KIGAS Vol. 16, No. 1, February, 2012
(Journal of the Korean Institute of Gas)

유기물의 자연발화점 예측을 위한 부분최소자승법과 SVM의 비교

†이기백
충주대학교 화공생물공학과
(2011년 12월 7일 투고, 2012년 2월 26일 수정, 2월 26일 채택)

Comparison of Partial Least Squares and Support Vector Machine for the Autoignition Temperature Prediction of Organic Compounds

†**Gibaek Lee**
Department of Chemical and Biological Engineering, Chungju National University
(Received December 7, 2011; Revised February 26, 2012; Accepted February 26, 2012)

요 약

화학물질의 화재위험을 나타내는 가장 중요한 물성의 하나인 자연발화점의 실험 데이터는 그 필요에도 불구하고 데이터를 얻는 것이 어려운 경우가 많다. 이 연구에서는 DIPPR 801에서 얻은 503개 유기물의 자연발화점 실험데이터로부터 자연발화점을 예측하는 부분최소자승법(PLS) 및 support vector machine(SVM) 모델을 만들고 비교하였다. 그룹기여법을 이용하여 59개 작용기가 이 예측모델의 독립변수가 되었다. 두 모델에서 결정해야 할 매개변수는 교차검증으로 계산된 오차를 이용하여 결정되었고, SVM모델은 그 매개변수가 많아 particle swarm optimization을 이용한 최적화를 이용하였다. 전체 데이터에 대해 계산된 평균절대오차는 PLS가 58.59K였고, SVM이 29.11K여서 SVM이 PLS에 비해 매우 우수한 예측성능을 보였다.

Abstract - The autoignition temperature is one of the most important physical properties used to determine the flammability characteristics of chemical substances. Despite the needs of the experimental autoignition temperature data for the design of chemical plants, it is not easy to get the data. This study have built and compared partial least squares (PLS) and support vector machine (SVM) models to predict the autoignition temperatures of 503 organic compounds out of DIPPR 801. As the independent variables of the models, 59 functional groups were chosen based on the group contribution method. The prediction errors calculated from cross-validation were employed to determine the optimal parameters of two models. And, particle swarm optimization was used to get three parameters of SVM model. The PLS and SVM results of the average absolute errors for the whole data range from 58.59K and 29.11K, respectively, indicating that the predictive ability of the SVM is much superior than PLS.

Key words : autoignition temperature, property estimation, group contribution methods, partial least squares, support vector machine, particle swarm optimization

Ⅰ. 서 론

자연발화점(autoignition temperature)은 인화점, 폭발상하한 등과 같이 액체의 화재위험을 나타내는 중요 물성의 하나이며, 외부 점화원 없이 공기 중에서 연소되는, 즉 인화되는 최저의 온도를 말한다[1]. 많은 물질에 대한 자연발화점 데이터가 요구되고 있으나 실제로 데이터를 확보하는 것은 매우 어렵다.

†주저자:glee@cjnu.ac.kr

특히 유독성, 폭발성 물질의 경우에는 실험이 더욱 어렵다. 산업적으로 많이 쓰이는 물질의 경우에도 자연발화점에 대한 실험 데이터를 얻을 수 없는 경우가 있어서 수천만 개 이상의 알려진 물질 중에서 단지 천 개 미만의 물질에 대한 데이터만 찾을 수 있다. 또한, 값이 알려진 경우에도 그 값이 실험에 의한 것인지 예측에 의한 것인지 불확실하고, 같은 물질에 대해 실험한 결과조차 실험자에 따라 평균 ±30K의 오차가 있다[1]. 따라서 화학공장의 안전한 설계 및 조업을 위해 신뢰할 수 있는 실험데이터를 이용하여 자연발화점을 예측하는 방법이 요구된다.

유기물의 자연발화점 예측법에 대한 몇 개 연구가 발표되었다. 물질구조에 대한 분자표현자와 물성을 관련짓는 정량적 구조-특성 관계(quantitative structure-property relationships, QSPR)를 이용한 Suzuki는 6개의 분자표현자를 이용하여 유기물 250개에 대해 33.6K의 평균절대오차(average absolute error, AAE)를 얻었다[1]. Tetteh 등도 6개의 분자표현자를 입력으로 사용한 인공신경망을 이용하여 248개 성분에 대해 자연발화점을 예측하였다[2]. Albari와 George는 작용기(functional group) 58개를 입력으로 사용한 인공신경망을 이용하여 DIPPR에서 얻은 490개 성분의 데이터에 대해 예측모델을 제시하였다[3]. 이들은 훈련 및 테스트 데이터에 대해 각각 17.8K, 16.7K의 좋은 결과를 얻었으나, 이 연구는 데이터 중 96%를 훈련용으로 사용하였다는 단점을 가지고 있다. 2008년에 Pan 등은 6개의 분자표현자로 90개 성분에 대한 예측모델을 SVM (support vector machine)으로 구축하였다[4]. 이 연구팀에는 2009년 3개의 데이터베이스로부터 얻은 데이터로부터 446개 성분에 대한 자연발화점 데이터를 정리하고, 이 데이터에 대한 예측모델로서 9개의 분자표현자를 입력으로 하는 다중회귀분석(multiple linear regression)과 SVM모델을 제안하였으며, 테스트 성분 90개에 대해 각각 32.7K와 28.88K라는 AAE를 얻었다[5].

Table 1은 이 연구들을 요약한 것으로, 표에서 사용된 AAE, RMS(제곱오차)는 다음 식과 같이 정의되며 R은 상관계수이다. 여기서 n은 자료의 수(여기서는 유기물의 수), y_p는 예측치, y_e는 실험치이다.

$$AAE = \frac{\sum_{i=1}^{n} |y_p - y_e|}{n} \quad (1)$$

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

그러나 AAE의 차이가 큼에도 불구하고 각 연구에서 사용된 성분과 그 실험값들이 다르고 입력변수가 다르므로 어떤 방법이 우월하다고 말하기는 어렵다.

이 연구에서는 신뢰할 수 있는 데이터베이스인 DIPPR 801에서 얻은 유기물에 대해 작용기를 입력변수로 자연발화점을 예측하는 부분최소자승법(PLS, partial least squares)과 SVM모델을 각각 만들고 그 결과를 비교하였다. 또한 SVM의 최적 계수들을 얻기 위해 PSO(particle swarm optimization)를 이용한 최적화를 적용하였다.

Ⅱ. 방법론

2.1. 그룹기여법(Group contribution method)

그룹기여법은 분자를 구성하는 각 구성요소들이 분자의 물성에 일정하게 기여한다는 가정을 이용하여 분자의 물성을 예측하는 방법이다. 가장 기본적인 구성요소는 탄소, 산소 등의 원자와 단일, 이중 등의 화학결합이며, 이보다 복잡한 것은 원자와 화학결합으로 구성된 작용기이다. 그룹기여법은 수백만 개 분자의 물성 예측을 위해 수백 개에 불과한 작용기를 이용하여 필요한 정보의 양이 크게 줄어든다는 장점이 있다. 그러나 작용기가 지나치게 단순화되거나[11], 물성에 대한 데이터베이스가 충분하지 않은 경우에는 예측된 물성이 큰 오차를 나타낼 수 있다[15]. 따라서 신뢰할 수 있으면서도 충분한 양의 자연발화점 데이터를 확보하고 적절한 작용기를 사용하는 것은 정확한 예측을 위해 필수적이다.

자연발화점에 대한 자료를 많은 데이터베이스, 논문, 핸드북 등에서 찾을 수 있으나 이 연구에서는 미국화학공학회에서 추천되어 가장 신뢰할만한 물성 데이터베이스인 DIPPR 801 (2009년 버전2)의 자연발화점 데이터를 이용하였다[8]. 이 데이터베이스의 1973개 물질 중 유기물은 1765개이며, 자연발화점 데이터가 있는 1141개 유기물 중 자연발화점이 실험에 의해 결정된 것은 503개이다.

Lee 등이 인화점 예측에 사용한 65개 작용기와 Albari와 George가 자연발화점 예측에 사용한 58개 작용기를 검토한 후 실험 데이터가 있는 503개 유기물의 작용기를 분석함으로써 Table 2와 같은 59개의 작용기를 독립변수로 선택하였다[9, 3]. 작용기는 ending group 17개, middle group 23개, aliphatic ring group 12개, aromatic ring group 7개로 구분된다.

2.2. 부분최소자승법

PLS는 독립변수 X와 종속변수 Y 각각에 대해 주성분분석(principal component analysis, PCA)으로

Table 1. Previous works and their results

Works	Inputs	Chemometrics method	AAE [K] (train/test)	R (train/test)	RMS [K] (train/test)	No. of whole data	Ratio of training data
Suzuki (1994)	6 molecular descriptors	multi-parameter regression	train 33.0/whole 33.6	0.951/-	-	250	40%
Tetteh et al. (1996)	6 molecular descriptors	neural network	17.1/30.1	0.976/0.913	-	248	34%
Albari et al.(2003)	58 functional groups	neural network	17.8/16.7	0.98/0.98	-	490	96%
Pan et al.(2008)	6 molecular descriptors	SVM	-	0.963/0.953	29.82/30.99	90	58%
Pan et al(2009)	9 molecular descriptors	multiple linear regression	30.98/32.70	0.932/0.925	37.99/36.86	446	80%
		SVM	27.55/28.88	0.949/0.935	33.21/36.86		

Table 2. The functional groups of the model

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

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

주성분(principal components, PC)을 구한 후 X와 Y 각각의 주성분을 다시 선형으로 관련짓는다[10]. 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($t_i = Xp_i$).

$$X = \sum_{i=1}^{k} t_i p_i^T + E = TP^T + E \quad (3)$$

PLS에서는 종속변수 *Y*에 대해서도 PCA를 적용한 후 *X*와 *Y*의 관계를 식 5과 같이 선형식으로 나타낸다.

$$Y = UQ^T + F \tag{4}$$

$$U = TB \tag{5}$$

PLS모델의 계수를 계산하는 가장 일반적인 방법은 NIPALS(Nonlinear Iterative Partial Least Squares) 알고리즘으로 *Y*에 대한 예측능력을 최대화하면서도 *X*의 중요 변동을 나타내도록 한다. 독립변수 *X*에서 종속변수 *Y*를 예측하는 식은 식 6으로 표현되는데 이 식에서 B_{PLS}는 PLS모델의 계수를 나타내며, *W*는 NIPALS알고리즘에서 정의된 가중치이다.

$$Y = XB_{PLS} = XW^T(P^TW)^{-1}BQ^T \tag{6}$$

NIPALS알고리즘에서 PLS모델을 얻으면 최종적으로 PC수를 결정해야 한다. PC가 많아지게 되면 주어진 데이터를 잘 나타낼 수는 있으나 과적합으로 인해 추정오차가 커질 수 있다. 따라서 통계적으로 의미있는 PC수를 찾기 위해 이 연구에서는 교차검증(cross-validation)을 사용하였다. 교차검증은 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.

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[11].

SVM은 원래 분류문제를 위해 개발된 방법으로, 하나의 집단과 다른 집단을 분류하는 최적의 분리경계면을 찾는다. 최적의 경계면은 분류할 두 집단으로부터 가장 멀리 떨어진 초평면으로 정의되며, 경계면에 가장 가까이 있는 데이터를 support vector라 한다[11]. 손실함수를 이용하면 SVM을 회귀에 적용할 수 있는데 이를 SVR(support vector regression)이라 부른다. SVR의 목적은 모든 데이터에서의 거리를 최소로 하는 초평면을 찾는 것이다. *n*개의 데이터 (x_i, y_i)가 주어진 선형 SVR문제는 *y*를 예측하는 최적의 초평면, $f(x) = \omega \cdot x + b$를 찾는 문제가 된다. 초평면에서 각 데이터 사이의 거리를 ε보다 작아지게 하는 ε-insensitive loss function을 사용하면, 이 문제는 $y_i - \omega x - b \le \epsilon$와 $\omega x + b - y_i \le \epsilon$의 제약조건에서 $\frac{1}{2}\|\omega\|^2$을 최소화시키는 문제가 된다.

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

$$\begin{aligned} min\ J(\omega, \xi_i, \xi_i^*) &= \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ subject\ to\ & y_i - wx - b \le \epsilon + \xi_i \\ & wx + b - y_i \le \epsilon + \xi_i^* \\ & \xi_i, \xi_i^* \ge 0 \end{aligned} \tag{7}$$

여기서 *C*는 오분류와 성능 간의 균형(trade-off)을 조절하는 비용 변수로 *C*가 커지게 되면 훈련데이터를 과적합하게 되고, *C*가 작아지면 풀이가 복잡해진다.

식 7에서 얻어지는 최적 회귀함수는 다음 식과 같으며, 이 식에서 $\alpha_i \ge 0,\ \alpha_i^* \le C$이다.

$$f(x) = (\omega x) + b = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)(xx_i) + b \tag{8}$$

비선형 SVR은 커널 함수(kernel function), $K(x, x_i)$를 이용하여 *x*를 고차공간으로 사상시켜 선형 SVR처럼 다루게 되는데 이때 식 8은 다음과 같이 바뀐다.

$$f(x) = [\omega\ x] + b = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(x, x_i) + b \tag{9}$$

이 연구에서는 $K(x, x_i) = exp(-\gamma|x - x_i|^2)$의 RBF (radial basis function)를 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.

이 연구에서 사용된 SVM모델은 Chang과 Lin이 Matlab 라이브러리로 개발한 LibSVM 버전2.91로 계산하였다[12].

2.4. PSO

PSO는 유전자 알고리즘(generic algorithm, GA), simulated annealing(SA)과 같은 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[13].

PSO는 원래 새 떼와 같은 동물 군집의 사회적 행동양식을 바탕으로 개발된 방법이다[13]. 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 10의 첫 번째 부분은 개체의 과거 속도이고 두 번째, 세 번째는 군집의 최적 위치 및 각 개체의 최적 위치와 개체의 현재 위치와의 거리이다. 식 10에서 입자의 새로운 속도를 계산하고, 계산된 속도를 바탕으로 식 11에서 새로운 위치로 이동하게 된다.

$$v_{p,d}^{k+1} = wv_{p,d}^{k} + c_1 r_1 (x_{p,d}^{ind} - x_{p,d}^{k}) + c_2 r_2 (x_{d}^{glo} - x_{p,d}^{k}) \quad (10)$$

$$x_{p,d}^{k+1} = x_{p,d}^{k} + v_{p,d}^{k+1} \quad (11)$$

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, c_1과 c_2는 각각 지식계수와 사회계수라 한다. w를 크게 하면 전역탐색에 많은 비중을 두게 되고 작게 하면 국부적인 탐색에 많은 비중을 두게 된다. PSO의 자세한 알고리즘은 Schwaab 등의 연구에서 찾을 수 있다[13].

여기서는 c_1과 c_2를 2, 개체를 20개, 반복횟수를 50으로 하여 SVM의 3개 매개변수를 결정하였다.

Ⅲ. 결 과

503개 유기물의 자연발화점 데이터 중 80%인 402개를 훈련 데이터로 선택하여 PLS와 SVM 각각에 대해 자연발화점 예측모델을 구성하고 101개 데이터를 테스트 데이터로 하여 그 예측성능을 비교하였다.

이 연구에 사용된 PC는 4GB 메모리, Intel Core2 Quad 2.4GHz CPU이며, PLS와 SVM의 계산에 사용된 시간은 각각 1.7초, 9분 1초였다. SVM의 계산시간이 PLS보다 훨씬 큰 이유는 PLS의 매개변수는 정수인 PC수로 1개이지만 SVM은 매개변수가 3개로 더 많고 그 값이 실수이기 때문이다. SVM의 계산시간이 매우 크지만 자연발화점 예측에서 사용할 통계모델이 정해진 후에는 예측을 위한 계산만 하므로 계산시간이 많은 것은 문제되지 않는다.

PLS에서 PC의 수는 20으로 결정되었고, SVM의 매개변수는 각각 C=13.518, ε=0.217, γ =0.025이다. Table 3은 예측 결과를 요약한 것으로, 각각의 결과는 훈련 데이터, 테스트, 전체 데이터의 것을 따로 비교하였다. AAE, R, RMS, 최대오차의 모든 결과에서 SVM이 우수한 성능을 나타내었다. SVM의 AAE는 29.11K로서 실험데이터의 ±30K이내였다[1].

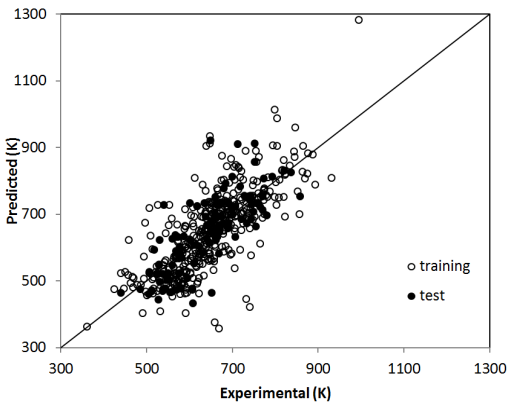
Table 3을 기존 연구의 예측 결과인 Table 1과 비

Table 3. Summary of autoignition temperature estimation results

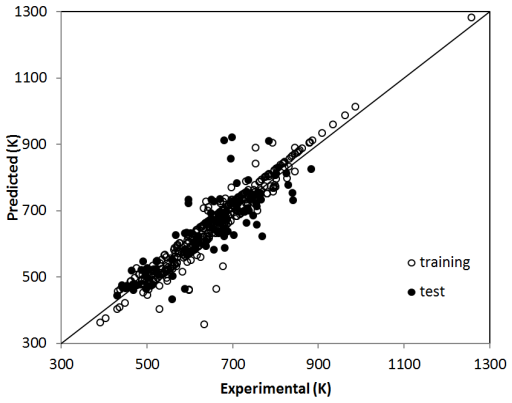
Parameter	PLS	SVM	Ratio (PLS/SVM)
AAE(train) [K]	59.34	25.80	2.30
AAE(test) [K]	55.63	42.26	1.32
AAE(whole) [K]	58.59	29.11	2.01
Max Error(train) [K]	319.98	275.14	1.16
Max Error(test) [K]	273.58	233.37	1.17
Max Error(whole) [K]	319.98	275.14	1.16
R(train)	0.751	0.961	0.78
R(test)	0.777	0.858	0.91
R(whole)	0.756	0.941	0.80
RMS(train) [K]	81.40	35.17	2.31
RMS(test) [K]	74.19	61.19	1.21
RMS(whole) [K]	80.00	41.72	1.92

Table 4. 10 compounds having the biggest absolute errors

Compound	Exp. Value [k]	Predicted Value [k]	Absolute Error [k]	Training or Test
Benzoyl Chloride	358.15	633.3	275.1	Training
Cyclopentadiene	913.15	679.8	233.4	Test
O-Dichlorobenzene	921	698.6	222.4	Test
5-Methyl-2-Hexanone	464.15	661.7	197.6	Training
Phthalic Anhydride	857	695.6	161.4	Test
Methylethanolamine	623.15	767.9	144.8	Test
1-Chloropentane	533	677.2	144.2	Training
2-Methyl-3-Ethylpentane	733.15	596.7	136.5	Test
Dimethyl Sulfate	461.15	597.5	136.4	Training
Aniline	890	753.7	136.3	Training



(a) PLS



(b) SVM

Fig. 1. Comparison between the predicted and experimental autoignition temperatures.

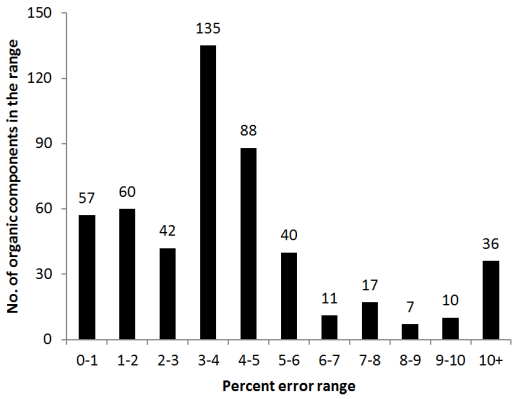


Fig. 2. The percent errors obtained by the SVM model and the number of organic compounds in each range.

교하였다. 거의 대부분의 데이터를 훈련용으로 사용한 Albari와 George의 연구를 제외하면 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 성분, 특히 실험값이 다르고 예측모델의 입력변수도 달라 공정하지 않다고 할 수 있다.

자연발화점의 실험값과 예측값을 Fig. 1에 비교, 도시하였는데 이 그림에서도 SVM의 결과가 좋음을 확인할 수 있다.

503개 데이터 중에서 SVM모델에 의한 예측값 중 가장 큰 오차를 보인 유기물은 Benzoyl Chloride로 절대오차로 275K, 상대오차로 76.8%였다. 상대오차를 그 값에 따라 나누어보면 Fig. 2와 같은데 503개 데이터 중 1% 이하의 오차를 나타낸 성분은 11%인

57개, 5% 이하인 성분이 76%인 382개였고 7.2% 정도인 36개가 10%를 넘는 상대오차를 나타내어 전체적으로 오차가 작지 않고 일부 성분의 경우 오차가 매우 크다는 것을 보여준다.

Table 4는 절대오차가 가장 큰 10개 성분의 결과를 보여주고 있는데 이 10개 성분이 평균 오차에서 10% 이상을 차지하고 있으므로 차후 이 성분들의 오차를 줄이는데 연구를 집중하여야 할 것이다.

Ⅳ. 결 론

유독성, 폭발성 등 물질의 특성 등 여러 이유로 자연발화점에 대한 실험을 하기 어려운 경우에 대해 자연발화점을 예측하는 방법이 요구된다. 이 연구에서는 유독성, 폭발성인 물질을 포함하여 실험 데이터를 얻기 어려운 유기물의 자연발화점 실험데이터로부터 자연발화점을 예측하는 PLS와 SVM 모델을 만들고 비교하였다. 두 모델에서 얻어진 결과를 비교하여 다음 결론을 얻었다.

(i) 모든 결과에서 SVM이 PLS에 비해 우수한 예측성능을 보였다. 이 결과를 통해 비선형 예측법인 SVM이 선형예측법인 PLS에 비해 59개 독립변수와 자연발화점의 관계를 더 잘 나타낸다는 것을 확인하였다.

(ii) SVM에 의한 AAE는 29.11K로 실험데이터의 30K와 유사하였다. 그러나 SVM모델로부터 얻은 상대오차를 그 범위별로 구하였을 때 상대오차가 큰 데이터가 적지 않아 예측방법의 지속적인 개선이 요구되었다.

감사의 글

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

참고문헌

[1] Suzuki, T., "Quantitative Structure- Property Relationships for Auto-ignition Temperatures of Organic Compounds", *Fire and Materials*, **18**, 81-88, (1994)

[2] Tetteh, J., E. Metcalfe and S.L. Howells, "Optimisation of Radial Basis and Backpropagation neural networks for modelling auto-ignition temperature by quantitative-structure property relation- ships", *Chemom. Intell. Lab. Syst.*, **32**, 177-191, (1996)

[3] Albahri, T.A. and R.S. George, "Artificial Neural Network Investigation of the Structural Group Contribution Method for Predicting Pure Components Auto Ignition Temperature", *Ind. Eng. Chem. Res.*, **42**, 5708-5714, (2003).

[4] Pan, Y., J. Jiang, R. Wang and H. Cao, "Advantages of Support Vector Machine in QSPR Studies for Predicting Auto- ignition Temperatures of Organic Compounds", *Chemom. Intell. Lab. Syst.*, **92**, 169-178, (2008)

[5] Pan, Y., J. Jiang, R. Wang, H. Cao and H. Cao, "Predicting the Auto-ignition temperatures of Organic Compounds from Molecular Structure Using Support Vector Machine", *J. Hazard. Mater.*, **164**, 1242-1249, (2009)

[6] Constantinou, L. and R. Gani, "New Group Contribution Method for Estimating Properties of Pure Comp- ounds," *AIChE Jr.*, **40**, 1697-1710, (1994)

[7] Lee, C.J., 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.*, **25**, 568-574 (2008).

[8] http://www.aiche.org/dippr/

[9] 이창준, 고재욱, 이기백, "유기물의 인화점 예측을 위한 부분최소자승법과 SVM의 비교", *화학공학*, **48**, 717-724, (2010)

[10] 이희두, 이무호, 조현우, 한종훈, 장근수, "다변량 통계 분석 방법을 이용한 연속 교반 MMA-VA 공중합 공정 품질 변수 온라인 모니터링", 화학공학, **35**, 605-612, (1997)

[11] Vapnik, V.N., *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, NY(1995)

[12] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[13] Schwwab, M., E.C. Biscaia, J.L. Monteiro and J.C. Pinto, "Nonlinear Parameter Estimation through Particle Swarm Optimization," *Chem. Eng. Sci.*, **63**, 1542-1552, (2008)