Journal of the Korean Society of Safety
Vol. 36, No. 1, pp. 1-8, February 2021
https://doi.org/10.14346/JKOSOS.2021.36.1.1

ISSN 1738-3803 (Print)
ISSN 2383-9953 (Online)
http://www.kosos.or.kr/jkosos

Group Contribution Method 및 Support Vector Regression 기반 모델을 이용한 방향족 화합물 물성치 예측에 관한 연구

강하영* · 오창보** · 원용선*** · 유준*** · 이창준**†

Group Contribution Method and Support Vector Regression based Model for Predicting Physical Properties of Aromatic Compounds

Ha Yeong Kang* · Chang Bo Oh** · Yong Sun Won** · J. Jay Liu*** · Chang Jun Lee**†

†Corresponding Author
Chang Jun Lee
Tel : +82-51-629-6465
E-mail : changjunlee@pknu.ac.kr

Received : December 1, 2020
Revised : December 29, 2020
Accepted : December 30, 2020

Abstract : To simulate a process model in the field of chemical engineering, it is very important to identify the physical properties of novel materials as well as existing materials. However, it is difficult to measure the physical properties throughout a set of experiments due to the potential risk and cost. To address this, this study aims to develop a property prediction model based on the group contribution method for aromatic chemical compounds including benzene rings. The benzene rings of aromatic materials have a significant impact on their physical properties. To establish the prediction model, 42 important functional groups that determine the physical properties are considered, and the total numbers of functional groups on 147 aromatic chemical compounds are counted to prepare a dataset. Support vector regression is employed to prepare a prediction model to handle sparse and high-dimensional data. To verify the efficacy of this study, the results of this study are compared with those of previous studies. Despite the different datasets in the previous studies, the comparison indicated the enhanced performance in this study. Moreover, there are few reports on predicting the physical properties of aromatic compounds. This study can provide an effective method to estimate the physical properties of unknown chemical compounds and contribute toward reducing the experimental efforts for measuring physical properties.

Key Words : group contribution method, functional group, support vector regression, property estimation

1. 서 론

유해 화학 물질을 기반으로 반응이나 증류 등 위험한 설비가 포함된 공정을 설계하는 경우, 물질의 물성치 정보는 필수적이다. 이를 통해 사전에 반응에 대한 위험성 분석이 가능하며, 특히 예상치 못한 폭주반응과 같은 잠재적 위험 요인을 추정할 수 있다. 물성치 자료를 얻기 위해서는 실험이 필수적이지만, 사전에 이론적으로 물질의 물성치를 예측한다면 실험에 드는 비용과 노력을 줄일 수 있다[1]. 물질의 물성치를 추정하는 방법은 크게 두 가지로 나뉜다. 하나는 분자 구조와 분자 간의 인력, 척력, 극성 등을 이용하여 계산하는 분자모델링 기법이고, 다른 하나는 분자에 존재하는 작용기를 이용하여 물성치를 예측하는 GCM(Group Contribution Method)이다. 분자모델링 기법은 어느 정도 정확성은 확보되지만 모델 준비에 시간이 오래 걸리는 단점을 가진다. 반면, GCM은 분자모델링과 비교하여 시간과 비용을 줄일 수 있다는 장점을 갖고 있다.

*부경대학교 안전공학과 석사과정 (Department of Safety Engineering, Pukyong National University)
**부경대학교 안전공학과 교수 (Department of Safety Engineering, Pukyong National University)
***부경대학교 화학공학과 교수 (Department of Chemical Engineering, Pukyong National University)

하지만 이 방법을 이용하여 물성치를 예측하는 경우도 정확성과 적용성 측면에서 단점을 가지고 있다. 예컨대, 모델 구축에 필요한 데이터베이스가 충분하지 않거나 작용기 분류가 지나치게 단순화되어 있어서 물성치를 예측할 수 있는 물질의 개수가 한정되며 정확성에도 문제가 있다는 것이다. 따라서 이러한 단점을 보완하여 물성을 예측할 수 있는 효율적인 모델 개발의 필요성이 요구된다.

Stefanis 등[1]은 방향족을 제외한 순수 유기화합물 334개의 임계온도에 대한 예측 모델을 연구하였으며, Lydersen[2]은 Alkane, Alkene, Alkyne 총 396개의 물질에 대한 임계온도와 임계압력 예측 모델을 연구하였다. Gharagheizi 등[3]은 1,378개 순수 화합물의 인화점만을 예측하는 모델을 연구하였으며, Kincewicz와 Reid[4]는 199개 화합물의 임계온도, 임계압력, 임계부피를 예측하는 모델을 개발하였다. Joback[5] 인공신경망을 이용하여 480개 화합물의 물성을 예측하는 모델을 만들었으며, Fedors[6]는 199개 화합물의 임계온도를 예측하는 모델을 수립하였다. GCM에 기반한 선행연구를 보면 대부분 임계점이나 인화점 예측에 관한 모델에 치중하고 있으며, 공정모델링이나 반응예측에 필요한 다른 물성치에 대한 연구는 부족한 실정이다.

본 연구에서 벤젠고리를 포함하는 방향족 화합물의 물성을 예측하는 모델을 개발한다. Lee와 Lee[7] 연구에서 화합물의 물성에 영향을 주는 56개의 작용기를 정의한 바가 있다. 이를 기반으로 먼저 147개의 방향족 화합물의 작용기를 조사하고, 그 결과를 바탕으로 전혀 존재하지 않는 14개의 작용기를 제외한 총 42개의 작용기를 확정하였다. 그리고 각 작용기의 개수를 매겨서, 이를 입력데이터로 구성하였다. 출력변수는 방향족 화합물의 물성치이다. 생성된 입력데이터의 경우 42개 변수 중 대부분의 값이 0인 고차원의 sparse data 형태를 보이게 된다. 이러한 데이터의 경우 대부분 통계기법으로 모델을 생성하는 경우 큰 어려움을 겪는다. 따라서, 본 연구에서는 sparse data를 잘 처리하는 통계기법인 SVR(Support Vector Regression) 을 기반으로 입력데이터를 이용하여 출력변수인 화합물의 물성치를 예측하는 모델을 만들고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 연구의 대상 물질인 방향족 화합물에 대한 설명과 사용된 연구 방법의 이론적 배경을 기술하였으며 3장에서 입력데이터의 생성과 이를 이용한 예측 모델에 대하여 설명하였다. 마지막으로 예측된 방향족 화합물의 물성치의 예측값과 해석모델의 성능에 대해 논하고자 한다.

2. 대상 물질 및 이론적 배경

2.1 방향족 화합물 특성

본 연구에서는 벤젠고리를 포함하는 방향족 화합물을 대상으로 물성 예측을 수행하려고 한다. 방향족 화합물을 예측 대상 물질로 고려한 이유는 다음과 같다. 벤젠고리를 설명하는 방향성(aromaticity)은 화학적 화합물의 열역학적 안정성을 설명하기 위해 사용되어왔다. 벤젠고리 그 자체로는 안정된 상태이지만, 결합 유형과 치환기에 따라 반응성이 달라지는 특이성을 가질 수 있다[8]. 한 예로, 벤젠고리에서 수소 원자 하나를 메틸기로 치환한 톨루엔은 불포화 탄화수소의 특성을 가지며 반응성이 매우 좋게 된다. McMurry[9]의 연구에 따르면, 방향족 치환반응에서 반응성은 공명효과와 유발효과의 상호작용에 영향을 받는 것으로 밝혀졌다. 유발효과란 전기음성도의 결합을 통해 전자를 끌거나 밀어주는 효과이며, 공명효과는 치환기와 벤젠고리에 있는 오비탈 사이 전자를 끌거나 밀어주는 효과를 뜻한다. 방향족 화합물의 반응성은 유발효과와 공명효과를 통해 벤젠고리에 전자를 밀어줄 때 벤젠고리 전자가 풍부하게 되어 반응성이 증가하는 특징을 가진다. 이처럼, 방향족 화합물은 결합과 치환 유형에 따라 물리적 성질이 바뀌는 특성이 있어서 물성값 예측이 어려운 특성이 있다.

또한, 방향족 화합물은 국내 석유화학 산업에서 생산되는 기초물질로서 상당히 많은 생산량을 차지하는 물질이다. Table 1에 나타낸 바와 같이, 대표적인 방향족 탄화수소인 BTX(Benzene, Toluene, Xylene)의 국내 생산량은 2019년을 기준으로 기초유분 생산량 중 약 40.9%를 차지하고 있다[10]. 또한, 산업안전보건기준에 관한 규칙에서는 관리대상 유기화합물 117종 중 벤젠고리를 포함하는 방향족 화합물은 27종을 차지하고 있다. 이처럼, 방향족 화합물은 널리 사용되고 있으므로, 사업장에서 방향족 화합물의 잠재적 위험을 관리하기 위해서는 무엇보다 물성치를 파악하는 것이 무엇보다 중요하다.

Table 1. Production capacity of BTX in Korea taken from[10]

(Unit: Thousand ton)

Benzene	Toluene	Xylene	Total
6,565	1,972	4,457	31,753

2.2 GCM(Group Contribution Method)

GCM은 물질을 이루는 작용기가 물성에 유의한 영향을 끼친다는 개념으로부터 출발한 기법으로, 이 개

념을 바탕으로 많은 연구가 이루어지고 있다[11-15]. 물질 중 가장 높은 비율을 차지하고 있는 요소는 탄소, 수소, 산소와 같은 원자, 단일, 이중, 삼중 화학결합 등이며 이런 결합들보다 더 복잡한 것은 원자와 화학결합으로 이루어진 작용기이다. GCM은 수백 개 물질의 물성을 예측하기 위해 단지 수십 개 정도의 작용기를 이용하기 때문에 분자 모델링과 같은 방법에 비해 필요로 하는 정보의 양이 크게 줄어드는 장점이 있다. 하지만 작용기가 지나치게 단순화되거나[16], 물성에 관련된 데이터베이스가 충분하지 않으면 예측된 물성치는 큰 오차를 보일 수 있으며[15], 또한, 회귀분석 모델은 일반적으로 내삽에는 적합하지만, 외삽에는 한계점을 가지고 있으므로 회귀분석에 기반한 모델을 만들 때는 최대한 많은 데이터를 이용하는 것이 매우 중요하다. 따라서 신뢰할 수 있으면서도 충분한 양의 데이터를 확보하고 적절한 작용기를 분류하는 것은 정확한 물성치 예측을 위해서 필수적이다. 본 연구에서는 미국화학공학회에서 추천되어 최근 활발히 사용되는 물성 데이터베이스인 DIPPR801 데이터를 사용하였다[17].

2.3 SVR(Support Vector Regression)

SVM(Support Vector Machine)은 Vapnik에 의해 고안된 기계학습 이론으로, 본 연구에서 사용한 입력 데이터와 같이 데이터가 Sparse data이며 고차원인 경우에 적합한 기계학습 방법론 이다[18]. 원래 분류 문제(Classification Problem) 해결을 위해 개발된 방법론인 SVM은 하나의 집단과 다른 집단을 분류하는 최적의 결정 경계면인 초평면(Hyper-plane)으로 정의되며 경계면에 가장 가까운 데이터를 Support vector라고 한다[13]. 이때 손실함수(Loss function)를 이용하면 SVM을 회귀문제로 확장할 수 있는데 이를 SVR이라 한다. SVR은 모든 데이터로부터 거리를 최소로 만드는 초평면을 찾는 것이 그 목적이 있다[19].

n개의 데이터(x_i, y_i)로 주어진 SVR문제는 y를 예측하고자 하는 최적의 초평면 $f(x)=\omega x+b$를 찾는 문제가 된다. 이때, 다양한 손실함수 중에서 초평면으로부터 각 데이터까지 거리를 ϵ보다 작아지게 만드는 ϵ-intensitive loss function을 사용하게 되면, 제약조건 $y_i-\omega x-b\le\epsilon$와 $\omega x+b-y_i\le\epsilon$에서 $\frac{1}{2}\|\omega\|^2$를 최소화시키는 문제가 된다. 그리고, 손실함수에 슬랙 변수(Slack variable, ξ & ξ^*)를 포함하여 손실함수의 범위를 좀 더 넓혀주면 예측오차 계산이 가능하게 되며, 이를 토대로 풀어야 할 최적화 문제는 다음과 같다.

$$\min\ J(\omega,\xi_i,\xi_i^*)=\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{n}(\xi_i+\xi_i^*) \quad (1)$$

$$\text{subject to}\ \ y_i-\omega x-b\le\epsilon+\xi_i$$
$$\omega x+b-y_i\le\epsilon+\xi_i^*$$
$$\xi_i,\xi_i^*\ge 0$$

이때 C는 성능과 오분류 간의 균형(trade-off)을 조절해주는 비용변수로 C가 커질 경우 훈련데이터가 과적합하게 되고, 적을 경우 풀이가 복잡해진다.

식 (1)로부터 얻어지는 최적 회귀함수는 다음과 같으며, 여기서 $\alpha_i\ge 0, \alpha_i^*\le C$이다.

$$f(x)=(\omega x)+b=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)(xx_i)+b \quad (2)$$

비선형 SVR의 경우 커널함수(kernel function) $K(x,x_i)$를 이용해서 x를 고차원 공간으로 사상시킨 후 선형 SVR로 다루게 되는데 이 때 식 (2)는 다음과 같이 변형된다.

$$f(x)=(\omega x)+b=\sum_{i=1}^{n}(\alpha_i,\alpha_i^*)K(x,x_i)+b \quad (3)$$

본 연구에서는 $K(x,x_i)=(x^Tx_i+c)^d,\ c>0$의 polynomial 커널함수를 사용하였다. SVR기반 회귀모델을 사용할 경우 매개변수는 실수이며, 사용자가 결정하여야 한다. 결정하여야 할 매개변수는 C(비용변수), ϵ(Intensitive loss function의 파라미터), Polynomial 커널함수의 차수 d이다. 본 연구에서는 매개변수의 최적해를 찾기 위해 PSO를 이용하였다.

2.4 PSO(Particle Swarm Optimization)

PSO는 최적해에 대한 근사해를 찾는 경험적 최적화 기법으로, 동물군집의 사회적인 행동양식을 바탕으로 개발된 이론이다. 군집(swarm)마다 각 개체(particle)는 다차원의 탐색공간을 옮겨가며 다른 대체들과 정보를 교환하는데, 자신과 이웃의 경험을 기반으로 한 정보를 이용해 최적의 해로 이동한다[20]. 매개변수가 많을 때도 적용 가능하며 초기값에 민감하지 않고 목적함수의 미분값을 필요로 하지 않는 장점이 있다[20-22]. 본 연구에서 사용한 SVM의 경우 모델 자체의 미분값을 이용하는 것이 불가능하므로 PSO를 채택하였다.

PSO의 구체적인 알고리즘은 Schwaab 등의 연구에서 찾을 수 있다[23]. 본 연구에서는 2,000개의 SVM의 매개

변수 조합을 랜덤으로 생성한 후, 최적의 매개변수를 찾을 때까지 최적의 매개 변수들을 탐색하였다.

3. 예측 모델 개발

3.1 입력변수를 위한 작용기 선택

본 연구에서는 Lee & Lee[7] 연구에서 제안한 총 56개의 작용기 그룹 중 42개의 작용기를 이용하여 입력데이터를 생성하였다. 입력변수를 분류해보면, Ending group 작용기는 12개, Middle group의 작용기는 16개, Ring group의 작용기는 11개이며, 작용기 이외 다른 변수로는 분자량과 이성질체 분류를 위한 변수 2개가 존재한다. 이성질체는 분자식은 동일하나 분자의 입체구조와 배열이 달라 물성이 다른 물질을 의미한다. 기존 물성 예측 연구에서는 이성질체 고려가 불가능하지만, 본 연구에서는 이성질체를 고려할 수 있도록 설계하였다. Table 3~4는 작용기의 위치가 다른 구조이성질체에 대한 분류를 보여주고 있다.

3.2 모델 구축을 위한 입력데이터 변환

입력데이터를 생성하기 위해서는 방향족 화합물의 분자구조를 조사한 후, 42개 작용기에 대한 개수를 조사해야 한다. 이를 위해서는 화합물의 명칭, CAS번호, 분자구조와 분자량을 파악해야 한다. Table 5~7은 화합물의 분자구조를 파악한 후 본 연구에서 정의한 42개의 작용기에 따라 입력변수를 생성하는 예를 보여주고 있다. 모든 방향족 화합물은 아래 Table 5에서 보는 바와 같이 42개의 입력변수를 갖게 된다. Table 7은 Table 6에 나와 있는 물질을 Table 5에 따라 그 입력값을 정리한 결과를 보여주고 있다. 총 입력변수의 개수는 42개이지만, Table 7에 표시된 작용기 이외 모든 작용기의 값은 0이다.

Table 2. 42 Functional groups as the input variables

Ⅰ-Ending Group -CH3 =CH2 ≡N −NH2 −NO2 −F −Cl −Phenyl −CO2H =O −OH -H
Ⅱ-Middle Group >C< >C= =C= -C≡ -CH2- >CH- -CH= >N- =N- -NH- -O- -CO- -Si- o-B m-B p-B
Ⅲ-Ring Group CH2 CH -CH >C −C N CO O =C R-C-R R-CH-R
Ⅳ-Molecular Weight
Ⅴ-The Distinction of the structure isomers (consisting of 3-branch benzenes) The Distinction of the structure isomers (consisting of 4-branch benzenes)

Table 3. An example of structural isomers consisting of 3-branch benzenes

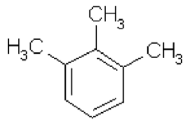
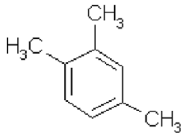
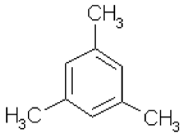
Structure [C9H12]			
Name	1,2,3-tri-methyl benzene	1,2,4-tri-methyl benzene	1,3,5-tri-methyl benzene
Input value	1	2	3

Table 4. An example of structural isomers consisting of 4-branch benzenes

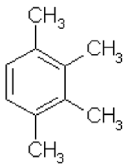
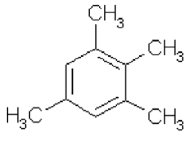
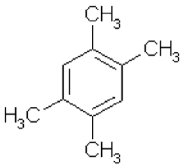
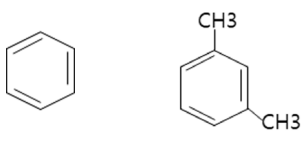
Structure [C10H14]			
Name	1,2,3,4-tetra-methyl benzene	1,2,3,5-tetra-methyl benzene	1,2,4,5-tetra-methyl benzene
Input value	1	2	3

Table 5. The input variables of the model

No.	Group	No.	Group	No.	Group
1(E1)	-CH3	13(M1)	>C<	29(R1)	CH2
2(E2)	=CH2	14(M2)	>C=	30(R2)	CH
3(E3)	≡N	15(M3)	=C=	31(R3)	-CH
4(E4)	-NH2	16(M4)	-C≡	32(R4)	>C
5(E5)	-NO2	17(M5)	-CH2-	33(R5)	-C
6(E6)	-F	18(M6)	>CH-	34(R8)	N
7(E7)	-Cl	19(M7)	-CH=	35(R9)	CO
8(E8)	-Phenyl	20(M8)	>N-	36(R10)	O
9(E9)	-CO2H	21(M9)	=N-	37(R11)	=C
10(E10)	=O	22(M10)	-NH-	38(R12)	R-C-R
11(E11)	-OH	23(M11)	-O-	39(R13)	R-CH-R
12(E12)	-H	24(M12)	-CO-	40	M.W
		25(M13)	-Si-	41	(1)
		26(M14)	o-B	42	(2)
		27(M15)	m-B	(1),(2) : Input variables for isomer compounds	
		28(M16)	p-B		

Table 6. Examples of the structure formula



[A] Benzene　[B] m-Xylene　[C] p-Chlorophenol

Table 7. Each functional group of examples in Table 6

Group	Functional group	A	B	C
Ending	-CH3	0	2	0
	-Cl	0	0	1
	-OH	0	0	1
Middle	m-B	0	1	0
Ring	CH	6	4	4
	-C	0	2	2

4. 모델 결과 분석

GCM과 PSO를 이용하여 147개 방향족 화합물의 8가지 물성치 예측하는 모델은 MATLAB을 이용하여 학습하였다. 입력데이터는 새롭게 정리한 42개의 작용기를 바탕으로 생성하였으며, 그 크기는 147×42이다. 출력데이터는 Table 8에 나와 있는 방향족 화합물의 8가지 물성치이며 따라서, 데이터의 크기는 147×8이다. SVR을 이용하여 예측 모델을 생성하고, 실제 물성치와의 비교를 통해 예측 모델의 성능을 분석하였다.

모델을 수립하는 과정은 일반적으로 Training이라 하며, 모델 검증은 Validation이라고 한다. 본 연구에서는 총 147개의 화합물 중 절반인 74개의 화합물을 이용하여 모델 Training을 실시하였고, 나머지 절반인 73개의 데이터를 이용하여 Validation 과정을 수행하였다. Table 9는 Benzene, m-xylenel, p-Chlorophenol의 실제값과 예측값의 비교 값을 보여주고 있다. 모델을 통해 예측된 물성은 실제값과 매우 유사함을 확인할 수 있다. Benzene의 경우 인화점과 임계온도의 오차가 다소 큰 것을 확인할 수 있으며, 이는 본 연구에서 제안한 모델의 문제점보다는 최적의 예측모델 파라미터 탐색의 문제이거나, 예측모델을 학습하는 경우 학습데이터 선택의 문제일 수 있다.

Fig. 1~4에는 예측 모델을 이용하여 방향족 화합물의 물성치를 예측한 값과 실제값의 비교를 보여주고 있다. X축은 실제값, Y축은 예측값이며, 따라서. 그림 가운데 표시한 실선 (Y=X)에 데이터가 물려 있을수록 예측 성능이 뛰어남을 의미한다. 결과를 살펴보면, 대부분 데이터가 이 실선 주변에 몰려 있음을 확인할 수 있다.

$$MAPE = \frac{\sum_{i=1}^{n} \frac{\left| Y_i - \hat{Y}_i \right|}{n} \cdot 100}{n} \quad (4)$$

Y_i : A real Value

$\hat{Y}_i$: A predictive value

Table 10은 본 연구에서 제안한 모델의 정확도를 측정하기 위해 식 (4)를 이용하여 예측값과 실제값의 MAPE(Mean Absolute Percent Errors)를 계산한 수치를 보여주고 있다.

MAPE는 실제값과 예측값의 모든 차이의 합을 퍼센트로 변환한 값으로서 오차를 직관적으로 알 수 있다는 장점이 있다. 이전 GCM연구에서 제시한 오차율과의 비교를 위해, Table 10에는 이전 연구의 MAPE와 본 연구에서 제안한 모델 간의 비교를 보여주고 있다. 기존 연구들은 대부분 특정 물성치만을 예측했으며, 데이터의 개수도 다르기 때문에 본 연구의 성능과 정확한 비교를 하기는 쉽지 않다. 그러나 기존연구의 MAPE 값보다 낮거나 유사한 값을 확인할 수 있으며, 기존 연구들의 경우 특정 물성치만을 예측한다는 점을 고려하면 본 연구에서 제안한 모델이 우수한 성능을 보여주고 있음을 확인할 수 있다.

Table 8. Physical properties from the proposed model

Symbol	Physical Properties	Units
FP	Flash Point	K
T_C	Critical Temperature	K
P_C	Critical Pressure	atm
V_C	Critical Volume	m^3
Z_C	Compressibility factor	
T_m	Melting Point	K
T_b	Boiling Point	K
H_C	Enthalpy of Combustion	$KJ/gmol$

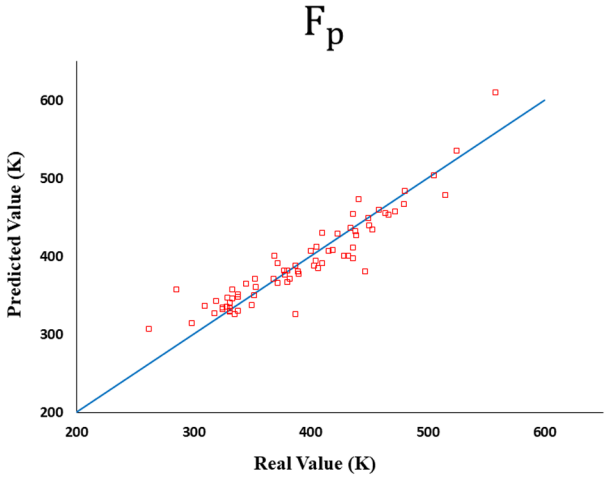


Fig. 1. The comparison of flash point between real and predicted values.

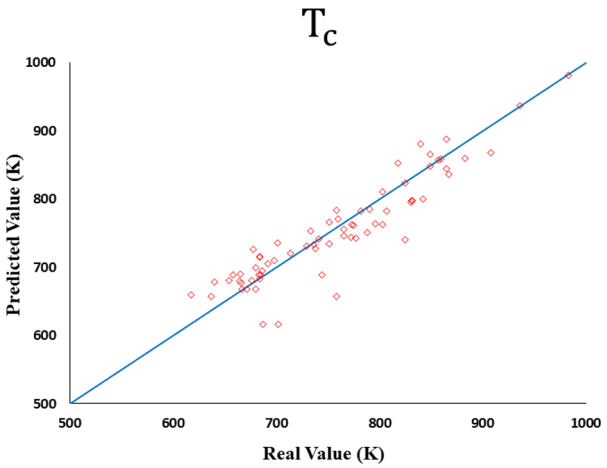


Fig. 2. The comparison of Critical Temperature between real and predicted values.

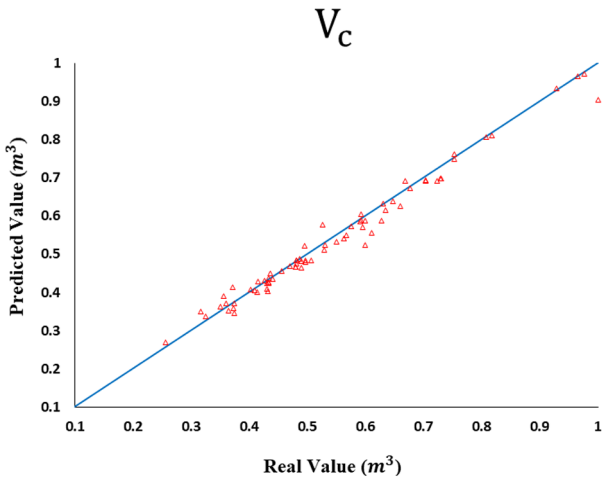


Fig. 3. The comparison of Critical Volume between real and predicted values.

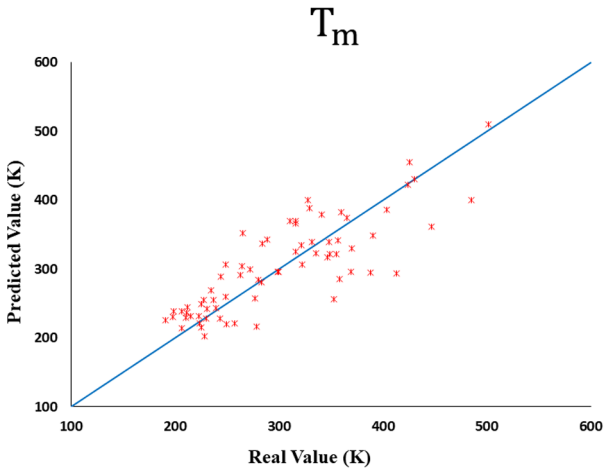


Fig. 4. The comparison of Melting Point between real and predicted values.

Table 9. Comparisons between real value and predicted values of benzen, m-xylene and p-Chlorophenol

	$FP(K)$	$T_C(K)$	$P_C(atm)$	$V_C(m^3)$	Z_C	$T_m(K)$	$T_b(K)$	$H_C(KJ/gmol)$
Benzene								
Real.	262	562.16	48.33	0.25	0.26	278.68	353.24	3136
Pred.	306.80	657.26	43.08	0.26	0.26	216.41	419.32	3383
Absolute Error	44.8	95.1	5.25	0.01	0	62.27	66.08	247
m-Xylene								
Real.	298.15	617.05	34.89	0.37	0.25	225.30	412.27	4331.8
Pred.	314.01	659.30	36.07	0.37	0.26	215.04	432.48	4038.54
Absolute Error	15.86	42	1.18	0	0.01	25.74	20.21	293.26
p-Chlorophenol								
Real.	389	738	52.50	0.32	0.28	316	493.11	2780
Pred.	380.43	727.48	44.94	0.33	0.27	324.65	494.02	2849.47
Absolute Error	8.57	10.52	7.56	0.01	0.01	8.65	0.91	69.47

Table 10. Mean absolute percent errors of previous studies and the proposed model

Method	Mean Absolute Percent Errors (%)			
	T_C	V_C	FP	V_P
Stefanis[1)]	-	-	-	10.82
Lydersen[2)]	-	8.9	-	-
Gharagheizi[3)]	-	-	9.94	-
Klincesicz and Reid[4)]	-	7.8	-	-
Joback[5)]	4.08	6.16	11.07	-
Fedors[6)]	5.0	-	3.15	-
Proposed	3.83	3.29	4.05	-

5. 결 론

본 연구에서는 GCM과 SVR을 기반으로 한 물성 예측 모델을 제안하였다. 기존연구의 공통적인 한계인 충분하지 않은 데이터베이스 개수와 작용기의 개수를 개선하기 위해 DIPPR801 데이터베이스 중 147개의 벤젠고리를 포함하는 방향족 화합물의 8개 물성을 이용하여 42개의 작용기를 이용한 입력데이터를 생성하였다. 입력데이터의 형태가 고차원의 sparse 데이터 특성을 갖기 때문에, 이를 가장 잘 다룰 수 있는 SVR을 이용하여 예측 모델을 만들었다. 예측 모델을 만드는 데 필요한 SVR의 매개변수는 경험적 최적화 기법 중 하나인 PSO를 이용하여 탐색하였다. 파라미터 최적화를 수행하였다. 실제값과 예측값의 차이인 오차는 MAPE를 기반으로 기존 연구와 비교하였다. 이때, 기존 연구들은 본 연구에서 이용한 데이터와 물질의 개수와 그 종류에 차이가 있으며, 따라서 기존연구의 결과와 비교하기에는 한계가 있다. 하지만, 대부분의 MAPE 값은 기존 연구와 비교하여 낮거나 유사한 값을 보여주고 있다.

기존 연구의 경우 특정한 물성치만을 예측하는 연구가 대부분인 점, 그리고 본 연구에서 다룬 물질의 개수가 다른 연구에 비해 다소 적은 점을 고려하면 기존 연구와 공정한 비교는 어렵다. 하지만, 본 연구에서 제안한 모델의 경우 다양한 물성치를 예측할 수 있다는 장점을 가지고 있다. 모델의 정확도의 경우 대상물질이 다르기 때문에 정확한 비교는 어렵지만, MAPE 수치를 통해 실제 물성치에 매우 근접한 물성치를 예측한다고 할 수 있다. 본 연구에서 물성 예측을 위해 개발한 모델은 향후 물성치를 실험하는 경우나 산업 현장의 공정설계 시 새로운 물질을 포함한 공정의 시뮬레이터를 개발하는 경우 제조 공정에 대한 중요한 물성 정보를 제공할 수 있다.

Acknowledgement: This research was supported by Pukyong National University Development Project Research Fund, 2020.

References

1) E. Stefanis, L. Constantinou, I. Tsivintzelis and C. Panayiotou, “New Group-Contribution Method for Predicting Temperature-dependent Properties of Pure Organic Compounds”, Int. J. Themophys., Vol. 26, No. 5, pp. 1369-1388, 2005.
2) A. L. Lydersen, “Estimation of Critical Properties of Organic Compounds by the Method of Group Contributions”, University of Wisconsin, 1955.
3) F. Gharagheizi, R. F. Alamdari and M. T. Angaji, “A New Neural Network - Group Contribution Method for Estimation of Flash Point Temperature of Pure Components”, Energy Fuels, Vol. 22, No. 3, pp. 1628-1635, 2008.
4) K. M. Klincewicz and R. C. Reid, “Estimation of Critical Properties with Group Contribution Methods”, AIChE J., Vol. 30, No. 1, pp. 137-142, 1984.
5) K. G. Joback, “A Unified Approach to Physical Property Estimation Using Multivariate Statistical Techniques”, Massachusetts Institute of Technology, 1984.
6) R. F. Fedors, “A Relationship between Chemical Structure and the Critical Temperature”, Chem. Eng. Commun., Vol. 16, No. 1-6, pp. 149-151, 1982.
7) C. J. Lee and J. M. Lee, “An Advanced Group Contribution Method for High-Dimensional, Sparse Data Sets”, Mol. Inform., Vol. 31, No. 1, pp. 41-52, 2012.
8) B. E. Poling, J. M. Prausnitz and J. P. O'connell “The Properties of Gases and Liquids”, Mcgraw-hill, Vol. 5, 2001.
9) J. E. McMurry, “Organic Chemistry; Nineth Edition”, Pearson, 2015.
10) Korea Petrochemical Industry Association, “Supply and Demand Status of Petrochemical Industry”, 2019.
11) L. Constantinou and R. Gani, “New Group Contribution Method for Estimating Properties of Pure Compounds”, AIChE J., Vol. 40, No10, pp. 1697-1710, 1994.
12) X. Wen and Y. Qiang, “A New Group Contribution Method for Estimating Critical Properties of Organic Compounds”, Ind. Eng. Chem. Res., Vol. 40, No. 26, pp. 6245-6250, 2001.
13) T. A. Albahri, “Structural Group Contribution Method for Predicting the Octane Number of Pure Hydrocarbon Liquids”, Ind. Eng. Chem. Res., Vol. 42, No. 3, pp. 657-662, 2003.
14) Z. Zbransk, J. Kukal, M. Zábranský and V. Růžička, “Estimation of the Heat Capacity of Organic Liquids as a Function of Temperature by a Three-Level Group Contribution Method”, Ind. Eng. Chem. Res., Vol. 47, No. 6, pp. 2075-2085, 2008.
15) C. J. Lee, G. Lee, W. So and E. S. Yoon, “A New Estimation Algorithm of Physical Properties based on a Group Contribution and Support Vector Machine”, Korean J. Chem. Eng., Vol. 25, No. 3, pp. 568-574, 2008.
16) L. Constantinou and R. Gani, “New Group Contribution Method for Estimating Properties of Pure Compounds”, AIChE J., Vol. 40, No. 10, pp. 1697-1710, 1994.
17) Design Institute for Physical Properties, “http://dippr.byu.edu/”, Retrieved on 09.30.2020.
18) V. N. Vapnik, “The nature of statistical learning Theory”, Springer-Verlag, 1995.
19) Y. Pan, J. Jiang, R. Wang, H. Cao and J. Zhao, “Quantitative Structure-Property Relationship Studies for Predicting Flash Points of Organic Compounds using Support Vector Machines”, QSAR Comb. Vol. 27, No. 8, pp. 1013-1019, 2008.
20) R. Poli, J. Kennedy and T. Blackwell, “Particle Swarm Optimization”, Swarm Intell., Vol. 1, No. 1, pp. 33-57, 2007.
21) S. Cha and C. J. Lee, “Study for the Plant Layout Optimization for the Ethylene Oxide Process based on Mathematical and Explosion Modeling”, J. Korean Soc. Saf., Vol. 35, No. 1, pp. 25-33, 2020.
22) P. J. Park and C. J. Lee, “The Research of Optimal Plant Layout Optimization based on Particle Swarm Optimization for Ethylene Oxide Plant”, J. Korean Soc. Saf., Vol. 30, No. 3, pp. 32-37, 2015.
23) M. Schwwab, E. C. Biscaia, J. L. Monteiro and J. C. Pinto, “Nonlinear Parameter Estimation through Particle Swarm Optimization”, Chem. Eng. Sci., Vol. 63, No. 6, pp. 1542-1552, 2008.