Printed in the Republic of Korea
ANALYTICAL SCIENCE & TECHNOLOGY
Vol. 28, No. 4, 270-277, 2015
http://dx.doi.org/10.5806/AST.2015.28.4.270

QSPR model for the boiling point of diverse organic compounds with applicability domain

Seong Eun Shin[1], Ji Young Cha[1], Kwang-Yon Kim[1,★] and Kyoung Tai No[1,2]

[1]Bioinformatics & Molecular Design Research Center, Seoul 120-749 Korea
[2]Department of Bioengineering, Yonsei University, 120-749 Korea

(Received June 18, 2015; Revised June 30, 2015; Accepted June 30, 2015)

다양한 유기화합물의 비등점 예측을 위한 QSPR 모델 및 이의 적용구역

신성은[1] · 차지영[1] · 김광연[1,★] · 노경태[1,2]

[1](사)분자설계연구소, [2]연세대학교 생명공학과

(2015. 6. 18. 접수, 2015. 6. 30. 수정, 2015. 6. 30. 승인)

Abstract: Boiling point (BP) is one of the most fundamental physicochemical properties of organic compounds to characterize and identify the thermal characteristics of target compounds. Previously developed QSPR equations, however, still had some limitation for the specific compounds, like high-energy molecules, mainly because of the lack of experimental data and less coverage. A large BP dataset of 5,923 solid organic compounds was finally secured in this study, after dedicated pre-filtration of experimental data from different sources, mostly consisting of compounds not only from common organic molecules but also from some specially used molecules, and those dataset was used to build the new BP prediction model. Various machine learning methods were performed for newly collected data based on meaningful 2D descriptor set. Results of combined check showed acceptable validity and robustness of our models, and consensus approaches of each model were also performed. Applicability domain of BP prediction model was shown based on descriptor of training set.

요 약: 비등점은 유기물의 물리화학적 성질을 특정하는데 있어 매우 근본적 요소 중 하나이다. 그러나 기존의 정량적 구조-물성 상관관계식들은 고에너지 물질 등과 같은 특정 물질 군에 대한 실험값들의 부족 등으로 인해 제한적인 응용성을 가지고 있었다. 본 연구에서는 서로 다른 출처로부터의 5,923개의 비등점 자료를 확보하였으며, 이에는 일반적 유기화합물과 더불어 특수목적을 가지는 분자들을 포함하였고, 이들 수집된 데이터 셋을 이용하여 새로운 비등점 예측모델을 개발하는데 사용하였다. 다양한 학습 방법을 이용하여 새로이 수집된 데이터 셋을 이용한 2차원 분자 표현자에 기반한 비등점 모델을 도출하였다. 개발된 예측모델의 적정성과 견고성을 확인하였고, 훈련 셋의 표현자에 기반한 비등점 예측모델의 적용구역을 도출하였다.

Key words: Boiling point, QSPR, machine learning, applicability domain

★ Corresponding author
Phone : +82-(0)2-393-9550 Fax : +82-(0)2-393-9554
E-mail : kykim@bmdrc.org

1. 서 론

비등점은 액체 물질이 고체로 변화하는 상변화 과정에서 증기압과 외부 압력이 같아져 상변화가 발생하는 온도를 나타내며 비등점이라고도 칭한다. 비등점과 같은 상변화 온도는 외부 압력에 따라 가변적이며 일반적인 정상 비등점의 경우 760 mmHg의 압력 즉 1 기압 하에서 측정된 끓는 온도를 말한다. 물질의 비등점은 물질의 고유한 특성으로 혼합물의 분리, 정제나 제조 공정 등에서 중요한 역할을 하며, 증기압이나 증기열,[1] 액체의 점도[2] 등과 관련한 기타 물성의 예측과 측정에 중요한 파라미터로 사용될 수 있다.

순수한 화합물질의 물성을 파악하는 가장 정확한 방법은 실험을 통한 측정이나, 측정 과정에서 시료의 정제, 실험환경 구축, 시료의 안정성 판단 등 번거롭거나 시간, 비용적 소모가 큰 과정이 동반 될 수 있다는 단점이 있다. 따라서 이와 같은 과정을 대체하기 위한 예측 방법의 모색이 진행되어 왔다. 화합물의 대표적인 물성인 비등점의 경우 오랜 시간에 걸쳐 연구되어 왔기 때문에 사전 데이터가 많다는 장점이 있으며 이는 새로운 미지물질에 대한 예측을 진행하는데 도움이 될 수 있다. 미지물질의 비등점 예측 방법은 화합물의 구조적 안정성에 기반한 양자역학적 계산 방법, 유사한 구조나 특정 부분 구조의 영향, 특정 그룹의 분류법 등을 통해 계산한 구조-활성 관계(SAR; Structure-Activity Relationship) 방법이나 유사상관성(Read-across), 그룹 기여도 방법(Group contribution)등이 있으며 실험데이터와 화합물의 구조, 물리 및 화학적 성질 등의 표현자들을 기반으로 예측을 진행한다. 액체에서 기체로의 상 변화 과정에서 주변에서 흡수한 열 에너지는 상 변화를 위해 사용되며 이때 물질은 평형상태로 존재하게 된다. 따라서 자유에너지의 변화량은 0으로 표현되며 열역학적 변수에 따라 다음과 같은 수식으로 표현될 수 있다.

$$\Delta G_b = \Delta H_b - T_b \Delta S_b = 0$$

따라서 비등점의 온도(T_b)는 다음과 같이 엔탈피(H)와 엔트로피(S)의 변화량으로 표현될 수 있다.

$$T_b = \frac{\Delta H_b}{\Delta S_b}$$

엔탈피(H)의 경우 내부 압력, 부피와 관련 되기 때문에 해당 화합물의 분자 사이의 결합, 즉 분자간 상호작용(inter-molecular interaction) 중 이온간의 힘, 쌍극자 힘에 따른 분극률(polarizability), 분산력, 수소결합과 같은 분자의 특성과 비등점을 관련 지어 설명할 수 있다. 엔트로피의 변화의 경우 분자의 위치, 회전, 구조적 유연성 등으로 설명할 수 있는데, 비등점의 경우 액체에서 기체로 변화는 비교적 자유로운 상 변화이므로, 분자의 대칭성이나 위치 등 보다 결합 형태나 세기에 대한 공유결합, 극성의 정도에 크게 영향을 받는다. 이와 같이 화합물의 비등점은 분자의 구조, 형태, 결합과 밀접한 연관을 지니고 있으며 이와 같은 물리적, 화학적 성질과 관련하여 구조에 따른 화합물의 비등점 예측이 다수 진행되어 왔다.

비등점을 예측하기 위한 다수의 모델 중 가장 널리 사용된 방법으로, 분자의 특정 작용그룹이나 결합에 따른 그룹 기여도(Group contribution method)를 사용한 모델[3]이 있다. 이는 분자의 특정 작용 그룹에 대해서 언급한 엔탈피 등의 변화에 따른 에너지 변화나 물성 값의 변화 통해 기여도를 설정하고 이들의 조합을 통하여 해당 분자의 비등점을 예측한다. 이와 같은 방법은 측정 부분 구조에 의해서 모델을 설정하기 때문에 해석이 용이하고 적용 방법이 간단하다는 장점을 지니나 특정 부분구조의 존재에 의지하기 때문에 적용 가능한 분자가 한정적이며 분자 결합구조에 따른 다형성(polymorphism)을 설명하기 어렵다는 단점을 지닌다.

따라서 본 연구에서는 정량적 구조-물성 관계 (QSPR, Quantitative Structure-Property Relationship)식을 통하여 구조에서 계산할 수 있는 다수의 분자 표현자와 해당 물성과의 상관관계를 표현한다. QSPR방법은 정량적 구조-활성 관계(QSAR, Quantitative Structure-Activity Relationship) 방법의 한가지로 예측 할 수 있는 다양한 활성 중 물성에 초점을 맞춘다는 점에 차이가 있다. QSPR 방법은 구조적 정보를 설명할 수 있는 설명인자인 표현자(descriptor)에 의해 의존하며 어떠한 표현자와 통계학적 계산 방법을 사용하는지에 크게 의존한다. 과거의 몇몇 모델의 경우 역시 그룹 기여도 방법에 비해 모델을 향상 시키기 위한 방법으로 QSPR방법을 사용하였으며 양자 역학적으로 계산한 에너지 변화를 표현자로 사용하는 방법, 물리화학적 성질 중 중요한 특성으로 꼽히는 작용기, 분자 골격구조를 중심으로 나뉜 클러스터에 따라 가중치를 부여하고 계산한 방법 등이 개발되어 왔다. 위와 같은 방법들은 기타 모델에 비해 예측의 정확도는 높을 수 있으나 특수한 화합물만을 훈련데이터로 모델을 생성하였기 때문에 적용 가능한 물질에 한계를 지니거나

양자역학적 계산방법과 3 차원 구조 안정화 과정을 포함하기 때문에 상대적 계산량이 많다는 단점을 지닌다. 이에 본 연구에서는 2 차원적 구조에서 계산할 수 있는 재현 및 설명이 가능한 표현자들을 사용하여, 별도의 데이터 분류가 없이 다양한 화합 물질에 대한 비등점의 예측이 가능한 QSPR 기반 예측모델을 개발하고자 하였다.

2. 연구 방법

Fig. 1은 본 연구에서 수행한 QSPR 예측모델 개발 과정을 도표로 나타낸 것이다.[4] 전체적으로 데이터 수집 및 정제, 분자 표현자 계산, 훈련 셋과 시험 셋의 분류, 적합한 표현자의 선택 및 단일/조합모델 생성, 외부검증, 적용 구역 설정 순으로 진행하였다.

2.1. 비등점 데이터

QSPR모델링을 진행하기 앞서 모델의 훈련과 검증을 위한 데이터 수집 과정을 진행한다. 본 모델에서는 무료로 접근할 수 있는 EPI Suite에서 제공한 SRC's PHYSPROP database[5]와 고에너지 물질과 같은 특수 기능화합물에 대한 ICT database,[6] DIPPR database[7]를

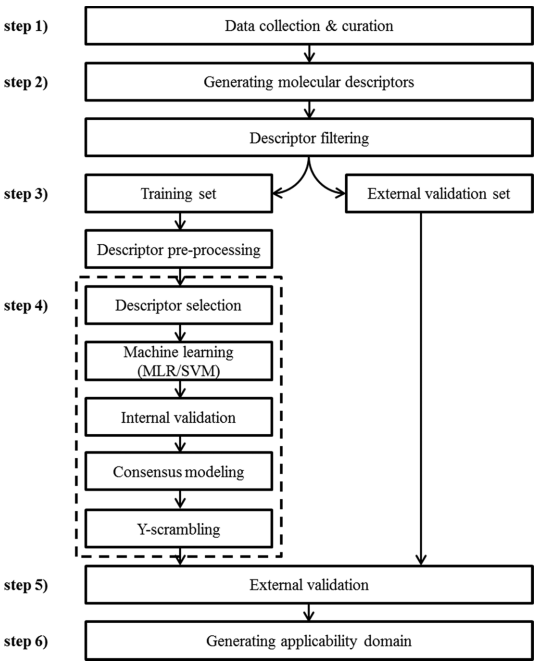

Fig. 1. Schematic diagram of QSAR model development for predicting boiling points.

통한 데이터 수집을 진행하였다. 비등점의 측정 실험 방법에 대해서는 특별히 제한을 두지 않았으나 동일 화합물질에 대해 반복된 실험이 존재하는 경우에는 표준오차가 10%이상인 데이터는 제거해주며 그 이하의 오차 값을 지닐 경우에는 정중값(median)값을 사용하여 빈도가 높은 값에 가중치를 주어 수집했다. 특히 본 모델에서는 순수한 단일화합 물질 데이터를 사용한 모델 개발을 목적으로 하였기 때문에, 2 가지 이상의 구성요소가 포함된 혼합물이나 잘 알려진 염, 화학적 특이성을 보일 수 있는 금속물질, 고분자복합체 등을 제외시키는 데이터 정제 과정을 진행하였다. 즉 수집한 데이터는 일반적인 유기화합물인 탄소(C), 수소(H), 산소(O), 질소(N), 황(S), 인(P)으로 구성된 화합물과 할로겐 화합물(F(불소), Cl(염소), Br(브롬), I(요드)을 포함하는 화합물)로 구성된다. 따라서 수집한 9 천여개의 화합물 데이터 중 위와 같은 거르기 작업을 통하여 5923 개의 데이터를 통해 모델을 생성, 검증하였다. 수집한 데이터는 통계적 모델링을 위하여 실험값의 분포가 정규분포와 유사하거나 대칭적인 구조를 나타낼 수 있어야 한다. 본 연구에서 수집된 데이터는 섭씨(ºC) 단위이며 수집한 데이터의 분포는 *Fig.* 2와 같이 정규분포와 유사하다고 판단하여 별도의 단위 변환 없이 예측을 위한 모델링에 사용하였다.

2.2. 분자 표현자(Molecular Descriptor) 계산

본 연구에서는 수집한 데이터에 입력한 이차원 평면구조에서 구조적, 정정기적, 기하학적, 물리화학적, 위상학적 표현자 등으로 구성되는 이차원적 분자 표현자를 계산하여 분자의 특징을 수치로 나타내었으며 이러한 표현자들의 값을 통해 화합물이 구별되게 된다. 본 모델에서 사용된 이차원 분자 표현자의 경우 PreADMET[8]에서 계산되었으며, 일부 훈련데이터에서 계산할 수 없는 표현자들을 제거해 주어 이후 통계학적 계산이 용이하도록 하였다. 그러나 예외로 다수의 화합물의 물성과 높은 상관관계를 지닐 것으로 예측되는 극성 편극도를 나타내는 표현자나 수소결합을 나타내는 표현자 등의 표현자가 10 개 내외의 화합물질에서 계산이 되지 않는 경우, 해당 화합물질의 구조적 특이성을 고려하여 표현자가 아닌 화합물 자체를 제거 하였다.

2.3. 훈련 셋(training set)과 시험 셋(test set)의 분류

모델의 검증을 위한 추가 데이터 수집이 어렵기 때

문에 수집한 5923개의 화합물 중 무작위로 선택한 약 50%, 즉 2962 개의 데이터를 훈련 셋(training set)으로, 나머지 2961 개의 데이터를 외부 검증 셋(external validation set)으로 분류하였다. QSPR 예측 모델은 훈련 셋 사이에서만 훈련되며 외부검증 셋은 최종으로 결정된 모델의 검증에만 사용된다. 훈련 셋으로 분류된 분자의 경우 데이터 예측에서 불필요한 시간 소모를 줄이기 위하여 표현자의 우선 여과 작업을 진행한다. 모든 훈련 셋에서 각 표현자의 분산이 0.01이하면 표현자가 해당 물성 값의 구분에 대한 변별력을 갖지 못한다고 판단하여 해당 표현자를 제거하며 물성과 표현자의 1:1 상관관계가 0.01 이하인 경우 모델을 생성할 때에는 다른 표현자와의 조합으로도 큰 영향을 주지 않을 것으로 판단하여 해당 표현자를 제거한다. 또한 표현자간의 상관관계가 0.9 이상이면 상호간에 의미의 중복이 발생될 수 있으므로 두 표현자 중 덜 의미있거나 해석이 용이하지 않다고 판단 되는 쪽을 제거한다.

2.4. 적합한 표현자의 선택 및 모델 생성

우선 여과(pre-filtering)된 210 개의 표현자 중에서 실제 모델에 사용되는 표현자를 선택하는 방법은 다음과 같다. 부트스트랩 선택 방법을 사용하여 무작위로 실제 훈련데이터와 내부 검증 데이터를 만들고 실제 훈련 데이터에서 전진선택(forward feature selection) 기법을 통하여 실제 모델에 사용될 표현자를 선별한다. 이 때 기준이 되는 값은 해당하는 기계학습법(machine learning)의 내부 검증 성능 값이며 적당한 표현자의 수에서 최적의 성능을 나타낼 때까지 반복된다. 내부 검증 방법으로 사용된 방법은 부트스트랩 선택방법으로서, 이는 전체 훈련 셋에서 무작위로 데이터를 추출하는 것을 반복하는 방법인데, 매 회 추출된 실제 훈련 셋 내에서 모델의 성능을 향상시키는 표현자를 순차적으로 추가해서 선택하는 전진선택 방법과 기계학습법을 진행한 부트스트랩 내부 검증 결과를 확인하는 것을 병행하여 표현자의 선택과 최적화 과정을 진행한다. 선택되는 표현자의 수가 증가하면 대부분 모델의 성능은 개선되나 모델의 예측에서 과-적합(overfitting)이 가능하므로 실제 훈련 데이터 셋의 10%를 넘지 않는 한도에서 표현자의 수를 결정하며 그 기준은 부트스트랩 선택 방법을 사용하여 내부검증 된 데이터의 RMSE (root mean square error) 값으로 한다. n 개의 화합물이 있고 모델에 의해 계산된 예측 값(y_{pred})이 존재할 때, RMSE의 결정 식은 다음과 같다.

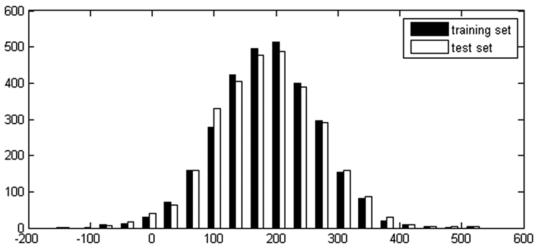

Fig. 2. Distribution of collected boiling point data after splitting into training set and test set.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} |y_{pred,i} - y_i|}{n}}$$

다중 선형 회귀식(multiple linear regression, MLR)의 경우 사용된 표현자들 사이에 강한 상관관계가 나타나는 다중 공선성 문제를 방지하도록 모델에 선택된 표현자 사이의 VIF(variance inflation factor)값을 계산한다. 다중 공선성이 발생하는 경우 표현자의 VIF값은 커지며 10이상인 경우 사용하지 않는 것이 좋은 것으로 알려져 있다.

비선형 모델인 서포트 벡터 머신(support vector machine, SVM)의 경우 RBF(radial basis function) 커널을 사용하여 회귀식을 생성하였으며, 표현자의 선택 외에 gamma, C, nu, epsilon값을 통하여 최적화를 진행하였다. C는 값의 오류에 대한 페널티 인자로, C값이 클수록 더 오류에 민감하게 반응하며 100을 초과하는 경우 과-적합 될 수 이으므로 1이상 100이하인 값에서 최적화를 진행하였다. Gamma는 초-평면을 결정하는 환경, 즉 벡터의 양을 결정하며 너무 적은 값에서는 예측률이 떨어지고 너무 높은 값에서는 모델링의 효율을 감소시키면서 과-적합을 발생 시킨다. nu 값은 훈련의 오류와 관련된 값으로, nu값이 크면 모델의 예외가 발생할 확률이 커진다. 일반적으로 gamma와 nu값의 최적화 이후에 epsilon 값의 최적화를 진행한다. 일반적으로 epsilon 값은 기본으로 설정된 0.001에서 변화시키지 않지만 큰 epsilon 값을 사용하여 계산 시간을 줄이고 다른 파라미터 값들을 최적화 할 경우 또는 훨씬 작은 값의 epsilon 값을 사용하여 더 세밀한 인식률을 사용해야 할 때 변경해 줄 수 있다. 본 모델에서는 epsilon 값을 0.001로 고정하고 C, gamma, nu 값을 순차적으로 변경하여 최적화 된 서포트 벡터 머신의 인자 값을 설정하고자 하였으며, 최적의 매개변수값에 대해 C = 100, gamma = 0.01, 그리

Table 1. Statistical performance of MLR, SVM and consensus results QSPR prediction models in this study.

Model (#descriptors)	Training Set (n=2962)			Bootstrapping validation set			External test set (n=2961)		
	R^2	RMSE[a]	MAE[b]	R^2_{boot}	$RMSE_{boot}$	MAE_{boot}	R^2_{ext}	$RMSE_{ext}$	MAE_{ext}
MLR(14)	0.890	26.840	17.800	0.888±0.011	27.016±1.222	17.898±0.502	0.875	29.730	18.976
MLR(17)	0.899	25.778	17.266	0.898±0.008	25.893±0.982	17.399±0.451	0.883	28.718	18.345
SVM(10)	0.842	32.714	21.270	0.838±0.009	32.805±1.264	21.429±0.521	0.831	34.913	22.355
opt-SVM(10)	0.910	24.354	15.829	0.868±0.011	29.510±1.279	18.923±0.555	0.871	30.107	18.631
MLR(14) +SVM(10)	0.892	27.116	17.737	0.886±0.009	27.785±1.061	18.274±0.530	0.883	28.950	18.681
MLR(14) + opt-SVM(10)	**0.915**	**23.633**	**15.363**	**0.898±0.008**	**25.868±1.013**	**16.748±0.470**	**0.892**	**27.499**	**17.128**
Y-Scrambling (MLR(14))	0.009	80.559	63.847	0.005±0.003	80.399±1.521	63.789±1.173	-	-	-
Y-Scrambling (opt-SVM(10))	0.001	84.859	67.350	0.001±1.005	80.951±1.333	64.265±0.005	-	-	-

[a]root mean square error, [b]mean absolute error.

고 nu = 0.5로 결정할 수 있었다.

표현자를 증가시키며 최적화된 각 기계학습 모델과 결과를 누적하고, 계산된 각 모델 중 두 가지 이상을 조합한 모델을 통하여 각 모델에서 발생하는 오차를 감소시키는 조합모델(consensus model)을 생성하여 상호 비교하면서 최종 모델을 결정하였다. 전체적인 통계기법에 기반한 표현자의 선택 및 모델의 도출은 RapidMiner[9]를 통해 수행되었다.

2.5. 적용구역

훈련 셋에서 사용된 구조 데이터 중 모델에서 사용된 표현자들을 통하여 모델에서 설명할 수 있는 분자에 대해 유사도를 평가한다. Tropsha[10] 등이 제안한 k-NN 기반의 적용 가능한 범위를 통해 화학물질간의 거리를 유클리드 거리로 유사도를 계산 하였으며 훈련 셋에서 계산된 거리의 평균(avg(D))와 표준편차(σ) 값을 통해 다음 수식에 따라 한계 값을 설정하였다.

$$D_{threshold} = avg(D) + Z\sigma$$

Z는 경험적으로 계산에 따라 얻는 값으로 본 연구에서는 0.5로 고정하였다. 외부 검정 데이터의 경우 훈련 셋에서 설정한 적용 가능한 범위에 포함되는 물질에 한해 신뢰할 수 있는 결과로 정의하고 성능을 평가하였다. 조합모델을 사용한 경우 각 해당하는 모델에 따른 적용범위에 따라 예측하며 중첩되는 범위에서 조합모델을 적용하였다.

3. 결과 및 토론

3.1. 학습 방법에 따른 모델 및 조합모델

본 연구에서 사용된 2가지 학습방법인 MLR 기법 및 SVM 기법에 대해 전진선택(forward selection) 기법을 이용하여 선택되는 분자표현자의 개수를 증가시키며 이들에 대한 부트스트랩 기법에서 내부검증데이터의 $RMSE_{boot}$ 값이 최소가 되는 지점의 모델을 최적 모델로 선택하고자 하였다. 이러한 훈련에 의한 결과로 도출된 MLR 및 SVM 모델의 성능결과를 *Table* 1에 나타내었다. MLR 모델은 14종 및 17종의 표현자가, 그리고 SVM 모델은 10종의 표현자 및 최적화된 커널에 의한 10종의 표현자가 포함되었을 때 최적의 성능을 나타내는 것을 확인할 수 있었으며, 각 모델에 대해 R^2_{ext} 값이 MLR 모델의 경우 0.875~0.883, SVM 모델의 경우 0.831~0.871의 범위를 보이는 것을 확인할 수 있었다.

MLR 및 SVM 모델 중에서 예측결과를 혼합하여 평균 처리한 조합모델을 생성하여 기존의 모델들과 동일하게 훈련데이터, 부스트래핑 방법, 외부테스트 데이터에 대한 성능결과도 함께 비교하였다. 결과적으로 복합모델들의 경우 MLR 또는 SVM개별 모델들에 비해 전체적인 예측성능이 더 향상된 것을 볼 수 있었으며, 조합모델 중에서 표현자 14종을 포함하는 MLR모델과 표현자 10종을 포함하는 최적화된 SVM 모델을 혼합한 조합모델의 내부검증 및 외부테스트 데이터의 예측결과도 가장 우수한 것을 볼 수 있었다. 또한 모델의 표현자가 우연하게 적합하였을 가능성을 검증하기 위하여 Y-scrambling 과정을 수행하였고 이 성능평가의 평균결과를 *Table* 1에 같이 나타내었다. 여기서 Y-scrambling이 수행된 모델들의 결과는 기존 MLR 및 SVM의 결과와 비교하여 볼 때 MLR의 경우 R^2_{boot}=0.005의 값을, SVM의 경우 R^2_{boot}=0.001의 매우 큰 차이를 보이는 것을 확인할 수 있었으며, 이에 본 연구에서 개발된 모델들이 우연적으로 생성된 관계가 아님을 확인할 수 있었다. 종합적으로 개발된 조합모

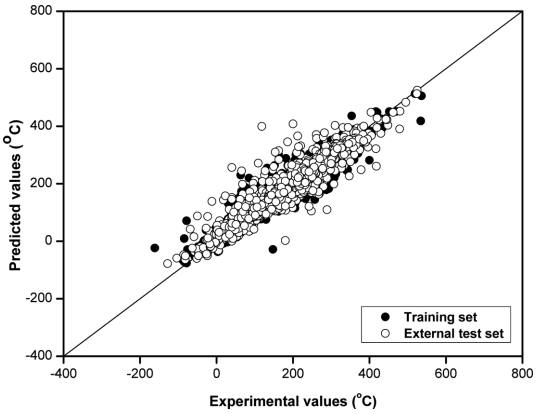

Fig. 3. Correlation plot of experimental versus predicted values of boiling point for consensus model.

델을 이용하여 훈련데이터와 외부테스트 데이터에 대해 그 실험값과 예측값을 *Fig.* 3에 비교하였으며 매우 우수한 상관관계를 보이고 있음을 확인할 수 있었다.

3.2. QSAR 모델 분석

MLR 모델은 원리적으로 각 분자 표현자들의 선형적인 합으로 표현되기에 타 학습기법인 SVM, ANN 기법에 비해 선택된 표현자들의 분석을 통해 그 분석이 용이한 면이 있다. 이에 본 연구에서 개발된 MLR 모델의 14 개의 선택된 표현자를 분석하여 분석하고자 하였으며, 이를 *Table* 2에 나타내었다. *Table* 2에 나타난 회귀계수(t-ratio) 중에서 가장 큰 표현자는 Polarizability_Miller이며 이 외에 Fraction_of_2D_VSA_HBond_all, No_single_bonds, Total_structure_connectivity_index의 순으로 그 기여도가 큰 것을 확인할 수 있었다. 가장 큰 역할을 하는 Polarizability_Miller 표현자는 Miller 기법에 의해 도출된 원자 분극성의 합을 의미하며, 분자간 비공유 결합에서의 분극성이 주요한 역할을 하는 것을 확인할 수 있다. 두 번째로 큰 영향을 미치는 요소인 Fraction_of_2D_VSA_Hbond_all은 van der Vaals(VDW) 표면에서의 수소결합에 대한 표면적 비율을 의미하는 것으로서 한 분자에 대해 수소결합이 가능한 표면분율이 높을 수록 분자간 인력이 증대되어 비등점이 증대될 수 있는 예상과 일치한다. 또한 No_single_bonds 및 Total_structure_connectivity_index는 분자의 유연성(flexibiligy) 및 가지화(branching)와 관련되며 음의 회귀계수를 가지는 것과 같이 고려

Table 2. Selected descriptors and its coefficients, t-ratio, and VIF in the best MLR model for boiling point prediction.

No.	Molecular descriptors	Description	Coeffi-cient	t-ratio	VIF[a]
1	Polarizability_Miller	The sum of the atomic polarizabilities (calculated by Miller method)	14.771	54.189	4.566
2	Fraction_of_2D_VSA_Hbond_all	Fraction of 2D Van der Waals H bond surface area	272.778	38.698	6.135
3	E_state_min	Minimum atomic E-state value	7.136	12.958	8.621
4	No_single_bonds	The number of single bonds	-3.684	-32.740	6.667
5	Total_structure_connectivity_index	Total structure connectivity index	-136.934	-24.857	7.519
6	Molecular_weight	Molecular weight	0.301	17.161	8.264
7	E_state_SssCH2	Sum of E-state for ssCH2 type	3.524	14.842	8.475
8	AlogP98_value	AlogP98 (calculated logP by Ghose method)	-12.631	-16.766	8.333
9	SCAA2	The atomic charge weighted acceptor atoms VDW surface area divided by total VDW surface area	-8.623	-20.742	7.937
10	2D_VSA_Hbond_acceptor	2D Van der Waals Hbond acceptor	-1.095	-18.918	8.130
11	Formal_charge	Formal charge	-25.276	-12.518	8.621
12	CATS_binary_Hyd_Acc_02	Chemical Advanced Template Search descriptor, binary (Hydrophobic-Acceptor) 2	-17.414	-11.420	8.696
13	WNSA1	The partial negative VDW surface area multiplied by the total VDW surface area and divided by 1000	-0.886	-9.195	8.850
14	CATS_binary_Hyd_Neg_02	Chemical Advanced Template Search descriptor, binary (Hydrophobic-Negative) 2	22.081	6.385	9.009

[a]valence inflation factor

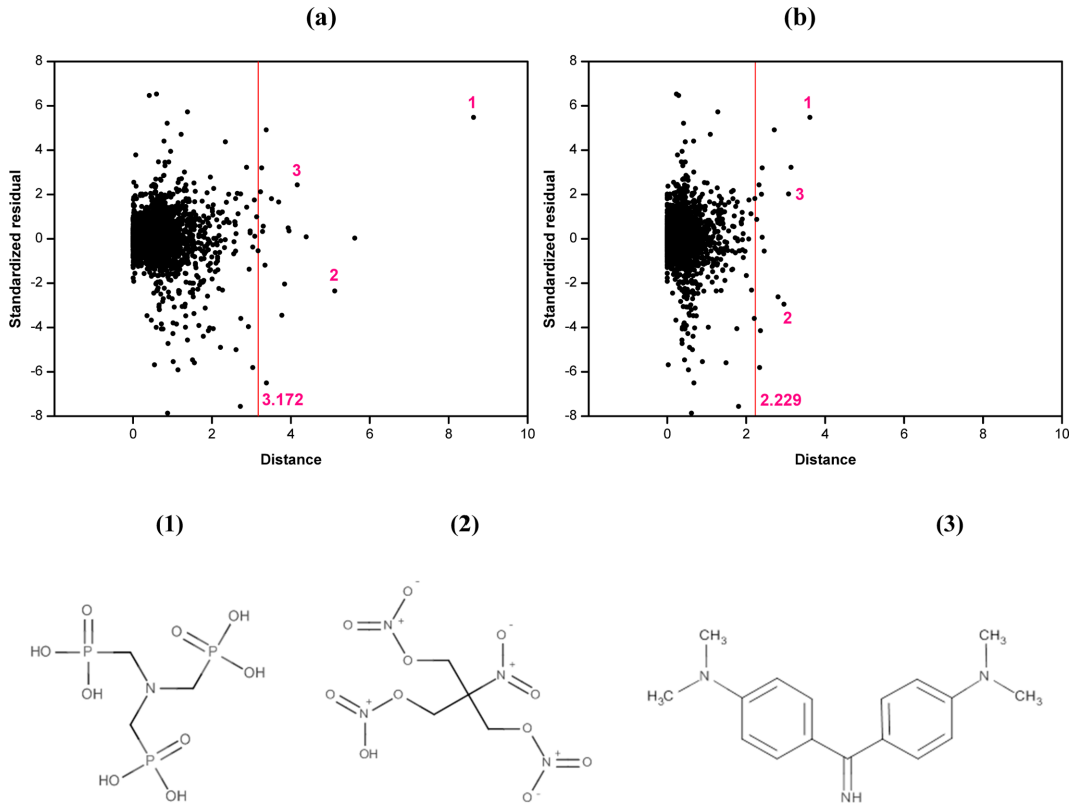

Fig. 4. Plots of standard residuals versus distance between external test set compounds using (a) 14-variable MLR and (b) 10-variable SVM model. Three representative outlied comcounds are shown below as (1), (2), and (3).

할때 선형의 유연성이 적은 분자형태일수록 분자간 인력이 증대되고 이에 따라 비등점이 증대될 수 있음을 예상할 수 있다. 따라서 MLR 모델에 의한 분석결과를 볼 때, 화학분자에 대한 비등점을 결정하는 요소로서 대해 분자의 분극성, 수소결합 가능 면적비율, 그리고 분자구조의 유연성 및 가지화 등이 가장 큰 역할을 하는 것을 확인할 수 있었다.

3.3. 적용 범위

개발된 모델들은 사용된 훈련데이터에 의해 학습되었기에 새로운 물질의 예측 시에는 사용된 훈련데이터와의 유사성이 보장되어야만 신뢰성 있는 예측결과를 부여할 수 있다. 모델의 예측신뢰성을 표현하기 위해 훈련데이터의 분자 표현자를 기반으로 나타낸 kNN기법을 이용하여 분자유사성으로 설정하고자 하였다. 이러한 설정된 적용범위를 적용한 결과를 *Fig.* 4에 나타내었다. 여기서 적용범위를 벗어난 화합물들의 구조를 확인한 결과 *Fig.* 4의 아래에 나타낸 바와 같이 다중 인산기 및 다중 질산기가 포함되어 있거나 디페닐메탄이민 계열이 포함되고 있음을 확인할 수 있었다. 이러한 경우는 비등점을 예측하는데 있어서 이러한 다중 음이온기 또는 방향족 메탄이민기 계열의 특성을 표현할 수 있는 표현자를 모델에 사용하지 못하였거나 선택된 표현자들로 설명되어지는 메커니즘과 상이한 특이성을 보일 수 있다는 점을 예상할 수 있고, 추후 예측모델의 적용범위를 고려할 때, 이러한 다중 음이온기 혹은 방향족 메탄이민기 계열의 분자들은 예측 시 제외되는 것이 타당하다고 예상된다.

4. 결 론

본 연구에서는 QSPR 기법을 통하여 2 차원 화학구조에 기반하여 활용 가능한 다양한 유기화합물에 대한 비등점을 예측 할 수 있는 적합한 모델을 개발하였다. 개발된 모델은 적합성, 우연성 신뢰성 및 예측능력 등을 확인하였으며, MLR, SVM 모델 각각 2 종

및 이에 의한 조합모델을 도출하였고, 최종적으로 개발되어진 최적의 조합모델의 성능을 확인할 수 있었다. 훈련 셋의 분자유사성에 기반한 적용구역을 확인하였고, 적용구역의 범위를 확인하였을 때 상당히 다양한 물질에 대한 예측범위를 가지고 있으나, 다중 음이온기 혹은 방향족 메탄이민기 계열의 분자에 대해서는 그 예측 신뢰성이 감소함을 확인하였다. 이에 향후 다양한 유기화합물에 대한 비등점의 예측에 대해 본 예측모델을 사용할 경우, 적절한 적용범위를 설정하고 이에 대한 신뢰성 있는 결과를 도출할 경우 QSPR 기반의 비시험적인 대량의 비등점 예측정보를 확보하고 이를 다양한 용도에 활용할 수 있을 것으로 예상된다.

감사의 글

본 연구는 민・군겸용기술과제(과제번호 13-SF-EB-03-mke)의 지원으로 수행되었으며, 이에 감사 드립니다.

References

1. B. Moller, J. Rarey and D. Ramjugernath, *J. Mol. Liq.*, **143**(1), 52-63 (2008).
2. R. Ceriani, R. Gani and A. J. A. Meirelles, *Fluid Phase Equilib.*, **283**(1), 49-55 (2009).
3. J. Marrero and R. Gani, *Fluid Phase Equilib.*, **183**, 183-208 (2001).
4. Y. S. Park, J. H. Lee, H W. Park and S. K. Lee, *Anal. Sci. Tech.*, **28**(3), 187-195 (2015).
5. Physical/Chemical Property Database (PHYSPROP), SRC Environmental Science Center, Syracuse Research Corporation, Syracuse, New York, 1994-2015.
6. H. Bathelt, F. Volk, and M. Weindel, The ICT-Database of Thermochemical Values, 8th update, Fraunhofer-Institut für Chemische Technologie (ICT), Pfinztal, Germany, 2008.
7. DIPPR(Design Institute for Physical Property Data), American Institute of Chemical Engineers, New Mexico, 2015.
8. PreADMET Ver.2.0.2.0, BMDRC, Seoul, Korea, 2007.
9. RapidMiner Ver.5.3.015, Rapid-I, Stockumer, Germany, 2014.
10. A. Tropsha, P. Gramatica and V. K. Gombar, *QSAR Combi. Sci.*, **22**(1), 69-77 (2003).