슷한 예측력을 유지할 수 있다. 이러한 사실은 군집화의 대표적인 알고리즘인 k-means 방법에서의 팔꿈치(Elbow) 방식과 유사하게 RMSE의 큰 증가가 일어나지 않으면서 예측 시간 효율성이 상대적으로 높은 기준으로 NI 값을 선정할 수 있음을 의미한다. 이는 본 연구에서 제안하는 DAG k-NN 회귀분석을 사용하는 데이터 분석자의 주관적 의지를 반영하는 방법으로서 두 성과지표의 상쇄관계(trade-off)를 고려한 최종 모형의 선정이 가능함을 뜻한다.

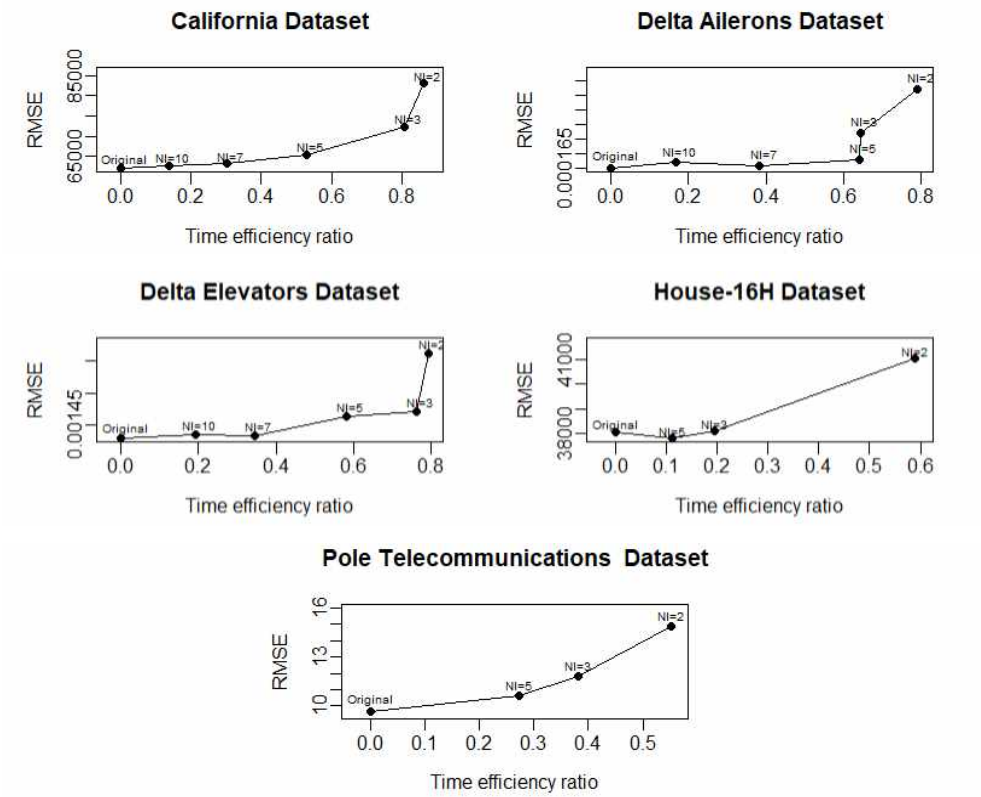

Figure 4. Trade-off between RMSE and time efficiency ratio in five different real-life data sets

5. 결 론

기존의 k-NN 회귀 분석은 단순한 구현에 비해 높은 성능을 보였으나 학습 데이터의 크기가 커지면 큰 규모의 메모리와 높은 계산 시간을 필요로 하는 문제점을 안고 있다. 본 연구에서는 기존의 k-NN 회귀 모형의 문제점을 개선하기 위해 밀도 적응 격자(DAG) 알고리즘을 제안하고 k-NN 회귀 모형과 결합하여 새로운 k-NN 회귀 모형을 제시하였다. 본 연구에서 제안된 방법은 분위수를 이용하여 격자를 생성하고 각 격자의 중심점을 활용함으로써 원형 데이터를 직접 사용하는 경우에 비해 데이터 저장 공간과 예측에 소요되는 시간을 크게 줄일 수 있다. 그럼에도 불구하고 예측 정확도는 수긍할만한 수준의 감소를 초래함으로써 데이터 분석가의 주관적인 판단으로 k-NN 회귀 모델을 선정할 수 있다는 장점이 있다. 본 연구에서는 다양한 실제 데이터 셋에 대한 실험을 통해 이러한 장점들을 확인해 보았다.

그럼에도 불구하고 본 알고리즘이 대용량 고차원 데이터(high dimensional)에 대해서도 그대로 사용 가능할 수 있는지에 대한 향후 연구가 필요하다. k-NN 알고리즘이 사용하는 거리(distance)의 개념이 다차원 공간에서는 차원의 저주(curse of dimensionality)에 의해 희소(sparsity) 문제를 야기하게 된다. 따라서 원형의 데이터 셋의 특징을 반영한 효과적인 다양한 차원 축소 알고리즘과의 결합하여 본 연구에서 제시된 알고리즘을 변형할 필요가 있을 것으로 예상된다. 또한 통계적 품질관리에서 사용되는 다변량 관리도에서도 k-NN 기법이 자주 사용되므로 회귀가 아닌 분류 문제의 일종으로 볼 수 있는 관리도에서의 k-NN 기법 활용 시 본 연구에서 제안된 DAG k-NN 알고리즘을 이용한 효과적인 프레임워크를 연구하는 것은 매우 흥미로운 연구가 될 것으로 보인다.

REFERENCES

Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L. and Herrera, F. 2011. Keel Data-mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. Journal of Multiple-Valued Logic & Soft Computing 17.

Al-Qahtani, F. H., and Crone, S.F. 2013, August. Multivariate k-nearest Neighbour Regression for Time Series Data—A Novel Algorithm for Forecasting UK Electricity Demand. In The 2013 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

Arnaiz-González, Á., Blachnik, M., Kordos, M., and García-Osorio, C. 2016. Fusion of Instance Selection Methods in Regression Tasks. Information Fusion 30:69-79.

Bhanu, C. V. K., Sudheer, G., Radhakrishna, C., and Phanikanth, V. 2008, October. Day-Ahead Electricity Price Forecasting Using Wavelets and Weighted Nearest Neighborhood. In 2008 Joint International Conference on Power System Technology and IEEE Power India Conference (pp. 1-4). IEEE.

Chen, P., Wu, S., Lin, J., Ko, F., Lo, H., and Wang, J. 2005, Virtual Metrology: A Solution for Wafer to Wafer Advanced Process Control, Proc. IEEE Int. Symp. on Semiconductor Manufacturing (ISSM 2005), 155-157.

Chen, Y.-T., Yang, H.-C., and Cheng, F.-T. 2006, Multivariate Simulation Assessment for Virtual Metrology, Proc. IEEE Int. Conf. on Robotics and Automation(ICRA 2006), 1048-1053.

Dimri, A.P., Joshi, P., and Ganju, A. 2008. Precipitation Forecast over Western Himalayas Using k-nearest Neighbour Method. International Journal of Climatology: A Journal of the Royal Meteorological Society 28(14): 1921-1931.

Eronen, A. J., and Klapuri, A. P. 2009. Music Tempo Estimation With k-NN Regression. IEEE Transactions on Audio,

Speech, and Language Processing 18(1):50-57.

Fernandez-Rodrıguez, F., Sosvilla-Rivero, S., and Andrada-Felix, J. 1999. Exchange-Rate Forecasts with Simultaneous Nearest-Neighbour Methods: Evidence from the EMS. International Journal of Forecasting 15(4):383-392.

Guillén, A., Herrera, L. J., Rubio, G., Pomares, H., Lendasse, A., and Rojas, I. 2010. New Method for Instance or Prototype Selection Using Mutual Information in Time Series Prediction. Neurocomputing 73(10-12): 2030-2038.

Hastie, T., Tibshirani, R., and Friedman, J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.

Jayawardena, A. W., Li, W. K., and Xu, P. 2002. Neighbourhood Selection for Local Modelling and Prediction of Hydrological Time Series. Journal of Hydrology 258(1-4):40-57.

Khan, A. A., Moyne, J. R., and Tilbury, D. M. 2007, An Approach for Factory-wide Control Utilizing Virtual Metrology, IEEE Transactions on Semiconductor Manufacturing 20(4):364-375.

Lin, T.-H., Hung, M.-T., Lin, R.-C., and Cheng, F.-T. 2006, A Virtual Metrology Scheme for Predicting CVD Thickness in Semiconductor Manufacturing, Proc. IEEE Int. Conf. on Robotics and Automation (ICRA 2006), 1054-1059.

Lora, A. T., Riquelme, J. C., Ramos, J. L. M., Santos, J. M. R., and Expósito, A. G. 2003, December. Influence of kNN-Based Load Forecasting Errors on Optimal Energy Production. In Portuguese Conference on Artificial Intelligence (pp. 189-203). Springer, Berlin, Heidelberg.

Lora, A. T., Santos, J. M. R., Expósito, A. G., Ramos, J. L. M., and Santos, J. C. R. 2007. Electricity Market Price Forecasting Based on Weighted Nearest Neighbors Techniques. IEEE Transactions on Power Systems 22(3): 1294-1301.

Rodríguez-Fdez, I., Mucientes, M., and Bugarín, A. 2013, July. An Instance Selection Algorithm for Regression and its Application in Variance Reduction. In 2013 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (pp. 1-8). IEEE.

She, D., and Yang, X. 2010. A New Adaptive Local Linear Prediction Method and its Application in Hydrological Time Series. Mathematical Problems in Engineering, 2010.

Song, Y., Liang, J., Lu, J., and Zhao, X. 2017. An Efficient Instance Selection Algorithm for k Nearest Neighbor Regression. Neurocomputing 251:26-34.

Su, A.-J., Jeng, J.-C., Huang, H.-P., Yu, C.-C., Hung, S.-Y., and Chao, C.-K. 2007, Control Relevant Issues in Semiconductor Manufacturing: Overview with Some New Results, Control Engineering Practice 15(10): 1268-1279.

Yang, S. 2006, August. Regression Nearest Neighbor in Face Recognition. In 18th International Conference on Pattern Recognition (ICPR'06) (Vol. 3, pp. 515-518). IEEE.

저자소개

유의기 동국대학교 경영학과에서 석사학위를 취득하였으며 현재 동국대학교 경영학과에서 박사과정에 재학 중이다. 관심분야는 품질경영, 통계적 품질관리 등이다.

정 욱 현재 동국대학교-서울캠퍼스 경영대학 정교수로 재직 중이다. Georgia Institute of Technology에서 산업시스템공학 박사 학위를 취득하고 삼성SDS SCM사업단에서 SCM 컨설턴트로 재직하였다. 현재 주요 연구관심분야는 품질경영, 기계학습, 공급사슬관리 등이다.