퍼지 및 지능시스템학회 논문지 2007, Vol. 17, No. 5, pp. 679-683

차분진화 기반의 Support Vector Clustering

A Differential Evolution based Support Vector Clustering

전성해
Sung-Hae Jun

청주대학교 바이오정보통계학과

요 약

Vapnik의 통계적 학습이론은 분류, 회귀, 그리고 군집화를 위하여 SVM(support vector machine), SVR(support vector regression), 그리고 SVC(support vector clustering)의 3가지 학습 알고리즘을 포함한다. 이들 중에서 SVC는 가우시안 커널함수에 기반한 지지벡터를 이용하여 비교적 우수한 군집화 결과를 제공하고 있다. 하지만 SVM, SVR과 마찬가지로 SVC도 커널모수와 정규화상수에 대한 최적결정이 요구된다. 하지만 대부분의 분석작업에서 사용자의 주관적 경험에 의존하거나 격자탐색과 같이 많은 컴퓨팅 시간을 요구하는 전략에 의존하고 있다. 본 논문에서는 SVC에서 사용되는 커널모수와 정규화상수의 효율적인 결정을 위하여 차분진화를 이용한 DESVC(differential evolution based SVC)를 제안한다. UCI Machine Learning repository의 학습데이터와 시뮬레이션 데이터 집합들을 이용한 실험을 통하여 기존의 기계학습 알고리즘과의 성능평가를 수행한다.

Abstract

Statistical learning theory by Vapnik consists of support vector machine(SVM), support vector regression(SVR), and support vector clustering(SVC) for classification, regression, and clustering respectively. In this algorithms, SVC is good clustering algorithm using support vectors based on Gaussian kernel function. But, similar to SVM and SVR, SVC needs to determine kernel parameters and regularization constant optimally. In general, the parameters have been determined by the arts of researchers and grid search which is demanded computing time heavily. In this paper, we propose a differential evolution based SVC(DESVC) which combines differential evolution into SVC for efficient selection of kernel parameters and regularization constant. To verify improved performance of our DESVC, we make experiments using the data sets from UCI machine learning repository and simulation.

Key Words : 차분진화, Support Vector Clustering, 커널모수, 정규화상수

1. 서 론

기계학습(machine learning) 알고리즘에서 발생되는 지역최적화(local optima)와 과대적합(over-fitting)의 문제점을 해결하기 위하여 Vapnik이 제안한 통계적 학습이론(statistical learning theory)은 분류(classification)를 위한 Support Vector Machine(SVM), 회귀(regression)를 위한 Support Vector Regression(SVR), 그리고 군집화(clustering)를 위한 Support Vector Clustering(SVC)으로 이루어진다[19]. 하나의 학습이론이 지도와 자율학습을 모두 수행할 수 있는 유연하고 강력한 알고리즘이다. 현재 데이터 마이닝을 포함한 다양한 분야에서 중요한 분석도구로 사용되어 우수한 성능을 보이고 있다[1-3,10,12,17].

통계적 학습이론은 커널함수(kernel function)와 정규화(regularization)를 위한 적절한 모수들의 사용을 통하여 학습데이터 집합을 주어진 데이터공간에서 고차원 형상공간으로 사상시켜 학습을 수행한다. 통계적 학습이론에서 커널모수(kernel parameters)와 정규화상수(regularization constant)의 적절한 선택은 학습결과에 직접적으로 영향을 미친다. 일반적으로 통계적 학습이론의 모수들은 분석가의 주관적 지식에 의해 결정되고 있다. 하지만 대부분의 주관적인 결정이 최적의 결과를 제공하지는 못한다[3-5]. 따라서 주관적 결정은 모형의 성능을 안정화 시키는데 있어서 저해요인일 뿐만 아니라 많은 시행착오(trial and error) 과정을 요구함으로써 최적의 모형구축에 많은 시간과 비용을 필요하게 한다. 주관적 결정이외에 격자 탐색과 같은 전역 탐색방법 기법들은 최적모수를 찾기 위하여 매우 많은 컴퓨팅 시간을 요구하기 때문에 일반적으로 사용하기에는 어려움이 있다.

본 논문에서는 차분진화(differential evolution(DE)) 알고리즘을 이용하여 통계적 학습이론 중에서 군집화에 사용되는 SVC의 커널모수와 정규화상수의 효율적이고 객관적인 결정 알고리즘을 제안한다. 즉, SVC의 주관적 모수결정을 차분진화를 이용한 해공간 탐색을 통하여 객관화 하였다. 본 논문에서는 이를 DESVC(differential evolution based Support vector clustering)라고 한다. 제안된 방법의 성능평가를 위하여 UCI machine learning repository와 시뮬레이션 데이터생성으로부터의 객관적인 학습데이터 집합을 이용하였다.

접수일자 : 2007년 5월 28일
완료일자 : 2007년 9월 25일

2. 관련 연구

2.1 차분진화

Rainer와 Kenneth에 의해 제안된 차분진화(DE, deferential evolution) 알고리즘은 간단한 구조를 갖으면서도 전역 최적해에 대한 수렴성이 뛰어나고 다른 진화 알고리즘들에 비해서 컴퓨팅 시간이 짧은 장점을 지닌다[6-7,16]. DE는 해집단 기반의 병렬탐색을 수행한다. 다른 진화알고리즘과는 달리 차분진화는 확률분포에 기반한 변이연산자(mutation operator)에 의존하지 않는다. 차분진화에서 사용되는 연산자는 주로 임의로 선택된 개체들 사이의 차이를 사용한다. 일반적으로 실수값을 갖는 문제해결에 사용되는 차분진화는 두 벡터의 차이와 또 다른 벡터와 가중합을 계산한다. 다음은 차분진화 알고리즘의 절차이다.

(1) 해집단의 초기화
(2) 목표벡터의 임의선택, (x_1,x_2,x_3)
(3) 가중된 차이벡터의 생성, $r(x_1-x_2)$
(4) x_3를 $r(x_1-x_2)$와 더하여 시행벡터 v를 생성,

$$v=x_3+r(x_1-x_2)$$

(5) 시행벡터와 현재벡터 x_i와 교차비 CR에 의해 교차연산 수행
(6) x_i의 교체 또는 유지

위의 절차에서 초기에 해집단을 생성한다. 초기에 결정된 집단의 크기는 진화과정에서 변화지 않는다. 해집단으로부터 3개의 목표벡터(target vector)을 임의추출한다. 차분진화는 두 개의 개체벡터의 차이에 가중치 r를 곱한 것을 나머지 한 개의 개체 벡터에 더해서 교차용 시행벡터(trial vector)를 생성한다. 시행벡터는 현재벡터와 교차비(CR, crossover rate) 값에 의해 교차연산을 수행한다. 최종적으로 x_i의 대체(exchange) 혹은 유지(retain)을 결정한다.

2.2 Support Vector Clustering

Vapnik이 제안한 통계적 학습이론은 분류, 예측, 그리고 군집을 위한 분석모형인 SVM, SVR, 그리고 SVC를 가진다. SVR과 SVC는 모두 SVM에서 확장되었다. SVM에서 목표변수(target variable) y와 입력벡터(input vector) x로 구성된 데이터집합 D는 다음과 같은 표현된다[3,17].

$$(x_i,y_i)_{i=1}^{l},\quad x_i\in R^N,\quad y_i\in\{-1,1\} \tag{1}$$

일반적으로 주어진 입력공간(input space)에서 서로 다른 클래스를 정확히 분류하는 초평면(hyperplane)을 찾는 것은 매우 제한적이다. 이러한 문제를 해결하기 위하여 SVM에서는 입력공간을 더 높은 차원의 특징공간(feature space)으로 사상(mapping)시키고, 특징공간에서 최적의 초평면을 찾는다. $z=\psi(x)$를 입력공간 R^N에서 특징 공간 Z로의 사상 ψ를 갖는 특징 공간 벡터로 표현하면, (w,b)의 쌍으로 이루어진 다음의 초평면을 구해야 한다.

$$w\cdot z+b=0 \tag{2}$$

위의 초평면식을 구하면 다음 함수에 의해 각 x_i를 분류한다.

$$\begin{cases}(w\cdot z_i+b)\geq 1, & \text{if } y_i=1\\ (w\cdot z_i+b)\leq -1, & \text{if } y_i=-1\end{cases}\quad i=1,2,\cdots,l \tag{3}$$

선형 분류가능(linearly separable) 집합 D는 이진 클래스를 갖는 학습 데이터의 사영(projection)들 사이의 마진(margin)을 최대화 하는 유일한 초평면을 구한다. S가 선형분류가능이 아니면 음이 아닌 여유변수(slack variable) ξ_i를 도입하여 다음과 같이 식 (3)을 일반화한다.

$$y_i(w\cdot Z_i+b)\geq 1-\xi_i,\quad i=1,2,\cdots,l \tag{4}$$

식 (4)에서 ξ_i는 식 (3)을 만족하지 않는 x_i이다. $\sum_{i=1}^{l}\xi_i$는 오분류(misclassification)의 양을 나타내는 측도로서 고려된다. 따라서 최적 초평면을 구하는 문제는 아래의 문제에 대한 해(solution)가 된다.

$$\begin{aligned}&\text{minimize}\ \frac{1}{2}w\cdot w+C\sum_{i=1}^{l}\xi_i\\ &\textit{subject to}\ \ y_i(w\cdot z_i+b)\geq 1-\zeta_1\end{aligned} \tag{5}$$

정규화상수 C는 조정모수(control parameter)이다. 이 상수값의 조정을 통하여 모형의 정확성(accuracy)과 복잡성(complexity) 사이의 균형을 맞출 수 있다. 식 (5)에서 최적 초평면을 찾는 것은 다음의 라그랑지 변환(Lagrangian transformation)을 통하여 풀 수 있다.

$$\begin{aligned}&\textit{maximize}\ W(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j z_i\cdot z_j\\ &\textit{subject to}\ \sum_{i=1}^{l}y_i\alpha_i=0\qquad 0\leq\alpha_i\leq C,\quad i=1,2,\cdots,l\end{aligned} \tag{6}$$

여기서 $\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_l)$는 식 (4)의 제약조건과 관련된 비음 라그랑지승수(multiplier) 벡터이다. 정규화상수 C와 함께 SVM에서 중요하게 고려되는 것은 데이터공간에서 형상공간으로의 사상(mapping)을 담당하는 커널함수이다. 일반적으로 SVM에서는 다음 표와 같이 다항학습기계(polynomial learning machine(PLM)), 방사기저함수(radial basis function(RBF)), 그리고 2층 퍼셉트론(multi layer perceptron(MLP))을 사용한다[9].

표 1. SVM의 커널함수

커널종류	함수식	커널모수
PLM	$(x^t x_i+1)^p$	p
RBF	$e^{\left(-\frac{1}{2\sigma^2}\|x-x_i\|^2\right)}$	σ^2
MLP	$\tanh(\beta_1 x^t x_i+\beta_0)$	β_0,β_1

위 표는 SVM의 커널함수와 함수식 그리고 각 커널함수에서 사용되는 커널모수를 나타내고 있다. PLM에서는 다항함수의 차수 p가 커널모수이고 RBF에서는 분산 σ^2가 커널모수가 된다. MLP에서는 편이(bias) β_0와 기울기 β_1이 커널모수이다.

비선형회귀(nonlinear regression) 문제를 해결하기 위하여 사용되는 SVM을 SVR이라 한다. SVR은 SVM과 같은 이론구조를 가지며 추가적으로 다음과 같은 ϵ-insensitive 손실함수(loss function)를 사용한다[19].

$$L(d,y)=\begin{cases}|d-y|-\epsilon, \text{for} |d-y|\geq\epsilon \\ 0 \qquad\quad , o.w.\end{cases} \tag{7}$$

통계적 학습이론을 이용한 군집화 알고리즘인 SVC는 SVR과 마찬가지로 SVM에 기반하기 때문에 주어진 학습데이터의 데이터 점들은 가우시안 커널(Gaussian kernel)에 의해 고차원의 특징공간으로 사상된다. 이 공간에서 주어진 데이터 점들을 그룹화 할 수 있는 최소 경계구면(minimal enclosing sphere)을 찾는다. 이 구면은 각 데이터점이 고차원의 특징공간에서 다시 주어진 데이터공간으로 사상될 때 데이터 점들의 분리된 군집을 결정할 수 있는 몇 개의 집단을 구분해 준다. 다음 그림은 SVC의 군집화 과정을 간단히 보여준다.

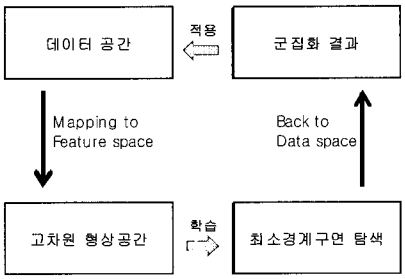

그림 1. SVC 절차

SVC에서 $(x_i)_{i=1}^{l}$가 l개의 데이터 집합일 때 데이터 공간 X로부터 고차원 형상공간으로의 비선형변환(nonlinear transformation) 사상 Φ를 이용하여 반지름 R의 최소 경계구면을 구한다. 이것은 다음과 같은 제한조건을 갖는다.

$$\|\Phi(x_j)-a\|^2 \leq R^2 \quad \forall_j \tag{8}$$

위식에서 a는 구면의 중심이며 여유변수(slack variable) ξ_j를 추가하여 다음 식과 같이 나타낼 수 있다.

$$\|\Phi(x_j)-a\|^2 \leq R^2+\xi_j, \ \xi_j \geq 0 \tag{9}$$

SVM과 마찬가지로 이 문제도 랑그랑지 방법을 사용하여 해결된다.

SVM, SVR, 그리고 SVC는 각각 분류, 예측, 그리고 군집화에 우수한 결과를 제공하지만 정규화상수 C와 커널모수의 선택이 분석가의 사전경험에 의해 주관적으로 결정되기 때문에 이들의 객관적인 최적 선택작업이 필요하다. 따라서 본 논문에서는 차분진화를 이용하여, 특히 SVC를 위한 최적의 정규화상수와 커널모수를 결정하는 알고리즘을 제안한다.

3. 차분진화를 이용한 SVC

이 절에서는 SVC 학습과정에서 사용자에 의해 주관적으로 결정되는 가우시안 커널모수와 정규화상수의 효율적이고 객관적인 결정을 위한 차분진화 기반의 SVC인 DESVC를 제안한다. DESVC에서는 SVC 학습과정에서 사전에 결정해야 하는 커널모수와 정규화상수를 차분진화에 의한 해공간 탐색을 통하여 결정한다. 차분진화를 통하여 결정된 커널모수와 정규화상수를 사용한 DESVC는 SVC를 포함한 기존의 학습 알고리즘들에 비해 향상된 성능을 제공한다.

3.1 DESVC의 적합도함수

일반적으로 군집화는 서로 유사한 개체들끼리 묶는 탐색적 자료분석(exploratory data analysis) 과정이다. 즉, 같은 군집내(within group)의 개체들의 분산은 최소가 되고, 군집간(between groups)의 분산은 최대가 되는 군집화 결과를 얻고자 한다. 본 논문에서는 DESVC의 적합도함수(fitness function)를 만들기 위하여 다음과 같이 군집화의 특성을 사용하였다.

$$Fit_{clusters}=f(Min(v_{within})+Max(v_{between})) \tag{10}$$

위 식에서 v_{within}과 $v_{between}$은 각각 군집내 개체들의 분산과 군집간 분산을 나타낸다. 따라서 DESVC의 적합도함수는 이와 같이 v_{within}은 최소(Min)가 되고, $v_{between}$은 최대(Max)가 되는 구조로 이루어진다. 다음은 DESVC를 위하여 본 논문에서 제안하는 적합도함수이다.

$$Fit_{DESVC}=\frac{1}{G}\sum_{i=1}^{G}v_i+\frac{G}{\log_{10}(N)} \tag{11}$$

위 식은 식 (10)을 이용하여 군집간의 분산은 최대로 그리고 군집내의 분산은 최소로 계산되는 식으로 정의하였다. G는 군집수이고, v_i는 i번째 군집에 속하는 개체들의 분산이다. 또한, N은 전체 학습데이터에 속하는 개체들의 총 수이다. 즉, DESVC를 위한 적합도함수 Fit_{DESVC}의 첫 번째 항은 G개 군집의 각각에 대한 분산들의 평균이다. 이 값이 작을수록 군집화 결과는 식 (10)에 근거하여 우수하다고 할 수 있다. 두 번째 항은 군집수를 전체 데이터수의 로그값으로 나눈 값이다. 이 값의 의미는, 예를 들어, 군집수를 전체 데이터 수만큼으로 결정하게 되면 식 (11)의 첫 번째 항의 값이 0이 되는 것에 대한 패널티(penalty) 항이다. 왜냐하면 군집수가 너무 크게 되면 군집화의 의미가 없기 때문이다. 즉, 두 번째 항은 군집수의 증가 속도를 줄이는 역할을 한다. 본 논문의 DESVC에서는 식 (11)의 적합도함수 값이 작을수록 좋은 군집화로 결정되고, 이에 맞춰 SVC의 커널모수와 정규화상수가 결정된다.

3.2 DESVC 알고리즘

본 논문에서 제안하는 DESVC는 최적의 커널모수와 정규화상수를 결정하는 DE단계와 이 단계의 결과를 이용하여 최적의 군집화를 수행하는 SVC단계, 그리고 마지막으로 최종 군집을 할당하는 Assign Clusters 단계로 이루어진다.

Step I (DE)
[I-1] Let $s=0$
Initializing CR and r
Initializing population X_s with P individuals
[I-2] Do
Selecting randomly $n_1,n_2,n_3 \sim \{1,...,P\}$,
with $n_1 \neq n_2 \neq n_3 \neq n$
Selecting randomly $i \sim \{1,...,I\}$
I : the number of parameters
Repeating
Do $j=1$ to I
if($U(0,1) < CR$ or $j=i$)

$$O_{s,nj} = X_{s,n_3 j} + r(X_{s,n_1 j} - X_{s,n_2 j})$$

else $O_{s,nj} = X_{s,nj}$

End Do

Selecting new population X_{s+1}

$$X_{s+1,n} = \begin{cases} O_{s,n} & \text{, for } Fit_{DE-SVC}(O_{s,n}) \le Fit_{DE-SVC}(X_{g,n}) \\ X_{s,n} & \text{, } o.w. \end{cases}$$

Until (Convergence)

Step I에서는 차분진화를 위한 교차확률 CR과 두 개의 벡터차에 대한 가중치 r을 결정하고 정규화상수와 커널모수들을 위한 후보해들(candidate solutions)로 이루어진 해집단(population)을 초기화한다. 3개의 교배용 벡터를 이용하여 시행벡터를 구하고 이를 현재벡터와 교배한 후에 현재벡터와 적합도 함수에 의해 비교한 후 최종적으로 생성되는 해를 선택한다. 이와 같은 수렴과정을 반복하여 최적의 정규화상수와 커널모수들을 결정하여 다음의 SVC 단계를 수행한다.

Step II (SVC)

[II-1] Mapping data points into feature space by Gaussian kernel function

$K(x_i, x_j) = e^{-q\|x_i - x_j\|^2}$

[II-2] Computing the minimal radius enclosing sphere

[II-3] Mapping back into input space

[II-4] Determining cluster boundaries using contour set

Step II에서는 Step I으로부터의 정규화상수와 커널모수들을 이용하여 SVC 군집화를 수행한다. 가우시안 커널함수를 통한 최종 군집화 결과를 이용하여 다음의 Assign clusters 단계를 수행한다.

Step III (Assign Clusters)

[III-1] Computing similarities between centers of clusters

$$\sum_{k=1}^{n}\sum_{l=1}^{m} \frac{K(s_{ik}, s_{jl})}{m \cdot n}$$

where,

$center_i = \{s_{i1}, \cdots, s_{\in}\}, center_j = \{s_{j1}, \cdots, s_{jm}\}$

$s_i, s_j \in$ {support vectors}

[III-2] Assigning data point x as the following

$$\sum_{k=1}^{n} \frac{K(x, s_{ik})}{n}$$

Step III 단계를 통하여 개개의 개체들에 대한 최적의 군집할당을 개체-군집간 거리(distance) 측도를 이용하여 수행한다.

4. 실험 및 결과

본 논문에서 제안하는 DESVC의 성능평가를 위하여 UCI machine learning repository의 기계학습 데이터와 설명변수들 간의 상관계수의 크기에 따른 시뮬레이션 데이터를 이용하여 기존의 기계학습 알고리즘들과 비교하였다.

UCI 기계학습 데이터 저장소로부터 Abalone data와 Cardiac Arrhythmia Database를 이용하였고 인공으로 생성된 시뮬레이션 데이터는 10개의 변수(attribute)와 1000개의 개체수를 갖는 구조로 연관성의 강도에 따라 3개의 서로 다른 데이터 집합들을 사용하였다[13,15,18]. 다음 표는 본 논문의 실험에 사용된 데이터 집합들의 변수의 개수와 개체수, 그리고 목표변수(target variable)의 클래스의 수를 나타내고 있다.

표 2. 실험에 사용된 학습데이터

data set	# of attributes	# of instances	# of class
Simulated	10	10000	5
Abalone	8	4177	29
Cardiac	279	452	13

위 학습데이터 중에서 시뮬레이션 데이터는 다음 표와 같이 변수간의 상관계수(corelation coefficient)에 따라 3가지 데이터 집합으로 구분하여 생성하였다[13,15]. 이는 설명변수들 간의 연관성의 정도에 따라 모형의 성능이 어떻게 나타나는지를 확인하기 위함이다.

표 3. 시뮬레이션 데이터

Simulated data	indep.	low	high
corelation coefficient	0	0.1-0.3	0.6-0.9

즉, 변수들 간의 상관관계(correlation)가 높은 경우(high)와 낮은 경우(low), 그리고 서로 독립인 경우로 나누어 표 3의 상관계수(correlation coefficient) 값에 의하여 실험하였다.

제안하는 DESVC와의 성능비교를 위한 기존의 분석도구들(analytical tools)로서는 현재 기계학습에서 널리 사용되고 있는 SVC, K-means clustering, SOM(self organizing maps), 가우시안 혼합모형(Gaussian mixture model), 그리고 K-nearest neighbor를 이용하였다[8,9,11,14]. 다음 표는 이들 분석도구들과 본 논문에서 제안하는 DESVC와의 성능평가 결과를 나타내고 있다.

표 4. 오분류율(%) 비교

	Simulated			Abalone	Cardiac
	indep.	low	high		
DESVC	0.9	1.2	1.7	2.8	1.6
SVC	1.5	2.1	2.9	3.4	2.5
K-means	1.9	2.8	3.3	5.0	3.1
SOM	2.3	2.5	3.9	9.2	3.8
Gaussian	2.1	2.5	3.4	8.4	2.9
K-nearest	2.7	3.1	3.7	6.7	3.9

위의 표는 비교되는 모형들 간의 성능평가를 위하여 오분류율(misclassification rate)을 사용하였다[8]. 분석결과를 통하여 DESVC의 성능이 다른 비교모형들에 비해 향상된 결과를 보여주고 있음을 확인할 수 있었다. 또한 시뮬레이션 데이터분석결과에서도 연관성이 높은(high) 데이터와 독립(indep.)인 데이터 간의 오분류율 차이도 다른 기계학습 알고리즘들에 비해 작음을 확인할 수 있었다. 즉, DESVC는 입력변수들 간의 연관성에도 다른 기법들에 비해 로버스트(robust)함을 알 수 있었다. 다음 표는 DESVC와 비교되는 알고리즘들 간의 컴퓨팅시간에 대한 결과이다.

표 5. 컴퓨팅시간(second) 비교

	Simulated	Abalone	Cardiac
DESVC	29.5	19.8	30.2
SVC	22.5	15.5	24.5
K-means	19.5	13.5	21.9
SOM	22.7	14.9	23.0
Gaussian	15.3	12.7	19.2
K-nearest	16.6	12.9	21.4

DESVC는 다른 분석기법들에 비해 진화검색의 과정이 추가되지만 컴퓨팅시간에 대한 비용은 상대적으로 크지 않은 것을 확인할 수 있다. 시뮬레이션 데이터는 3개의 데이터 집합에 대한 컴퓨팅시간들에 대한 차이가 크지 않아 평균 시간을 나타내었다.

5. 결론 및 향후 연구과제

본 논문에서 제안한 DESVC는 Vapnik의 통계적 학습이론 기반의 군집화 알고리즘인 SVC의 정규화상수와 커널모수의 최적 결정을 위하여 차분진화를 적용하였다. 실험결과를 통하여 모형의 성능이 기존의 알고리즘들에 비해 향상되었음을 확인할 수 있었다. 컴퓨팅시간은 정규화상수와 커널모수를 포함하는 해집단을 탐색하는 차분진화의 연산과정이 추가되었음에도 상대적으로 많이 걸리지 않았음을 알 수 있었다. 향후 연구과제로는 붓스트랩(bootstrap)의 표본추출(sampling)을 이용하여 기존의 DESVC의 성능은 유지하면서도 학습데이터의 크기를 축소하여 DESVC의 컴퓨팅 비용을 줄이는 연구가 이루어질 수 있다. 또한 차분진화 알고리즘을 다른 통계적 학습이론인 SVM과 SVR에 적용하여 이들 알고리즘에서 필요로 하는 초기 모수들의 최적결정방안에 대한 연구도 진행될 것이다.

참 고 문 헌

[1] 최병인, 이정훈, "Support Vector Machines를 이용한 Convex 클러스터 결합 알고리즘", 한국퍼지 및 지능시스템학회 2002 추계학술대회 논문지 pp. 267-270, 2002.

[2] 최준혁, 전성해, 오경환, "통계적 학습이론을 이용한 최적군집화", 한국퍼지 및 지능 시스템학회 2005 추계 학술대회 논문지, 2005.

[3] A. Ben-Hur, D. Horn, H. T. Siegelmann, V. Vapnik, "Support Vector Clustering" , Journal of Machine Learning Research, vol. 2, pp. 125-137, 2001.

[4] C. J. Burges, "A Tutorial on Support Vector Machine for Pattern Recognition", Data Mining and Knowledge Discovery, vol. 2, no. 2, pp. 121-167, 1998.

[5] J. C. Chiang, J. S. Wang, "A Validity-Guided Support Vector Clustering Algorithm for Identification of Optimal Cluster Configuration", Proceeding of IEEE International Conference on Systems, Man and Cybernetics, pp. 3613-361, 2004..

[6] A. E. Eiben, J. E. Smith, Introduction to Evolutionary Computing, Springer, 2003.

[7] A. P. Engelbrecht, Computational Intelligence An Introduction, Wiley, 2002.

[8] J. Han, M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann, 2001.

[9] S. Haykin, Neural Networks A Comprehensive Foundation, Prentice Hall, 1999.

[10] S. H. Jun, Web Usage Mining Using Evolutionary Support Vector Machine, Lecture Note in Artificial Intelligence, vol. 3809, pp. 1015-1020, 2005.

[11] K. Krishna, K. Narasimha Murty, "Genetic K-means algorithm", IEEE Transactions on Systems, Man and Cybernetics, part B, vol. 29, no. 3, pp. 433-439, 1999.

[12] J. Lee, D. Lee, "An Improved Cluster Labeling Method for Support Vector Clustering", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 3, pp. 461-464, 2005.

[13] W. L. Martinez, A. R. Martinez, Computational Statistics Handbook with MATRAB, Chapman & Hall, 2002.

[14] G. Mclachlan, D. Peel, Finite Mixture Models, John Wiley & Sons, 2000.

[15] S. M. Ross, Simulation, Academic Press, 1997.

[16] R. Storn, K. V. Price, "Differential Evolution-a fast and efficient heuristic for global optimization over continuous spaces", Journal of Global Optimization, vol. 11, pp. 341-359, 1997.

[17] B. Y. Sun, D. S. Huang, "Support Vector Clustering for Multiclass Classification Problems", Proceeding of IEEE Evolutionary Computation Congress, pp. 1480-1485, 2003.

[18] UCI Machine Learning Repository, http://www.ics.uci.edu/~mlearn/MLRepository.html

[19] V. Vapnik, Statistical Learning Theory, John Wiley & Sons, 1998.

저 자 소 개

전성해(Sung-Hae Jun)

1993년 : 인하대 통계학과 (학사)
1996년 : 인하대 통계학과 (이학석사)
2001년 : 인하대 통계학과 (이학박사)
2007년 : 서강대학교 컴퓨터공학과 (공학박사)
2003년~현재 : 청주대학교 바이오정보통계학과 조교수

관심분야 : 진화연산, 통계적학습이론, 신경망
Phone : 043-229-8205
Fax : 043-229-8432
E-mail : shjun@cju.ac.kr