



(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2013년11월06일
(11) 등록번호 10-1325117
(24) 등록일자 2013년10월29일

(51) 국제특허분류(Int. Cl.)
G06F 19/00 (2011.01) G06F 17/10 (2006.01)
G06N 3/12 (2006.01)
(21) 출원번호 10-2011-0102080
(22) 출원일자 2011년10월06일
심사청구일자 2011년10월06일
(65) 공개번호 10-2012-0085172
(43) 공개일자 2012년07월31일
(56) 선행기술조사문헌
KR1020100042453 A
KR1020090832763 A

(73) 특허권자
주식회사 켄에쎈
서울특별시 금천구 가산디지털1로 196, 에이스테크노타워 11층 1106 (가산동)
(72) 발명자
전정재
경기도 의정부시 충의로57번길 7-20, 3층 (용현동)
권오형
서울특별시 서대문구 증가로26길 18 (북가좌동)
(뒷면에 계속)
(74) 대리인
이처영, 강상윤

전체 청구항 수 : 총 15 항

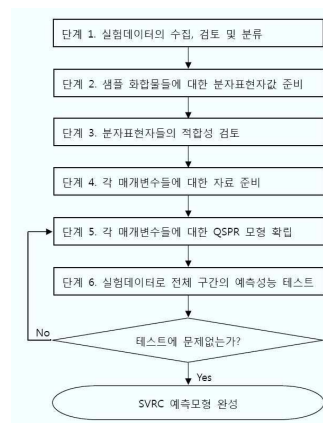
심사관 : 이수철

(54) 발명의 명칭 순수한 유기화합물의 액체의 열용량을 예측하는 SVRC 모형

(57) 요약

본 발명은 수소(H), 탄소(C), 질소(N), 산소(O), 황(S) 등 5가지 이내의 원소로 구성되고 수소를 제외한 원자의 개수가 25개 이하인 분자로 이루어진 순수한 화합물의 액체의 열용량(heat capacity of liquid)을 높은 정확도로 예측하는 수학적 모형을 제공한다. 상기의 모형은 SVRC(scaled variable reduced coordinate) 모형으로서, SVRC 수식을 통하여 각 온도에서의 액체의 열용량의 값을 알 수 있도록 하여 준다. 이 같은 수식계산에는 여러 매개변수들의 값이 필요한데, 이 값들은 각 화합물에 고유하게 주어지는 값들로 본 발명에서는 이들을 얻기 위해, 상기에 언급된 조건을 만족하는 다수의 화합물들의 액체의 열용량에 대한 실험값들을 바탕으로 다중선형회귀분석법과 인공신경망기법을 사용하여 각 매개변수들에 대한 QSPR(quantitative structure-property relationship) 예측 모형들을 확립하였다. 따라서 상기의 모형은, 모형에 포함된 분자표현자들의 구체적인 값만 알 수 있다면 그 어떤 분자든, 이 분자로 순수하게 이루어진 화합물의 액체의 열용량을 예측하여 준다. 이처럼, 본 발명은 실험값이 알려지지 않은 수많은 상기 조건의 화합물에 대해서도 신뢰할만한 액체의 열용량의 값을 예측할 수 있는 방법을 제공하여 줌으로써 실험에 드는 비용과 시간을 절약하게 해주어, 관련 산업의 연구개발활동을 용이하게 하는 등의 효과를 낳는다.

대표도 - 도1



(72) 발명자

권윤경

인천광역시 서구 가정3동 한신그랜드힐빌리지 39동 106호

김양수

서울특별시 영등포구 양산로 177, 108동 1401호 (영등포동7가, 경남아너스빌)

성애리

경기도 성남시 수정구 회망로466번길 23-13, 단대빌라 103호 (단대동)

정원천

서울특별시 강서구 화곡로15길 23-2, 503호 (화곡동)

조준혁

서울특별시 마포구 방울내로 31, 휴먼빌 101동 408호 (망원동)

박태윤

서울특별시 성북구 정릉로 305, 106동 1404호 (정릉동, 경남아파트)

특허청구의 범위

청구항 1

수집된 샘플 유기화합물들의 실험데이터를 입력하는 제1단계;

샘플 화합물들의 유기화합물의 액체의 열용량에 대한 분자표현자값을 준비하는 제2단계;

하기 식(1)에 기재되어 있는 SVRC 수식에 필요한 매개변수를 구하는 제3단계

$$\Theta(\epsilon) = \left(\frac{1 - A\epsilon^B}{1 - A} \right)$$

$$\alpha = \alpha_c$$

$$\epsilon = \left(\frac{T_b - T}{T_b - T_{0.55b}} \right)$$

$$C_{pL} = (C_{pL}(T_b)^\alpha - (C_{pL}(T_b)^\alpha - C_{pL}(T_{0.55b})^\alpha)\Theta)^{(1/\alpha)} \quad \text{식(1)}$$

[여기서 A, B는 각 물성에 고유하게 주어지는 보편상수들(universal constants)로서 액체의 열용량의 경우 그 값은 각각 1.33, 1이며, $C_{pL}(T_b)$ 은 정상끓는점에서의 액체의 열용량이고, $C_{pL}(T_{0.55b})$ 는 정상끓는점의 0.55배인 온도에서의($T_{0.55b}$) 액체의 열용량, α 는 스케일링 지수이며 α_c 는 임계점에서의 스케일링 지수];

상기 제3단계에서 얻은 상기 매개변수 정상끓는점과 정상끓는점의 0.55배인 온도에서의 액체의 열용량, α_c 에 대한 QSPR 모델을 구축하는 제4단계;

상기 실험 데이터로 예측성능을 테스트하는 제5단계; 및

상기 제5단계의 테스트를 만족하면 탐색된 모형에 의한 액체의 열용량 예측값을 액체의 열용량 값으로 채택하고, 만족하지 못하면 제4단계, 제5단계를 반복하는 제6단계를 포함하는 SVRC 모형에 의해서 유기화합물의 액체의 열용량을 구하는 방법.

청구항 2

제1항에 있어서, 상기 제4단계에서 상기 정상끓는점과 정상끓는점의 0.55배인 온도에서의 액체의 열용량에 대한 QSPR 모델을 구하는 방법은

상기 정상끓는점과 정상끓는점의 0.55배인 온도에서의 액체의 열용량, α_c 에 대한 최적의 분자표현자들을 추출하는 제4-0단계;

훈련집합과 시험집합으로 실험데이터를 분리하는 제4-1단계;

훈련집합에 대하여 최적의 다중선형회귀모형을 탐색하는 제4-2단계;

선택된 모형의 타당성을 검토하는 제4-3단계;

상기 제4-3단계에서 타당성이 없으면, 상기 제4-2단계, 제4-3단계를 반복하고, 타당성이 있으면 시험집합에 대하여 모형의 예측성능을 테스트하는 제4-4단계;

시험집합에 대한 상기 제4-4단계 테스트에서 성능이 기준을 만족하지 못하면 제4-2단계 내지 제4-4단계를 반복하고, 성능이 기준을 만족하면 샘플 표준화 후 3개 집합으로 분리하는 제4-5단계;

전체 샘플을 3개 집합으로 나눈 후에 최적의 인공신경망모형을 탐색하는 제4-6단계;

상기 제4-5단계에서 성능 테스트를 만족하는 상기 최적의 다중선형회귀모형에 의해서 구한 정상끓는점과 정상끓는점의 0.55배인 온도에서의 액체의 열용량 예측값과 상기 제4-6단계에서 탐색된 상기 최적의 인공신경망모형에

의해서 구한 정상끓는점과 정상끓는점의 0.55배인 온도에서의 액체의 열용량 예측값을 미리 설정된 과적합 방지 기준과 각각 비교하는 제4-7단계;

상기 제4-7단계에서 정상끓는점과 정상끓는점의 0.55배인 온도점에서의 액체의 열용량 각각의 경우, 인공신경망 모형에 의해서 얻어진 대응되는 온도점에서의 액체의 열용량 예측값이 각각의 경우에 해당하는 상기 과적합 방지 기준값의 85% 내지 115%를 벗어나면 상기 제4-5단계에서 얻은 다중선형회귀모형에 의한 대응되는 온도점에서의 액체의 열용량 예측값을 대응되는 온도에서의 액체의 열용량 값으로 채택하고 상기 과적합 방지 기준값의 85% 내지 115%에 속하면 상기 제4-6단계에서 탐색된 인공신경망모형에 의한 대응되는 온도에서의 액체의 열용량 예측값을 대응되는 온도에서의 액체의 열용량 값으로 채택하는 제4-8단계를 포함하고,

상기 제4단계에서 α_c 에 대한 QSPR 모형을 구하는 방법은

상기 제4-1단계 내지 제4-4단계와 시험집합에 대한 상기 제4-4단계 테스트에서 성능이 기준을 만족하지 못하면 제4-2단계 내지 제4-4단계를 반복하고, 성능이 기준을 만족하면 최적의 다중선형회귀모형으로 결정하는 제4-5-1단계;

상기 제4-5-1단계에서 성능 테스트를 만족하는 상기 다중선형회귀 모형을 통하여 α_c 값을 구하는 제4-6-1단계를 포함하는 SVRC 모형에 의해서 유기화합물의 액체의 열용량을 구하는 방법.

청구항 3

제2항에 있어서, 상기 제4-0단계에서 상기 정상끓는점과 정상끓는점의 0.55배인 온도에서의 액체의 열용량, α_c 에 대한 최적의 분자표현자는 모든 샘플 화합물에 대해 값이 동일하지 않은 독립적인 분자표현자인 것을 특징으로 하는 SVRC 모형에 의해서 유기화합물의 액체의 열용량을 구하는 방법.

청구항 4

제2항에 있어서, 상기 제4-1단계에서 훈련집합과 시험집합은 5:5 ~ 8:2의 비율로 나누는 것을 특징으로 하는 SVRC 모형에 의해서 유기화합물의 액체의 열용량을 구하는 방법.

청구항 5

제2항에 있어서, 상기 제4-2단계에서 상기 다중선형회귀모형은 상기 훈련집합에 대하여 유전적 알고리즘(genetic algorithm)을 적용하여 다중선형회귀모형을 탐색하는 것을 특징으로 하는 SVRC 모형에 의해서 유기화합물의 액체의 열용량을 구하는 방법.

청구항 6

제5항에 있어서, 상기 유전적 알고리즘(genetic algorithm)은 분자표현자들의 풀(pool)에서 일정한 개수의 분자표현자들을 무작위로 뽑아 만든 다수의 다중선형회귀모형들로 구성된 개체군(population)을 생성하는 단계; 각 개체(individual)들은 뽑힌 분자표현자들의 번호들을 조합하여 부호화하는 단계; 생성된 개체군으로부터 룰렛휠(Roulette Wheel) 방법에 의해서 두 개의 부모 염색체를 선택한 뒤 단순교배(single point crossover)법에 의하여 자손들을 생성하는 단계; 생성된 자손들의 염색체 일부를 일정 확률로 돌연변이(mutation) 시킨 뒤 기존 개체군의 일부를 이들로 교체하여 새 개체군을 생성하는 단계를 포함하는 것을 특징으로 하는 SVRC 모형에 의해서 유기화합물의 액체의 열용량을 구하는 방법.

청구항 7

제2항에 있어서, 상기 제4-2단계는 회귀모형의 결정계수 또는 평균절대오차에 의해서 예측성능을 판단하는 것을

포함하는 SVRC 모형에 의해서 유기화합물의 액체의 열용량을 구하는 방법.

청구항 8

제2항에 있어서, 상기 제4-3단계에서 타당성은 t 검정값에 의해서 타당성을 결정하는 SVRC 모형에 의해서 유기화합물의 액체의 열용량을 구하는 방법.

청구항 9

제2항에 있어서, 상기 제4-5단계에서 시험집합에 대한 예측성능이 훈련집합에 대한 예측성능과 유사하면 다중선형회귀모형이 결정되고, 시험집합에 대한 예측성능이 훈련집합에 대한 예측성능과 차이 나면 훈련집합과 시험집합을 다시 분류하는 SVRC 모형에 의해서 유기화합물의 액체의 열용량을 구하는 방법.

청구항 10

제2항에 있어서, 상기 제4-6단계에서 상기 인공신경망에 의한 탐색범위는 입력층과 출력층 사이에 하나의 은닉층을 가지며 전방향(feed forward)으로 만 연결되어 있는 것을 특징으로 하는 SVRC 모형에 의해서 유기화합물의 액체의 열용량을 구하는 방법.

청구항 11

제10항에 있어서, 상기 은닉층의 활성화 함수로는 시그모이드(Sigmoid) 함수를 사용하는 것을 특징으로 하는 SVRC 모형에 의해서 유기화합물의 액체의 열용량을 구하는 방법.

청구항 12

제2항에 있어서, 상기 제4-7단계에서 상기 과적합 방지 기준값은 선형회귀 모형에 의해서 얻은 값인 것을 특징으로 하는 SVRC 모형에 의해서 유기화합물의 액체의 열용량을 구하는 방법.

청구항 13

제2항에 있어서, 상기 제4-0단계에서 추출된 상기 정상끓는점에서의 액체의 열용량에 대한 최적의 분자표현자는

P_1 : 정상끓는점에서의 이상기체의 열용량(heat capacity of gas at T_b),

P_2 : 수소원자의 최대 부분전하(Max partial charge for a H atom),

P_3 : 전체전하 가중 부분음전하 표면적 분율(FNSA-2 Fractional PNSA (PNSA-2/TMSA)),

P_4 : 전체전하 가중 부분음전하 표면적 분율 가중치(WNSA-2 Weighted PNSA (PNSA2*TMSA/1000)),

P_5 : 동경 분포 함수-14.0/원자량 가중(Radial Distribution Function - 14.0 / weighted by atomic masses),
및

P_6 : 동경 분포 함수-8.5/원자 편극도 가중(Radial Distribution Function - 8.5 / weighted by atomic polarizabilities)를 포함하고,

상기 제4-0단계에서 추출된 상기 정상끓는점의 0.55배인 온도에서의 액체의 열용량에 대한 분자 표현자는

P_1 : 정상끓는점*0.55에서의 이상기체의 열용량(heat capacity of gas at $T_{0.55b}$),

P₂: N,O,S,P 극성 계수로 계산된 조각 기반 극성 표면적(fragment-based polar surface area calculated using N,O,S,P polar coefficients),

P₃: 동경 분포 함수-14.0/원자량 가중(Radial Distribution Function - 14.0 / weighted by atomic masses),

P₄: 동경 분포 함수-2.0/원자 반데르발스 부피 가중(Radial Distribution Function - 2.0 / weighted by atomic van der Waals volumes),

P₅: 정보 내용 지수(주변대칭 차수5)(information content index (neighborhood symmetry of 5-order)), 및

P₆: 위상거리8 C-S 결합의 빈도(frequency of C - S at topological distance 08)를 포함하며,

상기 제4-0단계에서 추출된 상기 α_c 에 대한 최적의 분자표현자는

P₁: 위상거리3 N-N 결합의 존재여부(presence/absence of N - N at topological distance 03),

P₂: 2차 성분접근지향적 가중 포괄불변분자 지수/원자 전자위상 상태 가중(2nd component accessibility directional WHIM index / weighted by atomic electrotopological states),

P₃: Burden 행렬의 최고 고유값3/원자 Sanderson 전기음성도 가중(highest eigenvalue n. 3 of Burden matrix / weighted by atomic Sanderson electronegativities),

P₄: 분자 걸음 합계 차수10(molecular walk count of order 10),

P₅: 위상거리3 N-O 결합의 존재여부(presence/absence of N - O at topological distance 03),

P₆: #CR / R=C=R 기능기 수(#CR / R=C=R),

P₇: N히드라진의 수(number of N hydrazines), 및

P₈: 수소원자의 최소 결합차수 (>0.1)(Min (>0.1) bond order of a H atom)를 포함하는 것을 특징으로 하는 SVRC 모형에 의해서 유기화합물의 액체의 열용량을 구하는 방법.

청구항 14

제1항 내지 제13항 중 어느 한 항에 의해 유기화합물의 액체의 열용량을 구하는 방법을 컴퓨터에서 실행시키기 위한 프로그램으로 기록하고 컴퓨터로 읽을 수 있는 저장 매체.

청구항 15

하기 식(2)에 기재되어 있는 SVRC 모형에 대하여

$$\Theta(\varepsilon) = \left(\frac{1 - A^{\varepsilon^B}}{1 - A} \right)$$

$$\alpha = \alpha_c$$

$$\varepsilon = \left(\frac{T_b - T}{T_b - T_{0.55b}} \right)$$

$$C_{pL} = (C_{pL}(T_b)^\alpha - (C_{pL}(T_b)^\alpha - C_{pL}(T_{0.55b})^\alpha)\Theta)^{(1/\alpha)} \quad \text{식(2)}$$

[여기서 A, B는 각 물성에 고유하게 주어지는 보편상수들(universal constants)로서 액체의 열용량의 경우 그 값은 각각 1.33, 1이며, $C_{pL}(T_b)$ 은 정상끓는점에서의 액체의 열용량이고, $C_{pL}(T_{0.55b})$ 는 정상끓는점의 0.55배인 온

도에서의($T_{0.55b}$) 액체의 열용량, α 는 스케일링 지수이며 α_c 는 임계점에서의 스케일링 지수];

정상끓는점에서의 액체의 열용량에 대한 최적의 분자표현자는

P_1 : 정상끓는점에서의 이상기체의 열용량,

P_2 : 수소원자의 최대 부분전하,

P_3 : 전체전하 가중 부분음전하 표면적 분율(FNSA-2 Fractional PNSA (PNSA-2/TMSA)),

P_4 : 전체전하 가중 부분음전하 표면적 분율 가중치(WNSA-2 Weighted PNSA (PNSA2*TMSA/1000)),

P_5 : 동경 분포 함수-14.0/원자량 가중(Radial Distribution Function - 14.0 / weighted by atomic masses),
및

P_6 : 동경 분포 함수-8.5/원자 편극도 가중(Radial Distribution Function - 8.5 / weighted by atomic polarizabilities)를 포함하고,

정상끓는점의 0.55배인 온도에서의 액체의 열용량에 대한 분자 표현자는

P_1 : 정상끓는점*0.55에서의 이상기체의 열용량(heat capacity of gas at $T_{0.55b}$),

P_2 : N,O,S,P 극성 계수로 계산된 조각 기반 극성 표면적(fragment-based polar surface area calculated using N,O,S,P polar coefficients),

P_3 : 동경 분포 함수-14.0/원자량 가중(Radial Distribution Function - 14.0 / weighted by atomic masses),

P_4 : 동경 분포 함수-2.0/원자 반데르발스 부피 가중(Radial Distribution Function - 2.0 / weighted by atomic van der Waals volumes),

P_5 : 정보 내용 지수(주변대칭 차수5)(information content index (neighborhood symmetry of 5-order)), 및

P_6 : 위상거리8 C-S 결합의 빈도(frequency of C - S at topological distance 08)를 포함하며,

α_c 에 대한 최적의 분자표현자는

P_1 : 위상거리3 N-N 결합의 존재여부(presence/absence of N - N at topological distance 03),

P_2 : 2차 성분접근지향적 가중 포괄불변분자 지수/원자 전자위상 상태 가중(2nd component accessibility directional WHIM index / weighted by atomic electrotopological states),

P_3 : Burden 행렬의 최고 고유값3/원자 Sanderson 전기음성도 가중(highest eigenvalue n. 3 of Burden matrix / weighted by atomic Sanderson electronegativities),

P_4 : 분자 걸음 합계 차수10(molecular walk count of order 10),

P_5 : 위상거리3 N-O 결합의 존재여부(presence/absence of N - O at topological distance 03),

P_6 : #CR / R=C=R 기능기 수(#CR / R=C=R),

P_7 : N히드라진의 수(number of N hydrazines), 및

P_8 : 수소원자의 최소 결합차수 (>0.1)(Min (>0.1) bond order of a H atom)를 포함하는 것을 특징으로 하는 SVRC 모형에 의해서 유기화합물의 액체의 열용량을 구하는 방법.

명세서

기술 분야

[0001] 본 발명은, 물성예측이라는 물리화학의 한 분야에 속하는 것으로 화합물의 여러 물성 중 하나인 액체의 열용량을 높은 정확도로 예측하기 위한 방법에 관한 것이다.

배경 기술

[0002] 화합물의 여러 물성의 정확한 값을 구체적으로 아는 것은 그 물질의 용도의 타당성을 검토하거나 합성 및 정제 과정을 설계하고 보관, 운반, 사용, 폐기의 방법과 조건을 설정하는 등, 생산과 소비의 전 과정에 걸친 제반의 사결정 사항들에 결정적이기 때문에 산업적으로나 학문적으로 매우 중요한 문제이다. 관심 있는 화합물의 관심 있는 물성의 값을 가장 정확하게 알 수 있는 방법은 역시 실험이겠으나 정제된 시료의 준비, 정확한 측정을 위한 환경의 구축 등 여러 가지 측면에서 상당한 비용과 시간이 드는 것이 사실이며 경우에 따라서는 불가능할 수도 있다. 따라서, 그 대안으로 오래 전부터 많은 연구자들이 화합물의 여러 물성의 정확한 값을 예측하고자 노력을 기울여 왔다. 이처럼 물성 예측은 오랜 역사를 가지며 끊임 없이 새로운 예측 방법들이 등장하여 현재는 물성 별로 정확도와 적용범위 등이 서로 다른 여러 예측모형들이 공존하고 있는 상황이다.

[0003] 본 발명의 관심 물성인 액체의 열용량에 대해서도 현재까지 여러 예측모형들이 제안되었다. 액체의 열용량(heat capacity of liquid)이라 함은 순수한 물질이 포화상태, 즉 밀폐된 용기 속에서 기체상태와 액체상태가 평형을 이루며 공존하고 있을 때 단위 부피당 액체의 물질량을 말한다. 액체의 열용량의 예측에 대한 그간의 연구결과들은 문헌[Poling B. E., Prausnitz J. M., O'Connell J. P., *The Properties of Gases and Liquids*(5 ed.), New York, McGraw Hill, (2000).]에 간략히 소개되어 있다.

[0004] 현재 액체의 열용량을 예측하는 모형으로 잘 알려지고 널리 쓰이는 것들은 주로 대응상태의 원리에 기반하여 개발된 수식 모형들이다. Bondi 방정식으로 알려진 다음의 수식은 가장 고전적이면서 다른 모형들의 기초가 되었던 모형으로서[A. Bondi, *J&EC Fundamentals*, Vol. 5 No. 4, 442-449(1966).], 기체의 열용량(Heat Capacity of Gas) C_{pG} , 임계온도(Critical Temperature) T_c , 이심인자(Acentric Factor) ω , 기체상수 R 값을 이용하여 임의의 온도에 대한 액체의 열용량(Heat Capacity of Liquid) C_{pL} 의 값을 계산해 준다.

$$C_{pL}(T) = R \left(2.56 + \frac{0.436}{(1 - T/T_c)} + \omega \left(2.91 + \frac{4.28(1 - T/T_c)}{T/T_c} + \frac{0.296}{(1 - T/T_c)} \right) \right) + C_{pG}(T)$$

[0005]

[0006] 이 모형은 $T < 273K$ 인 경우 계산된 C_{pL} 의 값이 부정확하였고, T 값이 임계온도에 가까이 갈수록 상당히 높은 수치로 예측되는 문제점이 발견되었다. 또한 비교적 단순한 형태의 이러한 수식들로 다양한 분자구조를 가진 화합물들 하나하나의 액체의 열용량을 각각 잘 맞추려는 시도는 무리가 있어 보인다.

[0007] 또 다른 방법으로는 그룹기여(group contribution) 방법을 이용한 것이 있다. 순수한 화합물의 (온도나 압력에 따라 변하지 않는) 특정한 한 물성을 예측하는 그룹기여 모형의 전형적인 형태는 아래와 같은 식으로 주어진다.

$$Y = a_0 + \sum_i a_i n_i$$

[0008]

[0009] 물성값 Y 를 구하기 위해서는 먼저 값을 알고자 하는 화합물의 분자를 미리 정해진 다수의 조각형식들에 맞추어 쪼갠 다음 각 조각형식들의 개수 n_i 를 구한다. 이를 다시 그 형식에 할당된 계수 a_i 와 곱한 것을 합산한 것이 예측값 Y 가 된다. 계수 a_0, a_i 들은 실험값이 존재하는 화합물들로부터 모형이 최선의 성능을 갖도록 통계적인 방법을 통해 결정된다.

[0010] Domalski등은 그룹기여 방법을 통해 액체의 열용량 C_{pL} 을 예측하기 위해 온도 T 에 의존하는 다음과 같은 모형을 제안하였다[V Ruzicka, ES Domalski, *J. Phys. Chem. Ref. Data.* (1993).].

$$C_{pL} = R \left(\sum_{i=1}^k n_i a_i + \sum_{i=1}^k n_i b_i \frac{T}{100} + \sum_{i=1}^k n_i d_i \left(\frac{T}{100} \right)^2 \right)$$

[0011]

[0012]

여기서 a_i, b_i, d_i 는 각 조각형식에 할당된 계수들로서 해당 조각형식이 각각 상수 항, 온도의 일차 항, 온도의 이차 항에 기여하는 정도를 반영하고 있다.

[0013]

이러한 그룹기여 방법은 그동안 어느 정도 성공을 거둔 것이 사실이나 이론적 근거가 부족하고 때때로 조각형식에 맞추어 쪼개는 방식이 유일하지 않거나 심지어 존재하지 않는 경우가 발생하여 값의 계산이 불가능해 지기도 한다. 또한 예측성능을 높이기 위해 모형을 개선해 나갈수록 점점 더 복잡해지고 취급이 어려워지는 양상을 보인다.

[0014]

액체의 열용량의 예측모형을 구축하는데 있어서 그룹기여 방법의 대안이 될 수 있는 다른 방법들의 하나는 QSPR(quantitative structure-property relationship) 방법이다. 이 방법은 기본적으로 화합물의 물성은 그 분자의 구조적 특성들의 함수라는 가정에서부터 출발하고 있으며 서로 다른 여러 구조적 특성들을 반영하는 다양한 분자표현자(molecular descriptor)들을 이용한다. 현재까지 제안된 분자표현자들의 종류는 수천에 이르며 한 분자내의 탄소나 수소의 개수와 같은 단순한 것들로부터 분자의 모양이나 연결상태, 전기화학적 특성과 같은 복잡한 것들에 이르기까지 수많은 종류의 분자표현자들에 대한 계산법들이 개발되어 있다[Todeschini R., V. Consonni V., *Molecular Descriptors for Chemoinformatics: Second, Revised and Enlarged Edition: Volume I/II*, Wiley-VCH, 2009]. QSPR 예측모형은 이러한 분자표현자들 그리고 때로는 이에 더하여 화합물의 다른 물리화학적 물성들(이들 역시 구조적 특성들의 함수이다) 중 일부를 독립변수로 포함하는 함수의 형태로 제시된다.

[0015]

이때 이러한 함수의 꼴로 가장 빈번이 채택되는 것은 아래와 같은 표현자 X_i 들의 선형 결합 함수이며 각 계수 c_0, c_i 들은 주로 다중선형회귀분석을 통해 실험데이터로부터 결정된다.

$$Y = c_0 + \sum_{i=1}^n c_i X_i$$

[0016]

[0017]

QSPR 모형을 만드는 또 다른 방법은 인공신경망을 이용하는 것이다. 인공신경망 기법은 지능을 가진 인간의 신경세포를 모델링하여 인공적으로 지능을 가진 기계를 만들어 보고자 하는 인류의 오랜 연구결과의 하나로서, 20세기 중반에 처음 등장하여 현재 다방면으로 응용되고 있는 정보처리기술이다. 도 3은 인공신경망의 전형적인 한 예를 보여주고 있다. 여기서 볼 수 있듯이, 인공신경망에는 입력 데이터를 수용하는 입력층(input layer)과 출력데이터를 만드는 출력층(output layer), 이들 사이에 위치한 은닉층(hidden layer)이 존재하며 각 층은 하나 이상의 노드(node)들로 구성되어 있다. 은닉층의 각 노드들은 입력층과 출력층의 노드들과 연결되어 있으며 각 연결들에는 가중치(weight)라 불리는 양 w_{ij}, w'_{ij} 이 부여되어 있다. 은닉층과 출력층의 각 노드들은 전 단계의 노드들로부터 이러한 연결들을 통해 입력을 받은 뒤 이를 가공하여 출력값을 만드는데 이때 활성화 함수(activation function)라 불리는 함수 f_1, f_2 를 적용한다. 이러한 인공신경망을 실제로 활용하려면 먼저, 다양한 입력값과 그 입력값에 대응하는 출력값을 함께 묶어 놓은 샘플집합을 이용하여 인공신경망을 훈련시키는 과정이 필요한데 이는 주어진 입력에 대한 인공신경망의 출력과 원하는 출력의 차이가 최소가 되도록 역전파(back propagation) 알고리즘을 사용하여 각 연결의 가중치를 최적화 하는 것을 말한다. 이러한 훈련을 거친 인공신경망은 문제해결에 필요한 규칙이나 지식을 따로 제공하지 않아도 학습을 통해서 스스로 일반적인 규칙을 수립하여 미지의 입력에 대해서도 타당성 있는 출력을 내주므로 화합물의 물성예측과 같이 기반 이론이 결여되어 있는 분야에 매우 유용한 수단으로 널리 이용되고 있다.

발명의 내용

해결하려는 과제

[0018]

본 발명이 이루고자 하는 기술적 과제는 위에서 언급된 여러 모형들의 한계들을 극복하고 보다 폭넓고 보다 정확한 예측성능을 보이는, 수소(H), 탄소(C), 질소(N), 산소(O), 황(S) 등 5가지 이내의 원소로 구성되고, 수소

를 제외한 원자의 개수가 25개 이하인 분자로 이루어진 순수한 화합물의 액체의 열용량에 대한 SVRC 모델을 구축하는 것이다.

과제의 해결 수단

- [0019] 우리는 보다 많은 실험데이터를 바탕으로 보다 다양한 분자표현자들을 고려하여 SVRC 수식에 포함된 매개변수들의 값을 예측해주는 QSPR 모형들을 구축함으로써 이 목표를 달성하였다. 이것들은 다중선형회귀분석과 인공신경망 기법을 적절히 조화시켜 얻어진 다중선형회귀-인공신경망 혼성모형인데 특히 인공신경망은 다중선형회귀모형이 반영할 수 없는 독립변수와 종속변수의 비선형적인 함수관계를 반영할 수 있다는 장점이 있어 보다 높은 예측성능을 가진 모형을 구현할 수 있게 해준다. 그러나 인공신경망은 내부적으로 규칙수립의 자유도가 높아 안정성이 다중선형회귀모형보다 떨어지는 단점이 있다. 본 발명에서는 인공신경망모형의 예측값과 다중선형회귀모형의 예측값이 큰 차이를 보일 때 다중선형회귀모형의 예측값을 채택하는 방법으로 이러한 단점을 보완하여 예측성능과 안정성의 측면에서 다중선형회귀모형과 인공신경망모형의 장점만을 살린 우수한 예측모형을 확립하였다.
- [0020] 예측모형을 적용할 수 있는 화합물의 범위에 위에 언급한 바와 같은 제한을 두는 이유는 주로, 사용된 분자표현자들 중 그 값을 구하기 위해서는 양자역학적 계산이 필요한 것들이 존재하는 경우, 현재의 기술수준으로는 언급된 범위를 넘어서는 화합물에 대해서는 정확도와 계산시간의 측면에서 곤란한 문제가 발생한다는 사실에 기인하고 있다. 그러나 상기의 제한범위 내라 할지라도 대단히 많은 화합물들이 존재하며 산업적으로 중요한 화합물들이 상당부분 포함되므로 본 발명이 인류사회에 큰 유익을 끼칠 수 있을 것으로 판단된다.

발명의 효과

- [0021] 오늘날 인류는 플라스틱, 섬유, 고무, 도료, 비료, 의약품, 연료 등, 방대한 종류의 화합물에 의존하여 살아가고 있으며 이러한 경향은 더욱 심화될 것으로 예상된다. 미국 화학회(ACS)에 따르면 2010년 7월 기준으로 등록된 전체 화합물의 수는 54,000,000개를 넘는다고 한다. 이에 비해 물성값이 한가지라도 실험적으로 알려져 있는 화합물의 수는 고작해야 수만에 지나지 않는다. 화합물의 물성값은 신물질과 신약의 개발, 화학플랜트의 최적설계, 기존 설비의 생산성 향상, 자원의 개발과 절약, 안전성 확보, 환경보호 등 인류의 보다 나은 물질생활에 필수적인 요소이다. 특히 액체의 열용량은 화학플랜트의 최적설계 프로그램으로 잘 알려진 AspenPlus 나 Pro/II 와 같은 상용 프로그램이 그 정확한 값을 절실히 요청하는 물성이며 REACH와 같은 화합물의 생산과 소비에 관한 국제적인 규정들이 그 값을 요구하기도 한다. 그러나 현재 그 실험값이 알려진 화합물의 개수는 기껏해야 수천에 불과하며 화합물에 따라서는 독성, 불안정성, 정제의 어려움 등으로 실험을 통하여 데이터를 얻는 작업이 지난한 경우도 있다. 이런 관점에서 실험을 거치지 않고도 분자에 대한 정보만으로 수많은 화합물의 액체의 열용량을 높은 정확도로 얻게 해주는 본 발명은 실험에 드는 비용과 시간을 절감해줄 뿐만이 아니라 실험이 불가능한 경우에도 그 값을 짐작하게 해주어 관련 산업의 연구개발활동을 용이하게 함은 물론 더 나아가 학계와 관(官)계 등 그 값을 필요로 하는 모든 곳에 합당한 정보를 제공하여 그 활동을 보다 원활히 수행할 수 있게 해주는 효과를 낳는다 하겠다.

도면의 간단한 설명

- [0022] 도 1은 본 발명이 제공하는 액체의 열용량에 대한 SVRC 예측모형을 구축하는 과정을 흐름도로 나타낸 도면이다.
- 도 2는 SVRC 모형에 필요한 매개변수들 중 Q^a 에 대한 다중선형회귀모형을 구축하는 과정을 흐름도로 나타낸 도면이다.
- 도 3은 SVRC 모형에 필요한 매개변수들 중 $C_{pL}(T_b), C_{pL}(T_m)$ 에 대한 다중선형회귀-인공신경망 혼성모형을 구축하는 과정을 흐름도로 나타낸 도면이다.
- 도 4는 본 발명에 사용된 인공신경망의 구조를 나타낸 도면이다.
- 도 5~8은 예로 몇몇 화합물에 대해 기존의 예측방법들 중 Bondi모형과 본 발명이 제공하는 예측모형의 예측성능

을 비교한 도면들이다.

도 9는 Bondi 모형의 15517개의 실험데이터에 대한 히스토그램 도면이다.

도 10은 SVRC 모형의 15517개의 실험데이터에 대한 히스토그램 도면이다.

발명을 실시하기 위한 구체적인 내용

- [0023] 도 1은 액체의 열용량에 대한 SVRC 모형을 구축하는 과정을 흐름도로 간략히 표현한 것이다.
- [0024] SVRC 모형을 구축하는데 있어서 가장 먼저 해야 할 일은 단계 1에 지정된 바와 같이 실험데이터를 수집하고 검토 분류하는 일이다. 본 발명을 위해 각종 논문과 단행본, 인터넷 사이트 등을 망라하여 참고할 수 있는 모든 문헌과 자료에 대한 광범위한 조사를 벌인 결과, 총 1796개 화합물에 대한 32227개의 데이터가 수집되었다. 이렇게 수집된 데이터가 모형을 구축하는데 쓰일 수 있는 진정한 값인지 다방면으로 검토하였는데 실험값이 아니거나 데이터 표기에 오류가 있었거나 동일 화합물의 비슷한 온도에 대한 값들임에도 불구하고 차이가 많이 나거나 가까운 다른 화합물들의 값에 비해 신뢰하기 어려울 정도로 동떨어진 값이거나 분자표현자들에 대한 값들이 당장 준비되기 어려운 화합물에 대한 데이터인 경우 등에 대해 면밀한 분석을 거쳐 데이터를 수정 또는 삭제하여 최종적으로 총 841개의 화합물들에 대한 15525개의 데이터를 선정하였다. 또한, 본 발명에서 ‘유기화합물’ 또는 ‘화합물’은 수소(H), 탄소(C), 질소(N), 산소(O), 황(S) 등 5가지 이내의 원소로 구성되고, 수소를 제외한 원자의 개수가 25개 이하인 분자로 이루어진 물질을 지칭한다.
- [0025] 그 다음 단계는 이들 화합물들에 대한 분자표현자들의 값들을 준비하는 단계이다. 총 1978개에 달하는 다양한 분자표현자들에 대한 값들을 각 화합물들의 분자에 대한 정보를 담은 파일들로부터 컴퓨터를 이용하여 일괄적으로 계산한다. 분자의 전자구조 계산을 하기 위해서는 보통 순이론인 방법으로 슈뢰딩거(Schrodinger) 방정식을 풀어 전자에너지에 대한 해를 구하게 되지만 전자가 많은 계의 경우 전자상관관계(electron correlation)를 무시한 근사법을 적용한 하트리-포크(Hartree-Fock, HF) 방법[C. C. J. Roothan, Rev. Mod. Phys. 23, 69 (1951)]을 사용하여 해를 풀게 된다. 이런 근사법으로 인해 계산된 결과에서 근본적인 오차가 유발되어 다차원의 이론적인 섭동항을 추가한 포스트 하트리-포크(Post Hartree-Fock) 방법[C. Moller and M. S. Plesset, Phys. Rev. 46, 618 (1934)]들을 사용하여 더 정확한 해를 구하긴 하지만 상대적으로 엄청나게 많은 계산량이 요구된다. 이런 방식으로는 큰 분자를 계산하기에는 비용이나 시간의 측면에서 무리가 있는 상황이다.
- [0026] 또한 하트리-포크와 포스트 하트리-포크를 조합한 가우시안 방법[L. A. Curtiss, K. Raghavachari, G. W. Trucks, and J. A. Pople, J. Chem. Phys. 94, 7221 (1991); L. A. Curtiss, K. Raghavachari, P. C. Redfern, V. Rassolov, and J. A. Pople, J. Chem. Phys. 109, 7764 (1998)]은 에너지 예측에 있어 아주 적은 오차를 보이지만 여러 포스트 하트리-포크 방법에 대한 에너지 계산을 수행하기 때문에 더 많은 계산량이 요구된다.
- [0027] 많은 전자로 이루어진 분자에 대한 전자들간의 상관관계를 고려하기 위해 다차원의 섭동항이 추가된 파동함수 대신 전자 밀도함수를 써서 총에너지의 범함수를 이용해서 바닥상태를 구하는 밀도 범함수 이론(density functional theory)[R. Seeger and J. A. Pople, J. Chem. Phys. 66, 3045 (1977)]을 적용하여 계산했다. 밀도 범함수 이론의 장점은 전자밀도만 고려하면 되므로 하트리-포크(Hartree-Fock) 방법과 비슷한 수준의 계산량으로 더 정확한 결과를 얻어낼 수 있다. 전자들의 교환-상관에너지를 계산을 위해 교환 범함수들과 상관 범함수들의 조합을 사용하여 계산량을 늘이지 않고도 더 향상된 결과를 얻고 있다.
- [0028] 최적의 양자역학 계산방법을 선별하기 위해 선행하여 시도하였던 계산이론은 상기에 언급된 하트리-포크 방법, 다양한 포스트 하트리-포크 방법, 가우시안(G2, G3) 방법, 다양한 범함수 조합의 밀도 범함수 이론 등이다. 이 중에서 계산시간 대비 가장 성능이 우수한 밀도 범함수 이론의 한가지 방법을 선별하였다.
- [0029] 따라서 본 발명에서는 상용 양자역학 계산 프로그램을 이용하여 지정된 밀도 범함수 이론의 계산방법을 적용하여 분자구조에 대한 최적화 및 진동수 계산을 수행하게 된다.
- [0030] 최적화된 구조에서는 상기의 물성정보 뿐 만 아니라 분자의 특성을 반영하는 여러 의미있는 수치로 나타내는 분자표현자(molecular descriptor)들을 얻을 수 있다. 2차원 구조의 특징을 표현할 수 있는 분자표현자들도 있고 3차원 구조의 특징을 표현하는 분자표현자들도 있다. 크게 나누어 총 24개의 범주로 나누고 각 범주에 상세한 표현자들을 포함한다. 분자 표현자 값을 계산한 후에 이들 중 적합하지 않은 것, 즉 모든 샘플 화합물에 대해 값이 동일하게 나와 모형의 독립변수가 될 수 없는 것들을 추려 내었다. 이는 관련 없는 분자표현자가 예측모형에 포함되는 것을 막아 모형의 신뢰도를 높이는 동시에 이렇게 분자표현자의 개수를 줄임으로써 최적 모형을 찾

는 데 드는 계산 시간을 줄일 수 있기 때문이다.

[0031] 그 다음인 단계 4는 실험데이터를 토대로 각 매개변수에 대한 QSPR 모델을 확립하는데 필요한 자료를 준비하는 단계이다. SVRC 수식으로 액체의 열용량을 계산하기 위해서는 정상끓는점과 삼중점에서의 액체의 열용량 $C_{pL}(T_b), C_{pL}(T_m)$ 의 값과 α_c 의 값이 필요하다. 한편 삼중점의 온도를 정확히 예측하는 것은 일반적으로 매우 어려운 일임이 잘 알려져 있으며 따라서 본 발명에서는 삼중점 대신 정상끓는점의 0.55배($T_{0.55b}$)를 액체의 열용량 곡선의 시작 온도로 잡았는데, 샘플 화합물들에 대해 이 지점들의 평균은 삼중점의 평균과 거의 일치한다. 이제 SVRC 모델을 완성하기 위해서는 매개변수들인 $\alpha_c, C_{pL}(T_b), C_{pL}(T_{0.55b})$ 에 대한 QSPR 예측모형을 확립하여야 한다. 이러한 QSPR 예측모형을 확립하기 위해서는 각 매개변수별로 여러 화합물들에 대한 해당 값들의 집합을 마련하여야 하는데, $C_{pL}(T_b), C_{pL}(T_{0.55b})$ 에 대해서는 먼저 액체의 열용량의 전형적인 곡선을 각 화합물의 실험데이터에 맞추어, 그 곡선에서 온도가 $T_b, T_{0.55b}$ 가 되는 지점의 값을 취하였으며, α_c 에 대해서는 비선형방정식의 수치해법을 통하여 아래의 식에서 α 의 값을 각 온도에 대해 구한 뒤, 이러한 값들을 평균 내어 취하였다.

$$\Theta(\varepsilon) = \left(\frac{1 - A^{\varepsilon^B}}{1 - A} \right)$$

$$\alpha = \alpha_c$$

$$\varepsilon = \left(\frac{T_b - T}{T_b - T_{0.55b}} \right)$$

$$C_{pL} = (C_{pL}(T_b))^{\alpha} - (C_{pL}(T_b))^{\alpha} - C_{pL}(T_{0.55b})^{\alpha}) \Theta)^{(1/\alpha)}$$

[0032]

[0033] 여기서 Θ 와 α 는 각각 correlating 함수, scaling 지수라고 불리는 양들이며 ε 는 일종의 환산온도(reduced temperature)이다. 또한 A, B, C는 각 물성에 고유하게 주어지는 보편상수들(universal constants)로서 액체의 열용량의 경우 그 값을 각각 1.33, 1, 0으로 정하였다. 이렇게 QSPR 예측모형을 확립하기 위한 자료로 쓰기 위해 한 화합물의 실험데이터로부터 각 매개변수들의 값을 구할 수 있으려면 그 화합물의 실험데이터가 비교적 넓은 온도 범위에 걸쳐 고루 분포하면서 데이터에 노이즈(noise)가 적어야 하는데 상대적으로 이런 조건을 만족하는 실험데이터가 존재하는 화합물들의 개수는 많지 않으므로 실제 각 매개변수의 QSPR 예측모형에 샘플로 참여하는 화합물의 개수는 전체 화합물의 개수보다 상당히 줄어들게 된다.

[0034] 단계 5는 각 매개변수에 대한 QSPR 모델을 구축하는 단계이다. 도 2는 각 매개변수에 대한 QSPR 모델을 구축하는 과정을 흐름도로 간략히 표현한 것이다. 그 구체적인 세부 단계들은 다음과 같다.

[0035] 먼저 세부 단계 1에서는 샘플 화합물들을 예측모형을 탐색하는데 사용할 훈련집합(training set)과 결정된 모형의 예측성능을 시험하는데 사용할 시험집합(test set), 이렇게 두 부분으로 나누는 작업을 진행한다. 유사한 분자들이 한쪽 부분에만 치우쳐 분포하지 않도록 주의하면서 샘플을 5:5 ~ 8:2, 바람직하게는 6 대 4의 비율로 각각 나누었다.

[0036] 이후 훈련집합을 토대로 최선의 다중선형회귀모형(multiple linear regression model)을 찾는다. 여기서 ‘최선’이라 함은 상대적인 의미로서 비교적 짧은 시간 내에 구할 수 있으면서 절대적인 의미에서의 최적 해에 매우 근접한 성능을 갖는다는 의미로 쓰여진 것이다. 최적 해를 직접 구하지 않는 이유는 긴 계산시간 때문인데 예를 들어 1978개의 분자표현자들 중 적합한 분자표현자들의 개수가 1700개일 때, 이 중에서 5개를 뽑아 만들 수 있는 서로 다른 다중선형회귀모형들의 총 개수는 $\binom{1700}{5} = \frac{1700!}{1695!5!}$ 이며 이들을 다 조사하는 것은 현실적으로 불가능하다.

[0037] 한정된 시간 내에 유용한 결과를 얻기 위해 본 발명에서는 유전적 알고리즘(genetic algorithm) [Judson,

"Genetic Algorithms and Their Uses in Chemistry", Reviews in Computational Chemistry, Lipkowitz & Boyd, Eds., Vol.10, pp.1-73 (VCH Publishers, NY, 1997)]을 채택하였으며 그 상세한 방법은 다음과 같다. 먼저 분자표현자들의 풀(pool)에서 일정한 개수의 분자표현자들을 무작위로 뽑아 만든 다수의 다중선택회귀모형들로 구성된 개체군(population)을 생성한다. 예를 들어 1700개의 적합한 분자표현자들 중 5개를 무작위로 뽑아 만든 1000개의 서로 다른 다중선택회귀모형들로 개체군을 만들었다고 하자.

[0038] 이때 염색체(chromosome)라 불리는 각 개체(individual)들은 뽑힌 분자표현자들의 번호들을 조합하여 부호화한다. 예를 들어 1700개의 분자표현자중 45, 167, 684, 1033, 1502번째의 분자표현자들로 형성한 다중선택회귀모형의 염색체는 (45, 167, 684, 1033, 1502)와 같이 표현할 수 있다. 이렇게 생성된 개체군으로부터 두개의 부모 염색체를 선택한 뒤 교배(crossover)하여 자식들을 만들어 내는데 본 발명에서는 부모 염색체를 선택하는 선택 기법으로 Roulette Wheel 방법을 채택하였다.

[0039] Roulette Wheel 방법은 일반적으로 가장 많이 사용하는 선택 알고리즘으로 각 염색체의 적합도(fitness)에 비례하는 만큼 룰렛의 영역을 그 염색체에 할당한 다음, 룰렛을 돌려 해당된 영역의 염색체를 선택하는 방법이다. 따라서 이 방법에서는 적합도가 높은 개체일수록 선택될 확률이 높다. 선택확률을 결정짓는 각 염색체의 적합도 계산에는 회귀모형의 결정계수(coefficient of determination: R^2) 또는 평균절대오차(average absolute error: AAE)를 활용하였다. 즉 결정계수값이 크거나 평균절대오차값이 작은 것이 선택확률이 높도록 하였다.

[0040] 교배방법으로는 단순교배(single point crossover)법을 채택하였는데 이는 가장 일반적인 교배 방법으로서 부모 염색체에서 임의로 1개의 교배점을 선택하여 그 지점 전후의 염색체부분을 서로 교환함으로써 자식을 생성하는 것을 말한다. 예를 들어 부모 염색체가 각각 (24, 262, 343, 789, 1290), (38, 454, 554, 1322, 1449)와 같이 주어지고 3번째와 4번째 요소 사이에 교배점이 놓이게 되면 자식 염색체는 각각 (24, 262, 343, 1322, 1449), (38, 454, 554, 789, 1290)와 같이 된다.

[0041] 이렇게 자식들이 생성되면 이들의 염색체 일부를 일정 확률로 돌연변이(mutation) 시키는 과정을 거치는데 이는 임의로 몇 개의 요소를 전혀 새로운 값으로 바꾸는 것으로 현재 집단에 존재하지 않는 새로운 정보로 초기 유전자 조합 이외의 공간을 탐색할 수 있게 해주어 초기 집합의 조합 내에 적절한 해가 없을 경우를 보완해주는 과정이다.

[0042] 이 같은 방법으로 새로이 구해진 개체들로 기존 개체군의 일부 또는 전부를 교체하여 새 세대의 개체군을 생성한다. 이 과정을 반복하여 그 세대수가 미리 정한 값(보통 10~1000사이에서 선택)에 이르면 가장 적합도가 큰 개체, 즉 예측성능이 가장 좋은 회귀모형을 선택하고 끝낸다.

[0043] 일단 이렇게 최선의 다중선택회귀모형이 선정되면 다음 단계로 이 모형의 타당성을 검토한다. 만일 모형에 포함된 분자표현자의 t검정값이 좋지 않다면 하는 문제점이 발견되면 이전 단계로 돌아가 다른 모형을 찾는다. 예를 들어 샘플 화합물의 수가 1005이고 선정된 모형이 5개의 분자표현자로 구성되어 있을 경우 그 중 한 분자표현자에 대한 t검정값이 3.3이상이면 이는 이 분자표현자가 해당 물성과 무관할 확률이 0.1%이하임을 뜻한다. 본 발명에서는 대략 3미만의 t검정값을 갖는 분자표현자가 존재할 경우 선정된 모형을 버리고 다른 모형을 찾았다. 또한 샘플 화합물들에 대한 한 분자표현자의 값들이 소수의 몇몇 화합물들을 제외하고는 모두 동일한 경우도 신뢰성 있는 모형이라고 볼 수 없어 마찬가지로 조처하였다. 일반적으로 모형에 포함되는 분자표현자의 개수를 늘리면 예측성능은 높아지지만 이와 같은 문제들이 발생하게 되므로 보통 최종 모형은 이 단계들을 분자표현자의 개수를 바꿔가며 여러 번의 시행착오를 거쳐 반복 수행함으로써 얻어진다. 선정된 모형에 더 이상 문제가 나타나지 않으면 다음 단계로 넘어간다.

[0044] 마지막인 세부 단계 4에서는 모형을 형성하는데 참여하지 않았던 시험집합을 이용하여 찾아낸 모형의 예측성능을 평가한다. 만일 훈련집합에서 보다 예측성능이 많이 떨어지거나 예측이 크게 벗어나는 샘플들이 보이는 등의 문제점이 발견되면 세부 단계 1로 가서 훈련집합과 시험집합을 재조정된 뒤 이후 세부 단계를 진행한다. 여기서 훈련집합과 시험집합의 차이가 훈련집합에 대해서 얻은 절대평균오차(AAE)의 20%를 넘지 않으면 예측성능이 만족되는 것으로 판단한다.

[0045] 이렇게 하여 한 매개변수에 대한 다중선택회귀모형이 일단 구축되면 인공신경망모형을 구축하기 위해 먼저 분자표현자들의 데이터와 매개변수의 데이터를 표준화하는 작업, 즉 각 값에서 해당 데이터의 평균을 뺀 뒤 표준편차로 나누는 작업을 진행한다. 이렇게 준비된 전체 샘플을 대략6:2:2의 비율로 훈련집합(training set), 검증집합(validation set), 시험집합(test set)으로 나눈다.

[0046] 이후 이들을 사용하여 최선의 인공신경망모형을 탐색한다. 이때 탐색 범위는 도 4에서처럼 입력층과 출력층 사이에 한 개의 은닉층을 가지면서 이 3개 층이 전방향으로(feed forward), 즉 입력에서 출력으로 향하는 방향으로만 연결되어 있는 구조를 갖는 신경망으로 제한하였다. 입력층은 이미 확립되어 있는 다중선형회귀모형에 포함된 각 분자표현자들의 값을 입력 받는, 같은 개수만큼의 노드들로 구성하였으며 출력층은 액체의 열용량을 출력하는 한 개의 노드로 구성하였다. 또한 은닉층의 활성화 함수로는 Sigmoid 함수 즉 $f_1(x) = (1 + e^{-x})^{-1}$ 을, 출력층의 활성화 함수로는 선형함수 즉 $f_2(x) = x$ 를 채택하였다. 따라서 입력층의 각 노드들이 받는 입력값들을 $I_1, I_2, \dots, I_l$ 라 할 때 은닉층의 j번째 노드의 출력값은 $H_j = (1 + \exp(-T - \sum_{i=1}^l w_{ij} I_i))^{-1}$ 와 같이 주어지며 은닉층이 m 개의 노드로 이루어져 있을 때 출력층 출력노드의 최종 출력값은 $O = T + \sum_{i=1}^m w'_{ij} H_i$ 와 같이 주어진다. 여기서 T 는 문턱 가중치(threshold weight)를 의미한다.

[0047] 탐색은 은닉노드의 수가 1개인 것부터 차례로 개수를 늘려가며 진행하는데 보통 입력노드 개수의 2배가 될 때까지 진행하지만 만족스러운 모형이 나오지 않을 경우 더 진행하여 탐색한다. 자세한 절차는 다음과 같다. 먼저 은닉노드의 각 개수 별로, 난수 발생 함수를 써서 생성한 가중치 T, w_{ij}, w'_{ij} 들의 다양한 초기값세트(보통1000 세트이내)를 마련하고, 훈련집합을 사용하여 각 세트로 초기화된 신경망을 역전파 알고리즘을 통해 반복 훈련함으로써 가중치 T, w_{ij}, w'_{ij} 들의 최적화된 값을 찾는다. 최적화에 대한 판단은 매 훈련 후 경신된 가중치들의 값으로 정해지는 모형을 검증집합에 적용하였을 때 그 평균제곱오차(mean square error)의 값이 최소가 되는 것으로 한다. 보통은 3000~5000번의 반복훈련 내에 이러한 시점이 나오게 된다. 이렇게 얻어진 각 초기값세트에 대응하는 최적화된 신경망모형을 훈련집합, 검증집합, 시험집합에 각각 적용하여 그 평균제곱오차들이 모두 다중선형회귀모형의 그것들보다 작은 것만을 모은다. 이러한 것이 여러 개 있을 경우, 결정계수나 평균절대오차를 기준으로 가장 우수한 모형을 선택한다.

[0048] 이렇게 인공신경망모형이 선정되면 마지막으로 과적합(overfitting) 방지기준을 설정한다. 이는 과도한 훈련의 결과로 인공신경망이 미지의 입력에 대해 엉뚱한 답을 내놓는 불안정성을 개선하기 위한 조치로, 한 기준값(다중선형회귀모형 값)을 정하여 인공신경망모형의 예측값이 기준값인 다중선형회귀모형에서 얻은 값의 85% 내지 115% 사이에 존재하면 인공신경망모형의 값을 채택하고 존재하지 않으면 다중선형회귀모형의 예측값을 채택하게 하는 것을 말한다.

[0049] 이 같은 과정을 거쳐 각 매개변수에 대한 QSPR 모형이 구축되면 다음으로 단계 6에서는 액체의 열용량에 대한 각 화합물의 실험데이터 전체를 SVRC 수식을 통해 계산된 값과 비교하는 테스트를 진행한다. 이때 SVRC 수식으로 예측값을 계산하려면 정상끓는점과 정상끓는점의 0.55배 되는 지점의 값이 필요한데 이 정보에 대해서는 본 발명의 출원인에 의해 이미 개발되어 있는 정상끓는점을 예측하는 모형을 활용하였다. 만일 실험값과 예측값의 오차가 용인할 수 있는 수준보다 클 경우(예측오차가 실험평균오차보다 큰 경우가 대략 10%이상일 때), 단계 5로 돌아가 각 매개변수들에 대한 QSPR 모형을 다시 잡는다. 테스트를 통과한 경우 이를 완성된 SVRC 모형으로 채택한다.

[0050] 이러한 과정을 거쳐 확립된 SVRC 모형에 대한 결과는 표 1~4에 간략히 정리되어 있다. 표 1(0.55배의 정상끓는점에서의 액체의 열용량($C_{pL}(T_{0.55b})$)에 대한 QSPR 예측모형의 주요 내용), 표 2(정상끓는점에서의 액체의 열용량($C_{pL}(T_b)$)에 대한 QSPR 예측모형의 주요 내용), 표 3(a_c 에 대한 QSPR 예측모형의 주요 내용)은 각각 탄화수소에 대한 $C_{pL}(T_b)$, $C_{pL}(T_{0.55b})$, a_c 의 값을 예측하는 QSPR 모형을 간단히 기술한 것이며 이들을 바탕으로 액체의 열용량을 예측하는 SVRC 모형과 그 성능에 대한 결과는 표 4에 나와 있다.

표 1

[0051]	샘플 화합물들의 개수	548
	분자표현자들의 개수	6

분자표현자들의 이름	P_1 : 정상끓는점의 0.55배 온도에서의 이상기체의 열용량 P_2 : N,O,S,P 극성 계수로 계산된 조각 기반 극성 표면적 P_3 : 동경 분포 함수-14.0/원자량 가중(Radial Distribution Function - 14.0 / weighted by atomic masses) P_4 : 동경 분포 함수-2.0/원자 반데르발스 부피 가중(Radial Distribution Function - 2.0 / weighted by atomic van der Waals volumes) P_5 : 정보 내용 지수(주변대칭 차수5) P_6 : 위상거리8 C-S 결합의 빈도(frequency of C - S at topological distance 08)
회귀모형 결정계수	0.98474991
회귀모형 AEA	2.85590502 cal/mol.K
회귀모형	$C_{p-}(T_{c55b})(\text{cal/mol.K}) = c_c - \sum_{i=1}^6 c_i P_i$
인공신경망 결정계수	0.98909344
인공신경망 AEA	2.52679081 cal/mol.K
인공신경망모형	$C_{p-}(T_{c55b}) = T - \sum_{j=1}^5 w'_{j1} (1 - \exp(-T - \sum_{i=1}^6 w_{ij} P_i))^{-1}$
과적합 방지기준	15%

표 2

[0052]

샘플 화합물들의 개수	548
분자표현자들의 개수	6
분자표현자들의 이름	P_1 : 정상끓는점에서의 이상기체의 열용량 P_2 : 수소원자의 최대 부분전하 P_3 : 전체전하 가중 부분음전하 표면적 분율(FNSA-2 Fractional PNSA (PNSA-2/TMSA)) P_4 : 전체전하 가중 부분음전하 표면적 분율 가중치(WNSA-2 Weighted PNSA (PNSA2*TMSA/1000)) P_5 : 동경 분포 함수-14.0/원자량 가중(Radial Distribution Function - 14.0 / weighted by atomic masses) P_6 : 동경 분포 함수-8.5/원자 편극도 가중(Radial Distribution Function - 8.5 / weighted by atomic polarizabilities)
회귀모형 결정계수	0.9685054
회귀모형 AEA	5.8203887 cal/mol.K
회귀모형	$C_{p-}(T_b)(\text{cal/mol.K}) = c_c - \sum_{i=1}^6 c_i P_i$
인공신경망 결정계수	0.98396278
인공신경망 AEA	4.05659284 cal/mol.K
인공신경망모형	$C_{p-}(T_b) = T - \sum_{j=1}^5 w'_{j1} (1 - \exp(-T - \sum_{i=1}^6 w_{ij} P_i))^{-1}$
과적합 방지기준	15%

표 3

[0053]

샘플 화합물들의 개수	541
분자표현자들의 개수	8

분자표현자들의 이름	P₁: 위상거리3 N-N 결합의 존재여부(presence/absence of N - N at topological distance 03) P₂: 2차 성분접근지향적 가중 포괄불변분자 지수/원자 전자위상 상태 가중(2nd component accessibility directional WHIM index / weighted by atomic electrotopological states) P₃: Burden 행렬의 최고 고유값3/원자 Sanderson 전기음성도 가중(highest eigenvalue n. 3 of Burden matrix / weighted by atomic Sanderson electronegativities) P₄: 분자 걸음 합계 차수10(molecular walk count of order 10) P₅: 위상거리3 N-O 결합의 존재여부(presence/absence of N - O at topological distance 03) P₆: #CR / R=C=R 기능기 수(#CR / R=C=R) P₇: N히드라진의 수(number of N hydrazines) P₈: 수소원자의 최소 결합차수 (>0.1)(Min (>0.1) bond order of a H atom)
결정계수	0.36392153
평균절대오차	2.827778646
모형	$\alpha_c = c_c - \sum_{i=1}^8 c_i P_i$

표 4

샘플 화합물들의 개수	841
실험데이터의 개수	15517
결정계수	0.986031
평균절대오차	2.42 cal/mol.K
모형	$C_{pL} = (C_{pL}(T_b))^{\alpha} - (C_{pL}(T_b))^{\alpha} - C_{pL}(T_{0.55b})^{\alpha} \Theta)^{(1/\alpha)}$ $\Theta(\varepsilon) = \left(\frac{1 - A^{\varepsilon}}{1 - A} \right)$ $\alpha = \alpha_c \cdot \varepsilon$ $\varepsilon = \left(\frac{T_b - T}{T_b - T_{0.55b}} \right)^{\beta}$

[0054]

액체의 열용량을 예측하는데 있어서 기체의 열용량과의 상관관계가 아주 많은 것으로 알려져 있어서 본 발명에서도 $C_{pL}(T_b)$, $C_{pL}(T_{0.55b})$ 를 예측하는 SVRC 모형에 사용되는 $C_{pG}(T_b)$, $C_{pG}(T_{0.55b})$ 의 값은 이미 알려져 있는 값 또는 QSPR 모형에 근거한 계산 방법에 의해서 구한 정상끓는점과 정성끓는점의 0.55배인 두 지점의 값을 이용하여 활용하였다.

[0056]

본 발명이 기존 기술보다 우월함을 보이기 위해 이렇게 확립된 SVRC 모형과 널리 사용되는 기존의 모형 즉 Bondi 모형[*A. Bondi, J&EC Fundamentals*, Vol. 5 No. 4, 442-449(1966).]의 예측성능을 실험값이 알려진 841개의 화합물들의 데이터를 사용하여 비교하였다. 그 결과 Bondi 모형은 841개 화합물 전체에 대해서 0.050479의 결정계수값과 11.75 cal/mol.K의 평균절대오차값을 갖음을 알게 되었다. 반면 SVRC 모형은 841개 화합물 전체에 대해서 0.986031의 결정계수값과 2.42 cal/mol.K의 평균절대오차값을 가져 Bondi 모형보다 우수함을 알게 되었다. 도 5~8은 예로 몇몇 화합물에 대해 각 모형의 예측성능을 비교한 도면들이다. 이도면들로부터 본 발명의 모형이 기존 모형보다 우수한 성능을 가짐을 눈으로 확인할 수 있다. 한편 15517개의 실험데이터에 대해 실험값과 예측값 사이의 오차를 히스토그램으로 그린 것이 도 9, 10이다. 이 도면들은, Bondi 모형은 71.19%, 본 발명의 모형은 91.01%의 확률로 평균확률오차인 10%의 오차 이내로 액체의 열용량을 예측하고 있음을 보여주어 본 발명의 모형이 보다 정확함을 증명해준다.

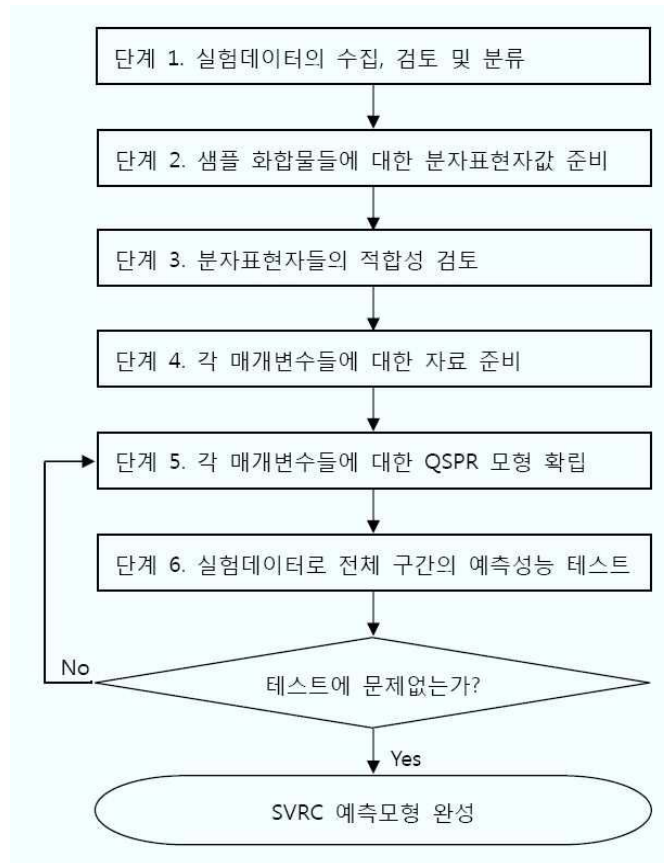
[0057]

본 발명은 상기한 실시예에 한정되지 아니하며, 청구범위에서 청구하는 본 발명의 요지를 벗어남이 없이 당해 발명이 속하는 기술분야에서 통상의 지식을 가진 자라면 누구든지 다양한 변형실시가 가능한 것은 물론, 그와

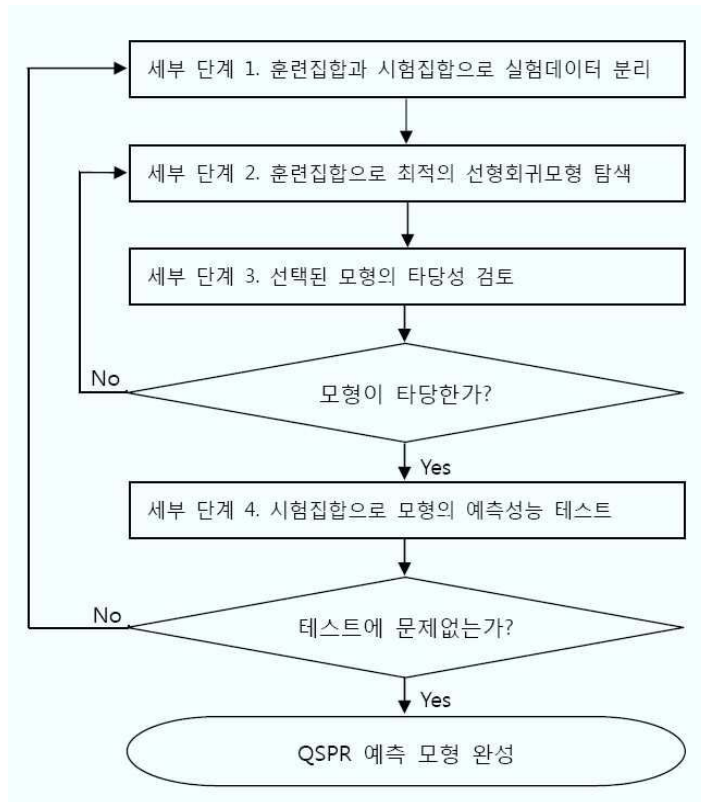
같은 변경은 청구범위 기재의 범위 내에 있게 된다.

도면

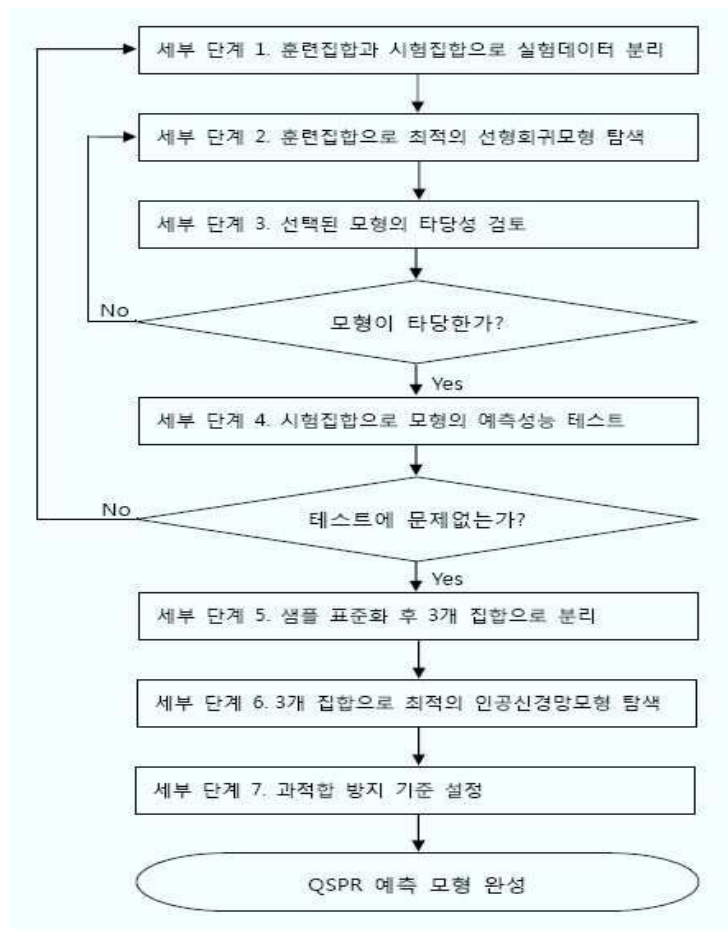
도면1



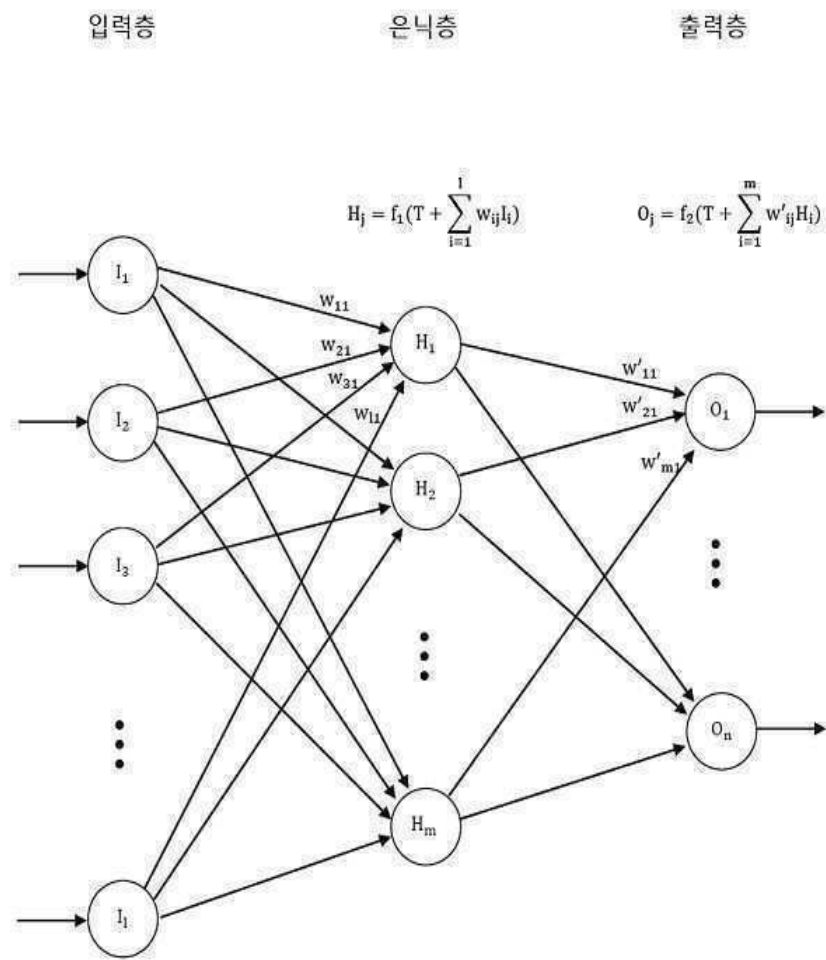
도면2



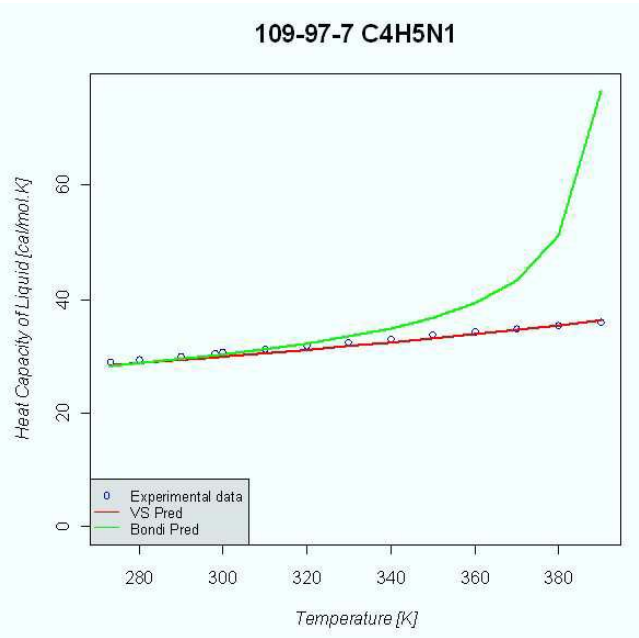
도면3



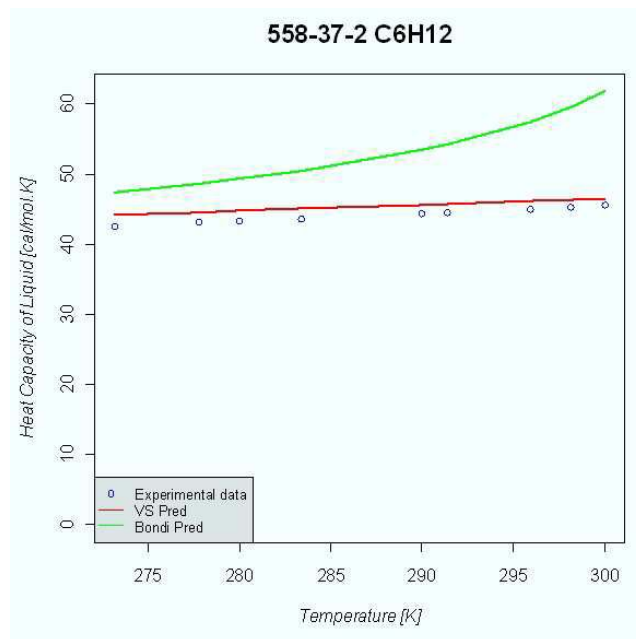
도면4



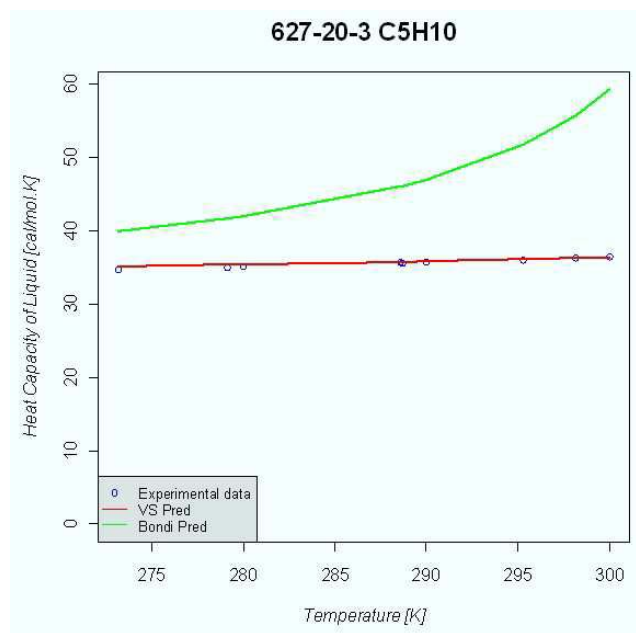
도면5



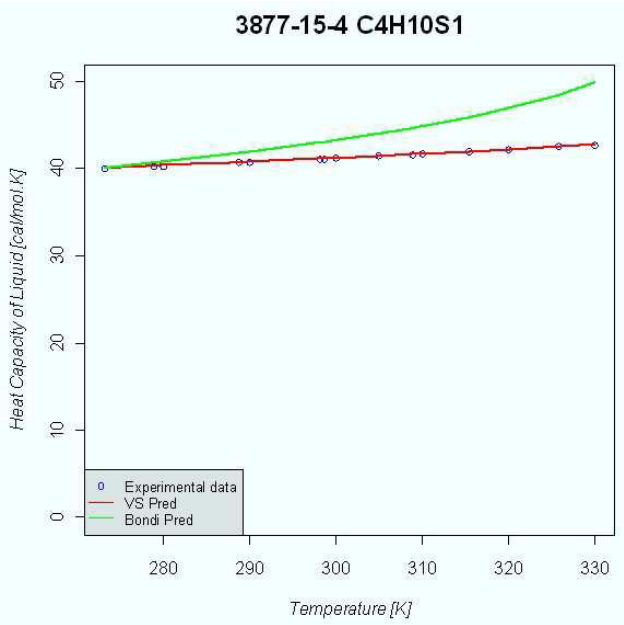
도면6



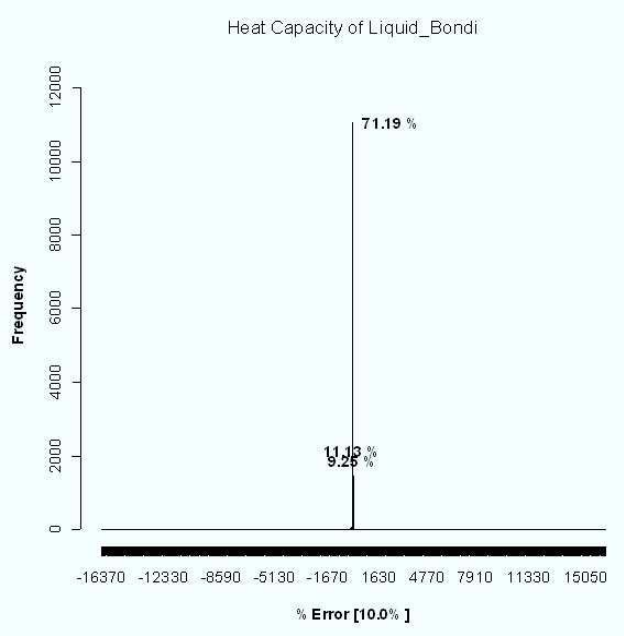
도면7



도면8



도면9



도면10

