Korean Chem. Eng. Res., Vol. 48, No. 6, December, 2010, pp. 717-724

유기물의 인화점 예측을 위한 부분최소자승법과 SVM의 비교

이창준 · 고재욱* · 이기백**,†

서울대학교 화학생물공학부
151-742 서울시 신림동 산56-1
*광운대학교 화학공학과
139-701 서울시 노원구 월계동 447-1
**충주대학교 화공생물공학과
380-702 충북 충주시 대학로 50
(2010년 7월 28일 접수, 2010년 9월 1일 채택)

Comparison of Partial Least Squares and Support Vector Machine for the Flash Point Prediction of Organic Compounds

Chang Jun Lee, Jae Wook Ko* and Gibaek Lee,†**

Department of Chemical and Biological Engineering, Seoul National University, San 56-1, Shilim-dong, Gwanak-gu, Seoul 151-742, Korea
**Department of Chemical Engineering, Kwangwoon University, 447-1 Wolgye-dong, Nowon-gu, Seoul 139-701, Korea*
***Department of Chemical and Biological Engineering, Chungju National University, 50 Daehak-ro, Chungju-si, Chungbuk 380-702, Korea*
(Received 28 July 2010; accepted 1 September 2010)

요 약

액체의 화재 및 폭발위험을 나타내는 가장 중요한 물성의 하나인 인화점의 실험 데이터는 그 필요에도 불구하고 실제로 데이터를 확보하는 것이 가능하지 않은 경우가 많다. 이 연구에서는 DIPPR 801에서 얻은 893개 유기물의 인화점 실험데이터로부터 인화점을 예측하는 부분최소자승법(PLS) 및 support vector machine(SVM) 모델을 만들고 비교하였다. 분자를 구성하는 각 구성요소들이 분자의 물성에 일정한 기여를 한다는 가정을 이용하여 분자의 물성을 예측하는 방법인 그룹기여법을 이용하여 65개 작용기가 이 예측모델의 독립변수가 되었고 분자량의 로그값이 추가되었다. 두 모델에서 결정해야 할 매개변수는 교차검증에서 계산된 오차를 이용하여 결정되었는데, SVM모델은 그 매개변수가 많아 particle swarm optimization을 이용한 최적화를 이용하였다. 훈련데이터의 선택이 예측성능에 영향을 줄 수 있어 임의로 100개의 데이터 세트를 생성하여 테스트하였다. 전체 데이터에 대해 계산된 평균절대오차는 PLS가 13.86~14.55였고, SVM이 7.44~10.26여서 SVM이 PLS에 비해 매우 우수한 예측성능을 보였다.

Abstract – The flash point is one of the most important physical properties used to determine the potential for fire and explosion hazards of flammable liquids. Despite the needs of the experimental flash point data for the design and construction of chemical plants, there is often a significant gap between the demands for the data and their availability. This study have built and compared two models of partial least squares(PLS) and support vector machine(SVM) to predict the experimental flash points of 893 organic compounds out of DIPPR 801. As the independent variables of the models, 65 functional groups were chosen based on the group contribution method that was oriented from the assumption that each fragment of a molecule contributes a certain amount to the value of its physical property, and the logarithm of molecular weight was added. The prediction errors calculated from cross-validation were employed to determine the optimal parameters of two models. And, an optimization technique should be used to get three parameters of SVM model. This work adopted particle swarm optimization that is one of heuristic optimization methods. As the selection of training data can affect the prediction performance, 100 data sets of randomly selected data were generated and tested. The PLS and SVM results of the average absolute errors for the whole data range from 13.86 K to 14.55 K and 7.44 K to 10.26 K, respectively, indicating that the predictive ability of the SVM is much superior than PLS.

Key words: Flash Point, Property Estimation, Group Contribution Methods, Partial Least Squares, Support Vector Machine, Particle Swarm Optimization

†To whom correspondence should be addressed.
E-mail: glee@cjnu.ac.kr

1. 서 론

인화점(flash point, FP)은 액체의 화재 및 폭발위험을 나타내는 가장 중요한 물성의 하나이며 최근 안전에 대한 관심과 함께 그 중요성이 커지고 있다[1]. 인화점은 액체의 표면에서 발생한 증기가 공기와 섞여서 가연성 혼합기체를 형성하고 여기에 불꽃을 가까이 댔을 때 순간적으로 섬광을 내면서 연소하는, 즉 인화되는 최저의 온도를 말한다. 인화점을 측정하는데 가장 일반적으로 사용되는 방법은 개방식 측정법(open cup)과 밀폐식 측정법(closed cup)이 있다. 방법에 상관없이 그 오차는 대략 5~8 ºC이다. 개방된 컵에 액체를 넣어 측정하는 개방식 측정법은 액체와 대기 사이의 물질전달속도를 알 수 없으므로 오차가 생기게 되지만 실제 상황과 유사한 방법이라 할 수 있다. 이에 비해 뚜껑이 있는 컵을 사용하는 밀폐식 측정법에서는 일관된 결과를 얻을 수 있다. 일반적으로 인화점이 보통인 경우에는 밀폐식 측정법을, 상대적으로 인화점이 높은 경우에는 개방식 측정법을 사용한다. 또한, 밀폐식 측정법으로 측정된 인화점은 개방식 측정법보다 낮게 나타난다[2].

많은 물질에 대한 인화점 측정 실험이 요구되고 있으나 실제로 데이터를 확보하는 것이 가능하지 않은 경우가 많다. 특히 유독성, 폭발성, 방사성 물질의 경우에는 그 실험이 매우 어려워진다. 산업적으로 많이 쓰이는 물질의 경우에도 인화점에 대한 실험 데이터를 얻을 수 없는 경우가 있어서 2천만개 이상의 알려진 물질 중에서 단지 수천 개의 물질에 대한 인화점만 찾을 수 있다. 또한, 값이 알려진 경우에도 그 값이 실험에 의한 것인지 예측에 의한 것인지 불확실한 경우도 많다. 따라서 화학공장의 안전한 설계 및 건설을 위해 유기물의 인화점을 예측하는 신뢰성 있는 방법이 요구된다.

유기물의 인화점 예측법에 대한 많은 연구가 발표되었으며, 2004년에 Vidal 등은 인화점과 연소한계의 예측법에 대한 리뷰논문을 발표하였다[3]. 특히, 물질구조에 대한 분자표현자(molecular descriptor)와 물성을 관련짓는 정량적 구조-특성 관계(quantitative structure-property relationships, QSPR)를 이용하는 방법이 많이 발표되었다. QSPR을 이용하여 인화점을 예측하는 첫 번째 연구에서는 2개의 분자표현자를 이용하여 유기물 400개의 인화점을 예측하였으며, 이 연구에서 400개 성분에 대해 얻어진 평균절대오차(average absolute error, AAE)는 10.3 K이었다[4]. Tetteh 등은 작용기(functional group) 25개와 분자연결지수(molecular connectivity index)를 입력으로 사용한 RBF(radial basis function) 인공신경망을 이용하여 400개 성분에 대해 인하점을 예측하였다[5]. Katritzky 등은 271개 성분에 대해 3개의 분자표현자로 표현된 선형 회귀식을 제시한 바 있다[1]. 이 연구팀에서는 2007년 4개의 분자표현자를 입력변수로 사용한 다중회귀분석(multiple linear regression)과 인공신경망으로 758개 성분의 예측법을 제안하여, 테스트 성분 158개에 대해 각각 13.9 K과 12.6 K이라는 AAE를 얻었다[6]. Gharagheizi와 Alamdari가 유전자 알고리즘(generic algorithm)에 기반한 다중회귀분석법을 이용하여 824개 훈련데이터에서 얻은 모델로 206개 테스트 성분에 대해 계산한 AAE는 10.2 K였다[7]. 발표된 연구 중 최저의 AAE를 보인 연구는 Pan 등이 57개 작용기를 입력으로 사용한 RBF SVM을 이용하여 1026개 훈련 데이터에서 모델을 만들고 256개 테스트 데이터에서 얻은 9.99 K이다[8]. 2009년에는 Patel 등이 16개의 분자표현자와 신경망을 이용하여 236개 용매에 대해 20.44 K의 AAE를 얻었다[9]. Table 1은 이 연구들을 요약한 것으로, 표에서 사용된 AAE, R^2(결정계수), RMS(제곱오차)는 다음 식과 같이 정의된다.

$$\mathrm{AAE} = \frac{\sum_{i=1}^{n} |y_p - y_e|}{n} \tag{1}$$

$$R^2 = \frac{\left(n\sum_{i=1}^{n} y_p y_e - \sum_{i=1}^{n} y_p \cdot \sum_{i=1}^{n} y_e\right)^2}{\left[n\sum_{i=1}^{n} y_p^2 - \left(\sum_{i=1}^{n} y_p\right)^2\right]\left[n\sum_{i=1}^{n} y_e^2 - \left(\sum_{i=1}^{n} y_e\right)^2\right]} \tag{2}$$

$$\mathrm{RMS} = \sqrt{\frac{\sum_{i=1}^{n} (y_p - y_e)^2}{n}} \tag{3}$$

여기서 n은 자료의 수(여기서는 유기물의 수), y_p는 예측치, y_e는 실험치이다.

그러나 2배 이상인 AAE의 차이에도 불구하고 각 연구에서 사용된 성분들이 다르고 입력변수가 다르므로 어떤 통계처리방법이 우월하다고 말하기는 어렵다. 또한, 훈련 데이터와 테스트 데이터가 구분되어 있어 훈련데이터에 따라 결과가 달라질 수 있다는 것도 이 연구들의 결과를 일반화하기 어렵게 하고 있다. 참고로 2007년 Katritzy 등의 연구부터 훈련 데이터와 테스트 데이터의 비율을 80:20으로 하고 있다.

이 연구에서는 DIPPR 801(2009년 버전2)에서 얻은 유기물에 대해 작용기와 분자량을 입력변수로 인화점을 예측하는 부분최소자승법(PLS, partial least squares)과 SVM(support vector machine) 모델을 각각 만들고 그 결과를 비교하였다. 또한 SVM의 최적 계수들을 얻기 위해 PSO(particle swarm optimization)를 이용한 최적화를 적용하였다. 2장에서는 사용된 방법론에 대해 소개하고 3장에서 두 방법에서 얻어진 결과를 비교하였다.

Table 1. Previous works and their results

Works	Inputs	Chemometrics method	AAE (train/test)	R^2 (train/test)	RMS (train/test)	No. of whole data	Ratio of training data
Tetteh *et al*.(1999)	molecular connectivity index and 25 functional groups	neural network	7.1/11.2	0.96/0.92	10.1/14.0	400	33%
Katritzky *et al*.(2001)	3 molecular descriptors	multi-parameter regression	-	0.95	12.2	271	100%
Katritzky *et al*.(2007)	4 molecular descriptors	neural network	-/12.6	0.88/0.98	-	758	79%
Gharagheizi *et al*.(2008)	4 molecular descriptors	GA-MLR	-/10.2	0.97/0.97	12.0/12.7	1030	80%
Pan *et al*.(2008)	57 functional groups	SVM	6.12/9.99	0.98/0.95	9.95/15.81	1282	80%
Patel *et al*.(2009)	16 molecular descriptors	neural network	20.44(whole)	0.90/0.66	-	236(solvents)	80%

2. 방법론

2-1. 그룹기여법(Group contribution method)

QSPR을 이용한 예측법에서는 분자표현자의 선택이 가장 중요하다. 분자표현자는 끓는 점, 녹는점, 밀도, 임계물성 등의 물성, 위상적 표현자(topological descriptor), 기하학적 표현자(geometrical descriptor), 작용기 등으로 구분된다[10]. 이 중 작용기를 이용한 물성 예측법은 그룹기여법으로 알려져 있는 방법으로 다양한 물성 예측을 위해 많이 사용되어 왔다[11-15]. 그룹기여법은 분자를 구성하는 각 구성요소들이 분자의 물성에 일정한 기여를 한다는 가정을 이용하여 분자의 물성을 예측하는 방법이다. 가장 기본적인 구성요소는 탄소, 수소, 산소 등의 원자와 단일, 이중, 삼중의 화학결합이며, 이보다 복잡한 것은 원자와 화학결합으로 구성된 작용기이다. 그룹기여법은 수백만 개 분자의 물성을 예측하기 위해 기껏해야 수백 개에 불과한 작용기를 이용하기 때문에 필요한 정보의 양이 크게 줄어든다는 장점이 있다. 그러나 작용기가 지나치게 단순화되거나[11], 물성에 대한 데이터베이스가 충분하지 않을 때 이 방법에 의해 예측된 물성은 큰 오차를 나타낼 수 있다[15]. 따라서 신뢰할 수 있으면서도 충분한 양의 인화점 데이터를 확보하고 적절한 작용기를 사용하는 것은 정확한 인화점 예측을 위해 필수적이라 할 수 있다.

인화점에 대한 많은 자료를 많은 논문, 핸드북, 데이터베이스 등에서 찾을 수 있으나 이 연구에서는 미국화학공학회(AIChE)에서 추천되어 최근 활발히 이용되고 있는 물성 데이터베이스인 DIPPR 801 (2009년 버전2)의 인화점 데이터를 이용하였다[16]. 이 데이터베이스의 1,973개 물질 중 유기물은 1,765개이며, 인화점 데이터가 있는 1,629개 유기물 중 인화점이 실험에 의해 결정된 것은 893개이다.

또한, 893개 유기물의 작용기를 분석하고 Lee 등의 55개 작용기와 Pan 등의 57개 작용기를 참조하여 Table 2와 같은 65개의 작용기를 독립변수로 선택하였다[15,8]. 작용기는 ending group 18개, middle group 24개, aliphatic ring group 13개, aromatic ring group 10개로 구분된다. 또한, 분자량은 Lee 등의 연구에서 사용된 것으로, 이 연구에서는 분자량이 인화점 예측에 미치는 영향을 정확히 하고자 하였다. Fig. 1(a)는 893개 유기물의 분자량과 인화점을 나타내고 있는데 선형이라고 보긴 어렵다(결정계수=0.3497). 몇 가지 비선형함수

Table 2. The functional groups of the model

No.	Group	No.	Group	No.	Group	No.	Group
1(E1)	-CH3	18(E18)	-H	35(M17)	-Al-	52(R10)	O
2(E2)	=CH2	19(M1)	>C<	36(M18)	-B-	53(R11)	=C
3(E3)	≡∏CH	20(M2)	>C=	37(M19)	>C<(-X)*	54(R12)	S
4(E4)	≡N	21(M3)	=C=	38(M20)	>C=(-X)*	55(R13)	>Si<
5(E5)	-NH2	22(M4)	-C≡	39(M21)	-CH2-(-X)*	56(A1)	=CH-
6(E6)	-NO2	23(M5)	-CH2-	40(M22)	>CH-(-X)*	57(A2)	=C<
7(E7)	-SH	24(M6)	>CH-	41(M23)	-CH=(-X)*	58(A3)	=C<(-X)*
8(E8)	-Br	25(M7)	-CH=	42(M24)	>Si<	59(A4)	>N-
9(E9)	-F	26(M8)	>N-	43(R1)	-CH2-	60(A5)	NH
10(E10)	-Cl	27(M9)	-N=	44(R2)	=CH-	61(A6)	O
11(E11)	-I	28(M10)	-NH-	45(R3)	>CH-	62(A7)	S
12(E12)	-COH	29(M11)	-O-	46(R4)	>C<	63(A8)	o-B, m-B, p-B
13(E13)	-COOH	30(M12)	-S-	47(R5)	=C<	64(A9)	3-branch benzene**
14(E14)	=O	31(M13)	-CO-	48(R6)	-N<	65(A10)	4-branch benzene***
15(E15)	-OH(alcohol)	32(M14)	-CO2-	49(R7)	NH	66	Logarithm of molecular weight
16(E16)	-OH(phenol)	33(M15)	-SO2-	50(R8)	=N-		
17(E17)	=S	34(M16)	-SO-	51(R9)	CO		

*-X: attached to halogen atoms, **3-branch benzene: (1,2,3), (1,2,4), or (1,3,5), ***4-branch benzene: (1,2,3,4), (1,2,3,5), or (1,2,4,5)

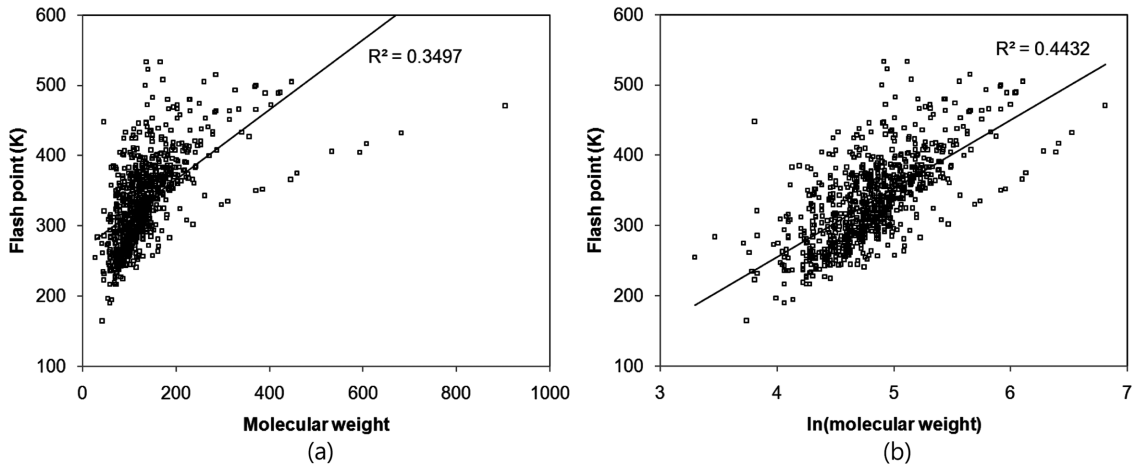

Fig. 1. Relationships of molecular weight and flash point.

로 계산해 본 결과 분자량의 로그가 가장 적합하다는 것을 확인하고 (Fig. 1(b), 결정계수=0.4432) 이를 예측모델의 독립변수로 추가하여 독립변수는 모두 66개가 되었다.

2-2. 부분최소자승법

독립변수 X와 종속변수 Y의 두 자료행렬 사이의 선형관계를 찾는데 주로 사용된 방법은 다중선형회귀법(multiple linear regression)이다[17]. 그러나 각 행렬 내에 서로 상관이 많은 변수들이 포함된 경우 그 예측에 오류가 많아지게 되는데 이를 공선형 문제(collinearity problem)라 한다. 부분최소자승법(PLS)은 이 문제를 효과적으로 처리하여 예측력이 뛰어나 최근 넓은 분야에서 많이 사용되고 있다.

PLS는 X와 Y 각각에 대해 주성분분석(principal component analysis, PCA)을 이용하여 주성분(principal components, PC)를 구한 후 X와 Y 각각의 주성분을 다시 선형으로 관련짓는다[17]. PCA는 자료의 중요한 변동을 나타낼 수 있도록 원래 변수의 선형결합으로 표현되는 새로운 변수인 주성분을 찾는 방법이다. PCA는 다음 식과 같이 데이터행렬 X를 T(score matrix)와 P(loading matrix)로 분해하게 된다. 여기서, k는 PC의 수, p_i는 원래 변수와 PC사이의 선형관계를 나타내는 계수이며 t_i는 변환된 PC를 나타낸다(t_i=Xp_i).

$$X = \sum_{i=1}^{k} t_i p_i^T + E = TP^T + E \tag{4}$$

PLS에서는 종속변수 Y에 대해서도 PCA를 적용한 후 X와 Y의 관계를 식 (6)과 같이 선형식으로 나타낸다.

$$Y = UQ^T + F \tag{5}$$

$$U = TB \tag{6}$$

PLS모델의 계수를 계산하는 가장 일반적인 방법은 NIPALS(Nonlinear Iterative Partial Least Squares) 알고리즘으로 Y에 대한 예측능력을 최대화하면서도 X의 중요 변동을 나타내도록 한다. 독립변수 X에서 종속변수 Y를 예측하는 식은 다음 식으로 표현되는데 이 식에서 BPLS는 PLS모델의 계수를 나타내며, W는 NIPALS알고리즘에서 정의된 가중치이다.

$$Y = XB_{PLS} = XW(P^TW)^{-1}BQ^T \tag{7}$$

일단 NIPALS알고리즘에서 PLS모델을 얻으면 최종적으로 PC수를 결정해야 한다. PC가 많아지게 되면 주어진 데이터를 잘 나타낼 수는 있으나 과적합(overfitting)으로 인해 추정오차가 커질 수 있으므로 통계적으로 의미있는 PC수를 찾아야 한다. 교차검증(cross-validation)은 통계모델 검증에 가장 널리 사용되는 방법으로, 데이터를 여러 개의 그룹들(이 연구에서는 10개)로 나눈 후에 그룹들 중에서 하나를 제외시킨 데이터 세트를 각각 만든다. 이 데이터 세트들로 각각 PLS 모델을 만든 후에 각 모델을 만들 때 제외된 그룹이 각 모델의 검증 그룹이 되어 이 검증 그룹에 대한 Y 변수의 실제 값과 예측값의 차이를 계산한다. 모든 데이터 세트에서 얻어진 오차를 합산하여 그 모델의 예측력을 나타내게 되는데, 이 연구에서는 이를 최소화시키는 PC수를 사용하였다. 또한, 데이터 세트를 나누는 방법이 결과에 영향을 미칠 수 있으므로 각각 10번씩 임의로 데이터 세트를 만들도록 하여 그 결과를 합하도록 하였다.

2-3. SVM

대표적 선형 예측법인 PLS 외에도 이 연구에서는 최근 분류 및 예측 분야에서 널리 사용되고 있는 SVM을 이용하였다. SVM은 그룹기여법과 같이 주어진 데이터가 희소하고 차원이 큰 문제를 잘 처리한다는 장점이 있다[19].

SVM은 원래 분류문제를 위해 개발된 방법으로, 하나의 집단과 다른 집단을 분류하는 최적의 분리경계면을 찾는다. 최적의 경계면은 분류할 두 집단으로부터 가장 멀리 떨어진 초평면(hyper-plane)으로 정의되며, 경계면에 가장 가까이 있는 데이터를 support vector라 한다[19]. 손실함수(loss function)를 이용하면 SVM을 회귀에 적용할 수 있게 되는데 이를 SVR(support vector regression)이라고도 부른다. SVR의 목적은 모든 데이터에서의 거리를 최소로 하는 초평면을 찾는 것이다[8]. n개의 데이터(x_i, y_i)가 주어진 선형 SVR문제는 y를 예측하는 최적의 초평면, f(x)-ω·x+b를 찾는 문제가 된다. 여기서 초평면에서 각 데이터 사이의 거리는 ε보다 작아지게 하는 ε-insensitive loss function을 사용하면, 이 문제는 $y_i-\omega x-b\le\varepsilon$와 $\omega x+b-y_i\le\varepsilon$의 제약조건에서 $1/2\|\omega\|^2$을 최소화시키는 문제와 같다.

슬랙 변수(slack variable) ξ와 ξ*를 포함시켜 예측오차를 고려하면 풀어야 할 최적화문제는 다음과 같이 쓸 수 있다.

$$\min\ J(\omega,\xi_i,\xi_i^*) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\xi_i+\xi_i^*) \tag{8}$$

$$\text{subject to } y_i - wx - b \le \varepsilon + \xi_i$$

$$wx + b - y_i \le \varepsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \ge 0$$

여기서 C는 오분류와 성능 간의 균형(trade-off)을 조절하는 비용 변수로 C가 커지게 되면 훈련데이터를 과적합하게 되고, C의 값이 적다면 풀이가 복잡해진다.

식 (8)에서 얻어지는 최적 회귀함수는 다음 식과 같으며, 이 식에서 $\alpha_i\ge0$, $\alpha_i^*\le C$이다.

$$f(x) = (\omega x) + b = \sum_{i=1}^{n}(\alpha_i-\alpha_i^*)(xx_i) + b \tag{9}$$

비선형 SVR은 커널 함수(kernel function), $K(x,x_i)$를 이용하여 x를 고차공간으로 사상시켜 선형 SVR처럼 다루게 되는데 이때 식 9는 다음과 같이 바뀐다.

$$f(x) = [\omega\ x] + b = \sum_{i=1}^{n}(\alpha_i-\alpha_i^*)K(x,x_i) + b \tag{10}$$

이 연구에서는 $K(x,x_i)=\exp(-\gamma|x-x_i|^2)$의 RBF를 커널 함수로 사용하였다. SVR에서 사용자가 결정해야 할 매개변수는 비용변수 C, ε-insensitive loss function의 값 ε, RBF의 γ이다. 3개 매개변수의 최적값은 PSO를 이용한 최적화를 통해 계산되었으며, 최적화의 목적함수는 데이터를 10개의 그룹으로 나누는 교차검증을 사용하였다. SVM의 계산시간이 크기 때문에 leave-one-out 교차검증 등 다른 교차검증방법이나 임의의 데이터 세트에 의한 교차검증의 반복은 사용하지 않았다.

이 연구에서 사용된 SVM모델은 Chang과 Lin이 Matlab 라이브러리로 개발한 LibSVM 버전2.91로 계산되었는데, 연구자의 경험으로 보았을 때 이 라이브러리는 공개된 것 중 가장 빠른 속도를 가진 것

중 하나이다[20].

2-4. PSO

PSO는 유전자 알고리즘(generic algorithm, GA), simulated annealing (SA)과 같이 경험적 최적화기법이다. 경험적 최적화기법은 정확한 최적해를 찾지 못할 수도 있으나 최적해에 대한 훌륭한 근사해를 찾고자 하는 접근방식으로, 매개변수가 많은 문제에도 적용할 수 있고 초기값에 민감하지 않으며 목적함수의 미분값을 구하지 않아도 된다는 장점이 있다. 이런 장점에도 불구하고 경험적 최적화기법이 매우 많은 목적함수 계산횟수를 요구한다는 것은 잘 알려진 단점이다. SVM의 매개변수 최적화에서는 목적함수의 미분값을 얻는 것이 매우 어렵기 때문에 이 연구에서는 경험적 최적화기법을 채택하였다. 또한, SVM의 계산시간이 상대적으로 크기 때문에, GA나 SA에 비해 계산시간이 적다고 알려진 PSO를 SVM의 3개 계수를 결정하는 최적화문제에 적용하였다[21].

PSO는 원래 새 떼와 같은 동물 군집의 사회적 행동양식을 바탕으로 개발된 방법이다[21]. 군집(swarm)의 각 개체(particle)는 다차원 탐색공간을 옮겨 다니며 다른 개체들과 정보를 교환하게 되는데, 그들 자신과 이웃의 경험에 의한 정보를 이용하여 최적의 해로 이동해 간다. 이를 위해 개체는 이전에 경험했던 최적의 위치정보를 기억한다. 이를 식으로 나타내면 다음과 같다. 식 11의 첫 번째 부분은 개체의 과거 속도이고 두 번째, 세 번째 부분은 군집의 최적 위치 및 각 개체의 최적 위치와 개체의 현재 위치와의 거리를 통하여 입자의 새로운 속도를 계산한다. 계산된 속도를 바탕으로 식 12를 통해서 새로운 위치로 이동하게 된다.

$$v_{p,d}^{k+1} = w v_{p,d}^{k} + c_1 r_1 (x_{p,d}^{ind} - x_{p,d}^{k}) + c_2 r_2 (x_{d}^{glo} - x_{p,d}^{k}) \quad (11)$$

$$x_{p,d}^{k+1} = x_{p,d}^{k} + v_{p,d}^{k+1} \quad (12)$$

여기서 v는 개체의 속도, x는 입자의 위치이고, x^{ind}와 x^{glo}는 목적함수가 낮은 값을 가진 위치를 나타내는데 x^{ind}는 개체 자신이 찾은 최적의 위치, x^{glo}는 군집이 찾은 최적의 위치이다. 아래첨자인 p는 개체, d는 탐색방향(여기서는 SVM의 계수 3개), k는 반복횟수이다. r_1과 r_2는 [0, 1]의 범위에서 균등분포된 난수이다. 계수 w, c_1과 c_2는 PSO의 탐색 매개변수이며 w는 관성 가중치(inertia weight), c_1과 c_2는 각각 지식계수(cognition parameter)와 사회계수(social parameter)라 한다. w를 크게 하면 전역탐색에 많은 비중을 두게 되고 작게 하면 국부적인 탐색에 많은 비중을 두게 된다. c_1과 c_2는 각각 식 11에서 군집의 최적 위치와 각 개체의 최적 위치로 움직이게 하는 가중치이며, 이 연구에서는 다른 연구들과 같이 2를 사용하였다. PSO의 자세한 알고리즘은 Schwaab 등의 연구에서 찾을 수 있다[21].

이 연구에서는 개체를 20개, 반복횟수를 50으로 하여 SVM의 3개 매개변수를 결정하였다.

3. 결 과

893개 유기물의 인화점 데이터 중 80%인 714개를 훈련 데이터로 하여 PLS와 SVM 각각에 대해 인화점 예측모델을 구성하고 179개 데이터를 테스트 데이터로 하여 그 예측성능을 비교하였다. 훈련데이터를 임의로 선택하기 때문에 그 예측성능이 훈련데이터의 선택에 따라 달라지므로 이 연구에서는 각각 임의로 선택된 100개의 세트를 생성하여 테스트하였다.

Table 3은 그 결과를 요약한 것으로, 각각의 결과는 훈련 데이터와 테스트 데이터의 것을 따로 비교하였다. 표에서 average, min, max는 100개 데이터 세트에서 얻어진 평균값, 최소값, 최대값을 나타낸다. 테스트 데이터의 최대오차를 제외하면 AAE, R^2, RMS, 최대오차의 거의 모든 결과에서 SVM이 우수한 성능을 나타내었다.

Table 3을 기존 연구의 예측 결과인 Table 1과 비교하면 테스트 데이터의 최소 AAE 등 SVM에 의한 최고 성능은 대체적으로 기존 연구 결과보다 우수하거나 유사한 성능을 보였다. 그러나 이 연구의 예측성능과 기존 연구 결과의 비교는 각 연구에서 사용된 성분과 예측모델의 입력변수가 달라 공정하지 않다고 할 수 있다.

이 연구에 사용된 PC는 4GB 메모리, Intel Core2 Quad 2.4GHz CPU이었으며, PLS와 SVM의 계산에 사용된 총시간은 각각 3분 51초, 69시간 49분 50초였다. SVM 계산이 PLS보다 훨씬 큰 계산시간이 사용된 것은 PLS는 정해야 할 매개변수가 정수인 PC수이고, SVM은 매개변수가 3개이면서 그 값이 실수이기 때문이다. SVM의 계산시간이 매우 크지만 인화점 예측에서 사용할 통계모델이 정해진 후에는 893개 모든 실험 데이터를 사용하여 1번의 계산만 하므로 Table 3의 CPU time의 1/100이 필요하게 되므로 이 시간이 크다고 할 수는 없다. Table 4는 100개 데이터 세트에 대한 PLS와 SVM모델의 매개변수를 보이고 있다.

Table 4. Model parameters of PLS and SVM

Model	Parameter	Average	Min.	Max.
PLS	no. of PCs	14.47	8	25
SVM	C	48.80	6.58	99.02
	ε	0.0863	0.0049	0.1
	γ	0.0026	0.0008	0.0049

Table 3. Summary of flash point estimation results by PLS and SVM

Parameter	PLS			SVM			Rato(PLS/SVM)		
	Average	Min.	Max.	Average	Min.	Max.	Average	Min.	Max.
AAE(train)	13.83	12.56	14.86	7.47	5.60	9.43	1.85	2.24	1.58
AAE(test)	15.68	12.26	21.07	13.50	10.85	18.07	1.16	1.13	1.17
Max Error(train)	149.63	81.24	170.05	136.94	33.97	176.86	1.09	2.39	0.96
Max Error(test)	141.99	46.79	289.09	125.84	64.96	186.64	1.13	0.72	1.55
R^2(train)	0.891	0.875	0.917	0.962	0.937	0.984	0.93	0.93	0.93
R^2(test)	0.846	0.656	0.920	0.879	0.783	0.936	0.96	0.84	0.98
RMS(train)	19.81	17.34	21.16	11.56	7.64	15.28	1.71	2.27	1.38
RMS(test)	23.81	16.64	37.97	21.03	15.56	30.14	1.13	1.07	1.26
CPU time(s)		231			251,390			0.0009	

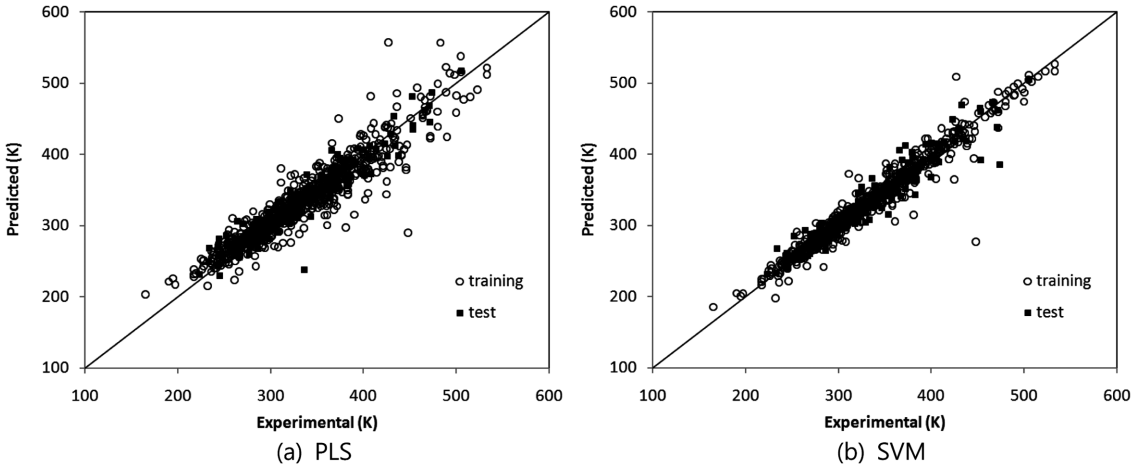

Fig. 2. Comparison between the predicted and experimental flash points.

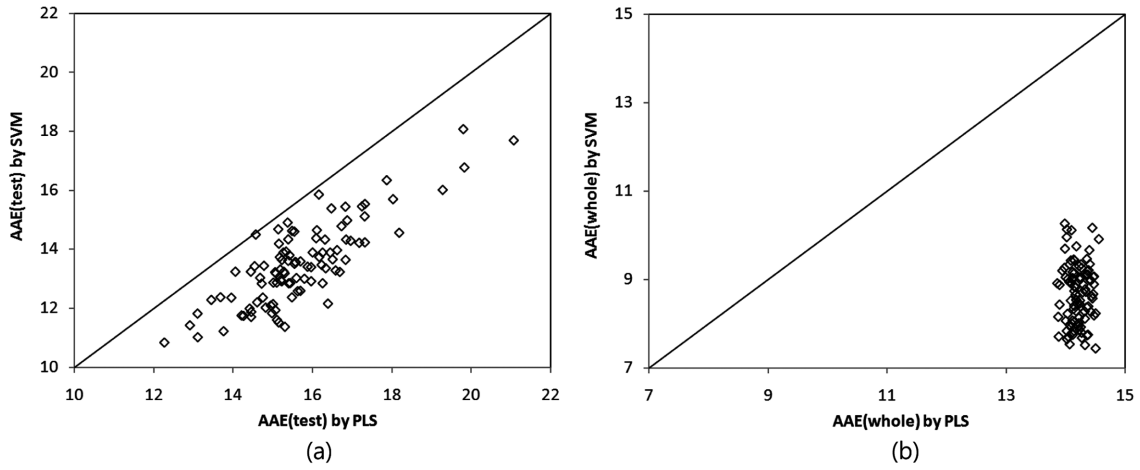

Fig. 3. Comparison of the AAEs calculated for test and whole data.

테스트 데이터에서 최저 AAE를 보인 데이터 세트에 대해 실험값과 예측값을 Fig. 2에 비교, 도시하였는데 이 그림에서도 SVM의 결과가 좋음을 쉽게 확인할 수 있다.

100개 데이터 세트에 대해 PLS와 SVM에 의한 테스트 데이터의 AAE를 비교해보면 Fig. 3(a)와 같은데, PLS에 의한 결과가 SVM보다 좋지만 몇 개 데이터 세트에서는 분명한 차이를 보여주지 못하고 있다. Fig. 3(b)는 100개 데이터 세트의 893개 데이터 전부에 대한 AAE를 계산하여 PLS와 SVM을 비교한 것으로 뚜렷한 성능 차이를 보여주고 있다. PLS에 의한 것은 평균 14.2, 최소 13.86, 최대 14.55 이었고, SVM은 평균 8.68, 최소 7.44, 최대 10.26이었다. PLS에 의한 최소 AAE가 SVM에 의한 최대 AAE보다 크다는 것은 데이터 세트에 무관하게 SVM이 PLS에 비해 우수하다는 것을 나타낸다.

전체 데이터의 평균절대오차가 7.44로 가장 작은 데이터 세트의 893개 데이터 중에서 SVM에 의해 얻어진 가장 큰 오차는 절대오차로 154K, 상대오차로 34.4%였다. 상대오차를 그 값에 따라 나누어 보면 Fig. 4와 같은데 893개 데이터 중 53%인 472개가 1% 이하의 오차를 나타내었고 2.7% 정도인 24개가 10%를 넘는 상대오차를 나타내어 평균 오차가 낮음에도 불구하고 일부 성분의 경우 오차가 매우 크다는 것을 보여준다.

66개 독립변수 중 분자량의 로그값이 예측성능에 미친 영향을 확인하기 위해 분자량의 로그값을 분자량으로 바꾼 66개 독립변수를

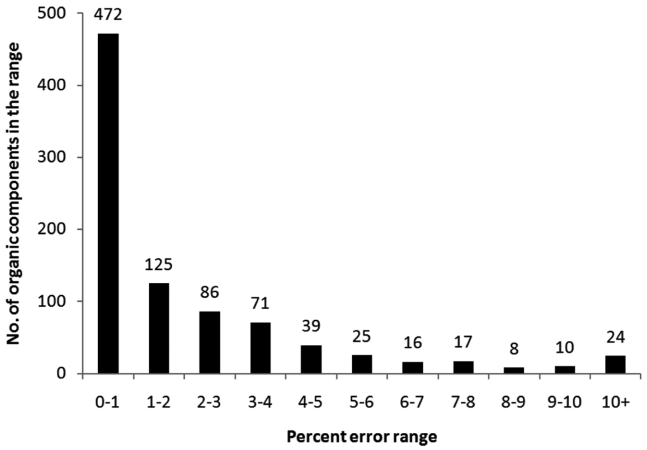

Fig. 4. The percent errors obtained by the SVM model and the number of organic compounds in each range.

이용하여 인화점을 추정하였다. Table 5는 그 결과를 요약한 것이며, 괄호 안의 숫자는 분자량의 로그값이 사용된 결과에 대한 비율을 %로 나타낸 것이다. 분자량의 로그값을 사용한 PLS는 분자량 자체를 사용한 PLS보다 대체적으로 좋은 결과를 나타내었으나, SVM의 경우에는 큰 차이가 얻어지지 않았다. 이것은 PLS가 선형 예측법이기

Table 5. Flash point estimation results after the logarithm of molecular weight is replaced with molecular weight

Parameter	PLS			SVM		
	Average	Min.	Max.	Average	Min.	Max.
AAE(train)	14.86(107%)	13.69(109%)	15.72(106%)	7.49(100%)	5.96(106%)	9.29(99%)
AAE(test)	16.78(107%)	12.74(104%)	22.17(105%)	13.55(100%)	10.91(101%)	17.41(96%)
Max Error(train)	138.65(93%)	78.44(97%)	150.88(89%)	131.93(96%)	35.64(105%)	167.02(94%)
Max Error(test)	138.05(97%)	64.84(139%)	241.99(84%)	127.99(102%)	63.91(98%)	178.80(96%)
R^2(train)	0.880(99%)	0.864(99%)	0.905(99%)	0.962(100%)	0.940(100%)	0.982(100%)
R^2(test)	0.837(99%)	0.703(107%)	0.913(99%)	0.877(100%)	0.794(101%)	0.935(100%)
RMS(train)	20.79(105%)	18.49(107%)	22.13(105%)	11.54(105%)	8.25(108%)	14.69(96%)
RMS(test)	24.58(103%)	16.98(102%)	38.05(100%)	21.23(101%)	15.53(100%)	30.61(102%)

때문에 분자량과 인화점의 관계를 비선형함수로 더 정확하게 나타냄으로써(Fig. 1) 예측 성능이 향상되었고, SVM은 비선형 예측법이기 때문에 독립변수를 비선형화해도 큰 차이를 만들지 못하기 때문이다.

4. 결 론

이 연구에서는 유기물의 인화점 실험데이터로부터 인화점을 예측하는 PLS와 SVM 모델을 만들고 비교하였다. 신뢰할 수 있는 인화점 예측 모델을 얻기 위해 DIPPR 801에서 얻은 893개 유기물의 인화점 실험데이터를 이용하였으며, 이 유기물의 작용기를 분석하여 예측모델의 독립변수로 65개의 작용기와 분자량의 로그값 등 66개를 사용하였다. 두 모델의 예측성능을 비교할 때 훈련 데이터의 선택이 영향을 줄 수 있어 임의로 100개의 데이터 세트를 생성하여 테스트하였다. 두 예측모델에 의한 결과를 비교하여 다음 결론을 얻었다.

(1) 테스트 데이터 및 훈련 데이터에 대해 얻어진 거의 모든 결과에서 SVM이 PLS에 비해 우수한 예측성능을 보였다. 이 결과를 통해 비선형 예측법인 SVM이 66개 독립변수와 인화점의 비선형 관계를 잘 나타낸다는 것을 확인하였다.

(2) 893개 데이터에 대한 상대오차를 그 범위별로 구하였을 때 10% 이상의 상대오차를 나타낸 데이터가 적지 않아 예측방법의 지속적인 개선이 요구되었다. 본 연구팀은 다양한 분자표현자 등 예측모델의 독립변수 조정 등을 통한 예측성능 향상을 시도하고 있다.

감 사

본 연구는 지식경제부의 에너지기술혁신 프로그램으로 지원되었으며 이 논문은 "차세대에너지안전연구단"의 연구 결과입니다(세부과제번호: 2007-M-CC23-P-02-1-000).

참고문헌

1. Katritzky, A. R., Petrukhin, R., Jain, R. and Karelson, M., "QSPR Analysis of Flash Points," *J. Chem. Inf. Comput. Sci.*, **41**(6), 1521-1530(2001).
2. Crowl, D. A. and Louvar, J. F., *Chemical Process Safety: Fundamentals with Applicatoins*, 2nd Ed., Prentice Hall, Upper Saddle River, NJ(2001).
3. Vidal, M., Rogers, W. J. Holste, J. C. and Mannan, M. S., "A Review of Estimation Methods for Flash Points and Flammability Limits," *Process Saf. Prog.*, **23**(1), 47-55(2004).
4. Suzuki, T., Ohtaguchi, K. and Koide, K., "A Method for Estimating Flash Points of Organic Compounds from Molecular Structures," *J. Chem. Eng. Jpn.*, **24**(2), 258-261(1991).
5. Tetteh, J., Suzuki, T., Metcalfe, E. and Howells, S., "Quantitative Structure-Property Relationships for the Estimation of Boiling Point and Flash Point Using a Radial Basis Function Neural Network," *J. Chem. Inf. Comput. Sci.*, **39**(3), 491-507(1999).
6. Katritzky, A. R., Stoyanova-Slavova, I. B., Dobchev, D. A. and Karelson, M., "QSPR Modeling of Flash Points: An Update," *J. Mol. Graph. Model.*, **26**(2), 529-536(2007).
7. Gharagheizi, F. and Alamdari, R. F., "Prediction of Flash Point Temperature of Pure Components Using a Quantitative Structure-Property Relationship Model," *QSAR Comb. Sci.*, **27**(8), 679-683(2008).
8. Pan, Y., Jiang, J., Wang, R., Cao, H. and Zhao, J., "Quantitative Structure-Property Relationship Studies for Predicting Flash Points of Organic Compounds using Support Vector Machines," *QSAR Comb. Sci.*, **27**(8), 1013-1019(2008).
9. Patel, S. J., Ng, D. and Mannan, M. S., "QSPR Flash Point Prediction of Solvents Using Topological Indices for Application in Computer Aided Molecular Design," *Ind. Eng. Chem. Res.*, **48**(15), 7378-7387(2009).
10. http://michem.disat.unimib.it/mole_db/
11. Constantinou, L. and Gani, R., "New Group Contribution Method for Estimating Properties of Pure Compounds," *AIChE Jr.*, **40**(10), 1697-1710(1994).
12. Wen, X. and Qiang, Y., "A New Group Contribution Method for Estimating Critical Properties of Organic Compounds," *Ind. Eng. Chem. Res.*, **40**(26), 6245-6250(2001).
13. Albahri, T. A., "Structural Group Contribution Method for Predicting the Octane Number of Pure Hydrocarbon Liquids," *Ind. Eng. Chem. Res.*, **42**(3), 657-662(2003).
14. Zbransk, Z. K. K. and Rika, V., "Estimation of the Heat Capacity of Organic Liquids as a Function of Temperature by a Three-Level Group Contribution Method," *Ind. Eng. Chem. Res.*, **47**(6), 2075-2085(2008).
15. Lee, C. J., Lee, G., So, W. and Yoon, E. S., "A New Estimation Algorithm of Physical Properties based on a Group Contribution and Support Vector Machine," *Korean J. Chem. Eng. (HWAHAK KONGHAK)*, **25**(3), 568-574(2008).
16. http://dippr.byu.edu/.
17. Lee, H. D., Lee., M. H., Cho, H. W., Han, C. and Chang, K. S., "Online Quality Monitoring Using Multivariate Statistical Methods in Continuous-stirred MMA-VA Copolymerization Process,"

HWAHAK KONGHAK, **35**(5), 605-612(1997).
18. Russell, E. L., Chiang, L. H. and Braatz, R. D., *Data-driven Techniques for Fault Detection and Diagnosis in Chemical Processes*, Springer-Verlag, London(2000).
19. Vapnik, V. N., *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, NY(1995).
20. http://www.csie.ntu.edu.tw/~cjlin/libsvm/.
21. Schwwab, M., Biscaia, E. C., Monteiro, J. L. and Pinto, J. C., "Nonlinear Parameter Estimation through Particle Swarm Optimization," *Chem. Eng. Sci.*, **63**(6), 1542-1552(2008).