논문 2004-41CI-1-3

패턴분류를 위한 통계적 RBF 모델

(Statistical Radial Basis Function Model for Pattern Classification)

최 준 혁*, 임 기 욱**, 이 정 현***

(Jun-Hyeog Choi, Kee-Wook Rim, and Jung-Hyun Lee)

요 약

인터넷의 발달과 데이터베이스의 구축이 보편화됨에 따라 막대한 양의 데이터 속에서 의사 결정에 필요한 지식을 찾아내는 작업은 결코 쉬운 일이 아니다. 본 논문에서는 대규모 데이터의 효율적인 분석을 위하여 지식의 탐사 이전에 데이터에 대한 축소 작업을 수행하기 위한 효과적인 차원 축소 전략에 의한 패턴분류 기법을 제안한다. 이를 위해 본 논문에서는 통계적 학습 모형인 Support Vector Machine의 VC-dimension에 기반한 RBF 신경망 모형을 제안한다. 기존의 RBF 신경망 모형은 주로 퍼셉트론 모형의 전처리 작업만을 수행하지만 제안하는 신경망 모형은 VD-dimension과 연계한 독자적으로 데이터를 분석할 수 있는 능력을 갖춘 모형을 구축하고 이를 바탕으로 개체들을 정확한 레이블로 분류한다. 기계 학습 데이터를 이용하여 본 논문에서 제안하는 모형의 성능을 비교 평가한 결과 기존의 여러 분류 알고리즘에 비해 우수한 성능을 보임이 실험을 통해 확인되었다.

Abstract

According to the development of the Internet and the pervasion of Data Base, it is not easy to search for necessary information from the huge amounts of data. In order to do efficient analysis of a large amounts of data, this paper proposes a method for pattern classification based on the effective strategy for dimension reduction for narrowing down the whole data to what users wants to search for. To analyze data effectively, Radial Basis Function Networks based on VC-dimension of Support Vector Machine, a model of statistical learning, is proposed in this paper. The model of Radial Basis Function Networks currently used performed the preprocessing of Perceptron model whereas the model proposed in this paper, performing independent analysis on VD-dimension, classifies each datum putting precise labels on it. The comparison and estimation of various models by using Machine Learning Data shows that the model proposed in this paper proves to be more efficient than various sorts of algorithm previously used.

Keywords: RBF model, Pattern classification, Support Vector Model, Neural Networks model

I. 서 론

현재, 우리는 관리하기 어려울 정도로 방대한 양의 정보화 사회속에 살고 있으며, 인터넷의 발달과 데이터베이스의 구축이 보편화됨에 따라 막대한 양의 데이터 속에서 의사 결정에 필요한 지식을 찾아내야 하는 매우 어려운 현실 문제에 직면해 있다.

인공 지능 기반의 많은 정보 추출 시스템들은 대규모 외부 정보를 인간의 두뇌와 같은 구조에서 처리할 수 있도록 하고 있으며, 이러한 인간의 두뇌활동을 모형화하는 대표적인 기법중의 하나가 신경망(neural networks) 기법을 이용하는 방법이다.

본 논문에서는 대규모 데이터의 효율적인 분석을 위하여 지식의 탐사 전에 데이터에 대한 축소 작업을 수행하기 위한 효과적인 차원 축소 전략에 의한 패턴분류(pattern classification) 기법을 제안한다. 제안하는

* 정회원, 김포대학 컴퓨터계열
(Division of Computer Science, Kimpo College)
** 정회원, 선문대학교 지식정보산업공학과
(Knowledge Information & Industrial Engineering Department, Sunmoon Univ.)
*** 정회원, 인하대학교 컴퓨터공학부
(School of Computer Science & Engineering, Inha Univ.)
접수일자 : 2003년4월25일, 수정완료일 : 2003년12월29일

방법은 통계적 학습(statistical learning) 모형인 Support Vector Machine의 VC-dimension에 기반한 Radial Basis Function(RBF) 신경망 모형이다. 기존의 RBF 신경망 모형은 주로 퍼셉트론 모형의 전처리 작업만을 수행하지만 제안하는 신경망 모형은 VC-dimension과 연계하여 독자적으로 데이터를 분석할 수 있는 모형을 구축하고 이를 바탕으로 개체들을 정확한 레이블로 분류하는 것을 가능하게 한다.

Ⅱ. 통계적 학습 기반의 RBF 신경망

대용량의 데이터로부터 직접 패턴분류 모형을 구축하는데 있어 가장 어려운 문제점은 다차원(higher-dimension)의 원인으로부터 발생한다. 전체 데이터의 각 개체의 레이블을 나타내는 목표 변수(target variable)에 대한 입력 변수들이 너무 많이 존재하면 이들 변수들 간의 종속성이 존재하게 되고, 최종의 패턴분류 모형의 입력 변수들 간의 종속성은 모형의 예측력을 떨어뜨리는 다중 공선성(multicollinearity)의 문제를 발생시킨다[2,7]. 따라서, 주어진 학습 데이터에 대한 패턴분류 작업을 수행하기 전에 입력 변수들에 대한 적절한 차원 축소 작업이 선행된다면 더 좋은 분류 결과를 얻을 수 있다. 이러한 차원 축소와 데이터 분류 기능을 동시에 갖고 있는 모형으로서 RBF 신경망 모델이 있다. 이 모형은 자율학습과 지도 학습을 동시에 수행하는 구조를 지니고 있으며 RBF의 자율학습 기능이 데이터에 대한 차원 축소를 담당하고 있다.

2.1 RBF 신경망

1989년에 Moody와 Darken에 의해 제안된 RBF 신경망은 입력층(input layer)과 은닉층(hidden layer) 그리고 출력층(output layer)의 구조로 구성되어 있다. 은닉 노드는 인스타(instar)-아웃스타(outstar) 모델 구조이며, 입력층과 은닉층에서는 자율학습(unsupervised learning) 방법인 경쟁 학습에 의해 각 은닉 노드의 평균과 표준 편차가 결정된다. 은닉층과 출력층에서는 지도 학습(supervised learning) 방법인 일반화된 델타 학습에 의해 연결 가중치가 결정되는 하이브리드 망(hybrid network) 구조이며, 은닉층의 처리 요소는 정규화된 가우시안(Gaussian) 함수를 갖는다. 가우시안 함수는 비선형적인 입력들에 대하여 입력 공간중의 하나인 좁은 영역에 대해서만 뛰어난 반응을 보인다. RBF 신경망은 역전파 신경망처럼 지도 학습에 따라 은닉층 노드값을 결정하도록 가중치를 조절할 수 있지만, 자율학습 방법에 따른 접근이 더 나은 결과를 생성함으로서 하이브리드 자율학습 방식으로 가중치가 결정된다.

RBF 신경망은 정확도의 측면에서 오류 역전파 신경망(backpropagation)보다 성능이 떨어지지만 3개 계층만으로 구성되어 있기 때문에 일반적인 MLP (Multi-Layer Perceptron)의 역전파 모형보다는 훨씬 쉽고 빠른 학습이 수행된다.

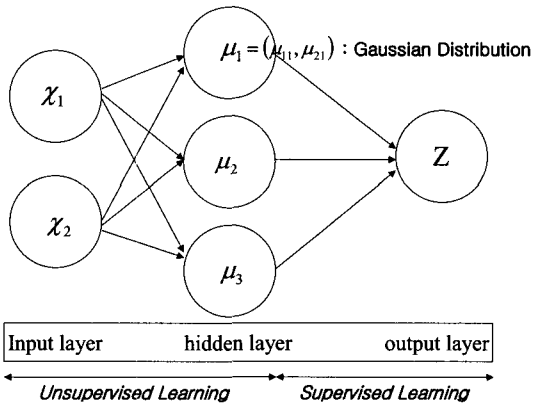


그림 1. RBF 신경망의 기본 구조
Fig. 1. RBF Architecture

그림 1은 2개의 입력 노드와 3개의 은닉 노드, 그리고 1개의 출력 노드로 구성된 RBF 신경망의 기본 구조를 나타내고 있다. 하나의 입력패턴이 존재하는 가수용 영역들의 경계를 벗어날 경우 기존의 은닉층은 그대로 유지하면서 이 입력패턴을 포함하는 새로운 가수용 영역 하나를 별도로 추가하는 것이 가능하다. 이러한 노드 성장 능력은 은닉층이 필요로 하는 최소 노드 수를 결정하는 문제를 혁신적으로 해결한 것으로 볼 수 있다.

RBF 신경망의 주어진 학습 데이터 구조는 입력값과 출력값의 쌍, $(x_i, y_i)\,(i = 1,...,n)$으로 구성되어 있다. RBF의 각 은닉 노드는 정규화된 가우시안 활성 함수를 갖고 있으며, 은닉 노드 q는 평균 m_q에 가장 가까운 입력 벡터에 대해 최대의 반응값을 보인다. 은닉 노드 q에 대한 출력 z_q는 식 (1)과 같이 정의할 수 있다.

$$z_q = \exp\left[-\frac{(x-m_q)^2}{2\sigma_q^2}\right] \quad (1)$$

식 (1)은 각 은닉 노드가 입력공간 안에서 평균 m_q를 중심으로 σ_q 만큼의 편차를 유지하는 가우시안 분포에서 입력 x에 대한 자신만의 가수용 영역(receptive field) $R_q(x)$를 갖고 있음을 의미하는 것으로 σ_q^2는 q의 분산을 나타낸다. 가수용 영역들은 특정 입력에 대해 그 중 하나만이 가장 잘 반응하는 특성을 갖고 있어야 한다. 이런 역할은 각 가수용 영역의 표준편차가 담당하는데, 입력 공간상에서 단절됨이 없이 적당한 중첩성을 유지하는 것이 RBF 신경망의 자율학습 부분이다.

2.2 SV-RBF 모형

기존의 RBF 모형이 Vapnik의 Support Vector Machine과 결합되면 식 (2)와 같은 구조를 갖는데, 이러한 모형을 SV-RBF(Support Vector machine based Radial Basis Function) 모형이라 한다[11].

$$f(x)=sign\left(\sum_{i=1}^{N}a_iK_\gamma(|x-x_i|)-b\right) \tag{2}$$

식 (2)에서 $|x-x_i|$는 두 벡터 간의 거리이고, a_i와 b는 추정해야 할 모수인 가중치이다. $K_\gamma(|\cdot|)$는 두 벡터간의 거리에 의존하는 함수로서, 보통은 식 (3)과 같은 가우시안 함수를 사용한다[8].

$$K_\gamma(|x-x_i|)=\exp\{-\gamma|x-x_i|^2\} \tag{3}$$

식 (3)의 모형을 구축하기 위해서는 x_i들의 수 N과 x_i, a_i, γ의 4개의 모수를 결정해야 한다. 이 4가지 모수 중에서 N과 x_i, a_i는 경험적으로 결정하지만, γ는 경험적 위험 범함수(empirical risk functional)를 최소화하는 모수로 결정한다. 그러므로 내적을 생성하는 커널함수로서 $K_\gamma(|x-x_i|)$를 선택하고, 이 커널을 이용하여 SV-RBF 모형을 생성한다. 기존의 RBF에 비해 이 모형은 4개의 모수를 자동적으로 결정할 수 있으며, 이 결정방법은 표 1과 같은 방법에 의해 결정할 수 있다.

표 1. SV-RBF의 모수 결정 방법
Table 1. Parameter determination method of SV-RBF

모 수	결정 방법
N	Support Vector의 개수
x_i	Support Vector
a_i	확장 계수, $a_i=\alpha_i y_i$
γ	범함수의 최소화

이 모형은 SVM에 기반하고 있기 때문에 이진 분류 데이터에는 우수한 성능을 보이지만 3개 이상의 다중 패턴에 대한 분류에는 한계를 보인다[13]. 따라서 이진 분류뿐만 아니라 3개 이상의 모든 패턴의 데이터에 대한 분류에서도 좋은 분류 결과를 보일 수 있는 모형이 필요하다.

본 논문에서는 이러한 문제를 해결하기 위해 통계적 학습 기반의 RBF 모형인 SL-RBF(Statistical Learning based Radial Basis Function)를 제안한다. 본 논문에서 제안하는 변형된 RBF 모형을 SL-RBF라고 명명한 이유는 SV-RBF가 이진 패턴의 데이터 분류에 특히 우수한 성능을 보이는 것에 반해, SL-RBF는 모든 다중 패턴 데이터에 대하여 안정적인 분류 결과를 제시한다.

Ⅲ. 통계 학습 기반의 RBF 모형

1980년대 초반 Vapnik이 처음 제안한 SVM은 여러 통계적 학습 이론 중에서도 매우 우수한 패턴분류 알고리즘으로 알려져 있다[1]. 특히, 이진패턴의 데이터에 대한 분류에 있어서는 기존의 다른 기계 학습 알고리즘들에 비해 매우 좋은 성능을 보임이 현재까지의 많은 연구들을 통하여 확인되어 왔다[3]. 하지만 SVM에서 사용하는 커널함수의 특성 때문에 패턴 수가 3개 이상인 데이터에 대한 분류 문제에는 적용하기가 쉽지 않다. 왜냐하면 SVM의 커널함수들은 이진 분류에 대한 작업만 수행하도록 설계되어 있기 때문이다. 따라서 3개 이상의 다중 패턴분류 문제에 대해 SVM을 적용하기 위해서는 반복적인 이진 분류 작업을 통해서 해결 가능하게 된다. 이러한 SVM의 문제점을 해결하기 위하여 Zhu는 3개 이상의 부류를 갖는 데이터의 패턴분류를 위하여 Import Vector Machine(IVM)을 제안하였다[13].

본 논문에서는 RBF 커널함수를 IVM에 적용하여 가장 성능이 우수한 통계적 학습 이론 기반의 분류 모형을 구축하게 된다. 제안 모형에서 사용되는 RBF 커널함수는 은닉층이 한 개인 다층 신경망 구조를 띠고 있으며, 은닉층의 결합 함수로는 식 (4)의 원형 기준 함수를 사용한다.

$$H_i=exp\left(-\frac{(x_1-c_{1i})^2+\cdots+(x_k-c_{ki})^2}{r_i^2}\right) \tag{4}$$

RBF 커널 신경망에서 추정해야 할 모수의 수는 다층 신경망 모형과 동일하지만 그 추정 방법은 다소 다르다. 식 (4)에서 은닉 마디의 개수를 k라 할 때, 먼저 k개의 중점 $c_{1i}, c_{2i}, \ldots, c_{ki}$와 반경 $r_i(i=1,2,\ldots,k)$를 군집분석과 같은 자율학습 방법에 의해 추정하고, 다음으로 은닉층으로부터 출력층으로 전달되는 계수를 지도학습 방법에 의해 추정한다. 즉, 은닉층에서의 모수를 먼저 추정하고 그 다음에 출력층에서의 모수를 추정하는 2단계 추정 방법을 사용한다. 두 개의 패턴을 갖는 데이터를 가장 잘 분류하는 기존의 SVM에 비하여 IVM은 2개 이상의 레이블을 갖는 데이터에 대한 패턴분류를 통계적 학습이론에서 가능하게 해주는데, 두 개 이상의 클래스를 가장 잘 대표하는 Support Vector들을 이용하여 동시에 가장 잘 분리하는 분류평면을 구함으로써 최적 이상 평면(optimal hyper-plane)을 효과적으로 구분할 수 있게 된다.

SVM은 식 (5)와 같은 분류기 구조를 이용하여 이진 패턴을 분류한다.

$$y = Sign[f(x)], \; y_i \in \{-1, 1\} \tag{5}$$

식 (5)에서 y_i는 패턴의 레이블을 나타낸다. $Sign(g)$ 함수는 함수 값이 임계값 이상이 되면 1로, 임계값 이하면 -1로 분류하는 역할을 수행한다. $f(x)$에 기존의 SVM의 커널함수를 사용하지 않고 커널 로지스틱 함수를 사용하게 되면 다중분류가 가능해지는 Zhu의 IVM이 된다[15]. 커널 로지스틱 회귀함수는 SVM의 $(1-yf)_+$대신에 $ln(1+e^{-yf})$의 이항 변형자(binomial deviance)를 사용한다. SVM의 support vector와 유사하게 IVM은 학습 데이터의 fraction을 이용하는데, 이를 IVM에서는 import points라고 부른다.

기존의 연구에서는 전체 학습 데이터 중에서 SVM의 모형 구축에 사용되는 support points 수보다 IVM의 import points의 수가 항상 작기 때문에, 학습시간에 대한 비용측면에서도 IVM이 SVM보다 더 효율적임을 알 수 있다.

임의의 커널함수를 갖는 SVM은 식 (6)으로 표현할 수 있다.

$$\sum_{i=1}^{N}(1-y_i f(x_i))_+ + \lambda \| f \|_{H_K}^2 \tag{6}$$

식 (6)에서, $f = b + h$, $h \in H_K$, 그리고 $b \in R$ 이다. 또한, H_K는 커널함수 K에 의해 생성된 함수들의 Reproducing Kernel Hilbert Space(RKHS)이고, $\| f \|_{H_K}$는 모형의 roughness에 대한 패널티이다. 최종적인 모형은 라그랑지(Lagrange) 확장을 통하여 식 (7)과 같이 $K(x, x_i)$의 유한한 확장으로 표현된다[4,6].

$$\hat{f}(x) = \hat{b} + \sum_{i=1}^{N} \hat{\alpha}_i K(x, x_i) \tag{7}$$

식 (7)에서 $\hat{\alpha}_i$는 라그랑지 확장 계수(Lagrange multiplier)로서, SVM의 커널 전략으로 식 (8)과 같은 KLR(Kernel Logistic Regression)을 사용하게 되면 다중 패턴의 분류가 가능해 진다[5,9].

$$P(Y=1 \mid x) = \frac{e^{f(x)}}{1+e^{f(x)}} \tag{8}$$

본 논문에서는 Zhu의 IVM을 RBF 신경망에 적용하여, 다중 패턴분류를 위한 Import Vector의 결정(voting)을 위한 함수들을 위해 다수의 커널함수를 사용한다. 다수의 커널함수들의 평균을 이용하여 최적의 Import Vector를 결정하게 되며 새로운 다중 패턴 데이터에 대한 분류를 수행한다. IVM은 식 (6)에 KLR을 적용하면 식 (9)와 같이 정의할 수 있다.

$$-\sum_{i=1}^{N}[y_i f(x_i) - \ln(1+\exp(f(x_i)))] + \frac{\lambda}{2}\| f \|_{H_K}^2 \tag{9}$$

식 (9)는 SVM과 마찬가지로 라그랑지 확장을 통하여 모형의 최적 계수를 구하게 된다[12]. 최적의 다중 패턴분류를 위한 SL-RBF 모형의 알고리즘은 다음과 같다.

알고리즘 1의 SL-RBF 모형에서 사용된 커널함수는 Running-mean smoothers, Running medians and enhancements, Equivalent kernels, Regression splines, Cubic smoothing splines, 그리고 Locally-weighted running smoothers이다.

IV. 실험 및 결과

제안 알고리즘에 대한 성능 평가를 위하여 UCI Machine Learning Repository의 Iris 데이터와 Arrhythmia 데이터[14]를 이용하여, 통계적 학습모형에서의 커널함수의 형태에 따른 결과와 오분류율을 계산

알고리즘 1. SL-RBF 모형
Algorithm 1. SL-RBF Model

(step 1) Initialization : 초기 모수 및 모수 공간의 결정

1. $S=\varnothing$, $R=\{x_1,x_2,...,x_N\}$, $k=1$ (S: set of import points)
2. RBF의 원형 함수의 개수 결정(m)

(step 2) Learning and estimation

1. 다음의 식을 최소화하는 모수결정

$$f_l(x)=\sum_{x_j\in SU\{x_l\}}\alpha_j K(x,x_j)$$

2. 자율학습을 통한 RBF의 노드군집의 중심 추정

(step 3) Kernel Bagging and model heuristics

1. 다음의 조건을 만족하는 $x_{l'}$ 를 구한다.
   a) $\arg\min_{x_l\in R} H(x_l)$,
   여기서, $H(x_l)$는 식 (5)이다.
   b) $average[bagging_{k_l\in\{k_1,...,k_M\}} B_{k_l}(H(x_l))]$,
   다수의 커널함수에 대한 bagging 수행
2. 중심 v_j 에 대한 너비모수(width parameter) α_j 결정
   a) 최단 중심 거리 결정

$$r_j=\min_k\|v_k-v_j\| \text{ for all } k\neq j$$

   b) 2-a)를 만족하는 너비모수 결정

$$\alpha_j=\gamma r_j$$

   c) 2-a)와 2-b)를 통한 모형의 가중치 w_j 를 추정

(step 4) Repeat

주어진 조건이 만족될 때까지 (step 2)와 (step 3)를 반복함.

하여 비교하였다. 비선형 관점에서 이들 데이터를 임의의 비선형 집단으로 분리하여 초평면의 형태 및 오분류율을 계산해 보았고, 기존의 분류분석 기법들인 SVM, SV-RBF, 다층 신경망, 그리고 K-NN 등과도 비교하였다.

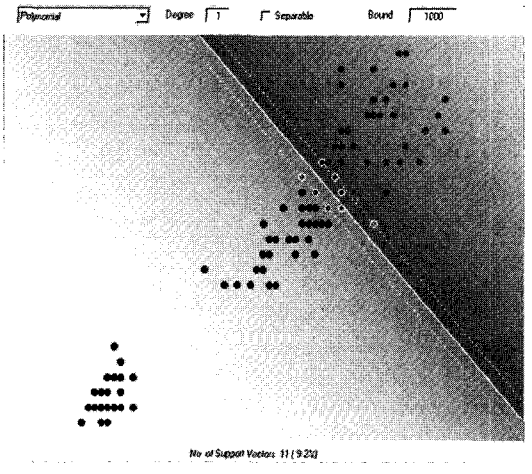
오분류율은 학습 데이터와 테스트 데이터로 구분하여, 학습 데이터로 모형을 구축하고 테스트 데이터를 이용하여 분류예측을 수행한 후에 계산하였다. 학습 데이터는 임의로 60%를 추출하였고 나머지 40%는 검증을 위한 데이터로 사용하여 각 방법마다 150번을 반복적으로 실험하였다. 본 논문의 실험을 위해서는 Matlab Ver. 6.1과 'The R Project for Statistical Computing'의 R 소프트웨어[15]를 사용하였다.

4.1 SVM을 통한 분류 실험

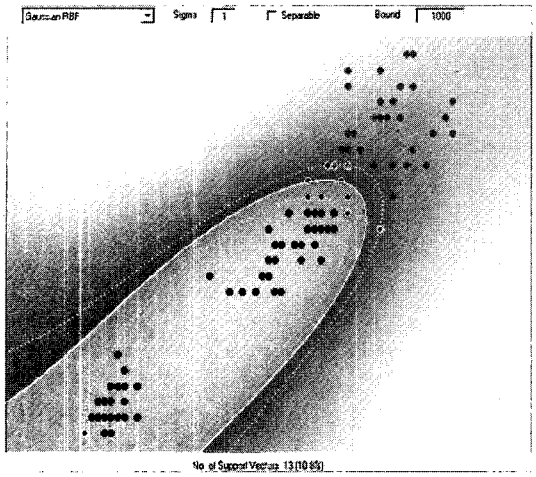
Iris 데이터는 총 4개의 입력변수와 1개의 목표변수로 이루어져 있다. 4개의 입력 변수들은 각각 붓꽃의 외형을 결정하고 있다. 또한 목표변수는 Setosa, Versilcolor, Viginica의 3개의 레이블을 갖고 있으며, 총 데이터의 갯수는 150개이다. 우선, 이진 분류모형인 SVM을 이용한 분류실험을 위하여 Versilcolor, Viginica 두 집단에 대한 분류분석을 실시한 후에 다시 나머지 Setosa에 대한 분류분석을 실시하였다.

그림 2는 선형 커널함수와 가우시안 커널함수를 사용했을 때의 분류 모형에 대한 결과를 나타낸다.

그림 2에서 알 수 있듯이 커널함수의 형태에 따라 분류의 성능은 차이가 나고 있음을 알 수 있는데, 선형 커널함수보다는 가우시안 커널함수가 더 좋은 분류결과를 얻을 수 있음을 알 수 있다. 그림 2의 커널함수에 따른 성능비교 결과는 표 2와 같다. 표 2의 결과 선형 커널함수에 비해 가우시안 커널함수의 마진이 훨씬 작고 모형의 설명력을 나타내는 Support Vector의 개수도 더 많음을 알 수 있다.



(a) Linear



(b) Gaussian

그림 2. 커널함수의 형태에 따른 SVM의 분류 평면
Fig. 2. Classification plane of SVM by Kernel function

표 2. 커널함수에 따른 성능 비교(Iris)
Table 2. Performance according to Kernel function (Iris)

	Linear	Gaussian
Margin	0.1356	0.0492
SupportVector수	11	13

다음으로는 Arrhythmia 데이터를 이용한 실험을 수행하였다. 이 데이터는 279개의 입력변수와 16개의 레이블을 갖고 있는 1개의 목표변수로 이루어 졌다. Iris 데이터와 같은 실험을 수행한 결과가 표 3에 나타나 있는데 실험을 위해 사용한 총 데이터의 개수는 452개이다. 표 3의 결과를 살펴보면 가우시안 커널함수를 이용한 결과가 기존의 선형 커널함수를 이용한 결과보다 분류율이 더욱 좋음을 알 수 있다.

표 3. 커널함수에 따른 성능 비교(Arrhythmia)
Table 3. Performance according to Kernel function (Arrhythmia)

	Linear	Gaussian
Margin	0.1692	0.0538
SupportVector수	19	24

4.2 제안 모형과 기존 분류 모형과의 비교 실험

이번 절에서는 제안하는 SL-RBF 모형과 Gaussian, SV-RBF 모형 그리고 기존의 K-NN, MLP와의 비교 실험을 통해 제안 모형의 분류 성능을 평가한다.

표 4. 5개의 분류 모형의 오분류율 비교
Table 4. Misclassification ratio of five models

모형	Iris	Arrhythmia	Average
Gaussian	0.2132	0.2316	0.2224
SV-RBF	0.1232	0.1501	0.1367
K-NN	0.2942	0.3262	0.3102
MLP	0.5621	0.6138	0.5880
SL-RBF	0.0843	0.1123	0.0983

표 4는 Iris 데이터와 Arrhythmia 데이터에 대한 분류 기법들의 오분류율과 두 데이터의 오분류율에 대한 평균을 나타낸다. 실험결과 본 논문에서 제안한 SL-RBF의 오분류율이 다른 모형에 비해 가장 적음을 알 수 있다.

그림 4는 주어진 학습 데이터에서 임의추출을 통해 서로 다른 표본을 통한 150번의 반복 실험 결과를 나타내는 것으로, x축은 학습데이터에 대한 반복 회수를, 그리고 y축은 테스트 데이터에 대한 오차율을 나타낸다.

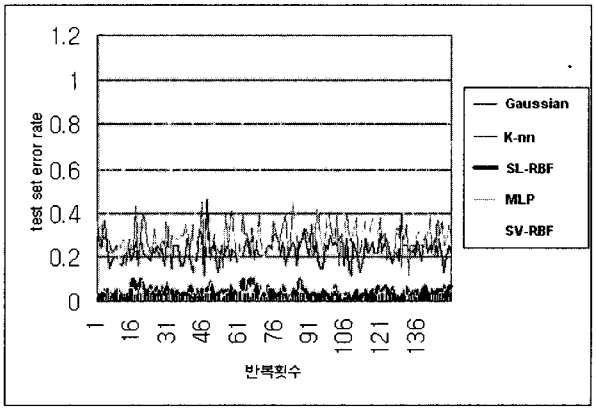


그림 3. 반복 실험을 통한 분류 기법간의 성능 평가
Fig. 3. Performance of classification by repeat

그림 3에서 SL-RBF와 SV-RBF가 다른 분류 기법에 비해 작은 오차율을 보이고 있으며, 또한 SV-RBF보다 SL-RBF 모형이 더 안정적인 실험 결과를 나타내고 있음을 알 수 있다.

V. 결론 및 향후 연구과제

본 논문에서는 이진분류를 포함한 모든 다중 분류 문제의 데이터에 대해 우수한 분류 결과를 얻을 수 있는 통계적 학습이론에 기반한 원형 기준 함수 모형인 SL-RBF 모델을 제안하였다. 본 논문에서 제안한 모형

은 기존의 분류 모형들에 비해 적은 오분류율 결과를 보이고 있으며, 반복된 실험에서도 비슷한 결과를 제공하는 안정된 모형이다. 그러나 한 개의 모형에서 자율학습과 지도학습을 모두 병행하면서 분류를 수행하게 됨으로써 학습에 대한 계산비용 문제가 발생할 수 있다. 향후 연구 과제로는 이러한 계산시간을 단축할 수 있는 방안에 대한 연구가 있어야 할 것이며, 이를 해결하기 위한 연구 방안 중의 하나로 Jackknife 기법 등을 적용하고자 한다.

참 고 문 헌

[1] C. Burges, "A tutorial on support vector machines for pattern recognition," In Data Mining and Knowledge Discovery, 1998.

[2] V. Cherkassky, F. Mulier, "Learning from Data: Concept, Theory, and Methods," John Wiley & Sons, Inc., 1998.

[3] T. Evgeniou, M. Pontil, T. Poggio, "Regularization networks and support vector machines," MIT Press, 1999.

[4] P. Green, B. Yandell, "Semi-parametric generalized linear models," Proceedings 2nd International GLIM Conference, 1985.

[5] G. Kimeldorf, G. Wahba, "Some results on Tchebyc- heffian spline functions," Math. Anal. Applic., 1971.

[6] X. Lin, G. Wahba, D. Xiang, F. Gao, R. Klein, B. Klein, "Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV," Technical Report 998, Department of Statistics, University of Wisconsin, Madison, 1998.

[7] R. H. Myers, "Classical and Modern Regression with Applications," Duxbury, 1990.

[8] M. J. D. Powell, "The theory of radial basis func- tions approximation in 1990," Advances in Numer-ical Analysis Volume II: Wavelets, Subdivision Algorithms and Radial Basis Functions, W. A. Light, ed., Oxford University, pp. 105-210, 1992.

[9] A. Smola, B. Scholkopf, "Sparse Greedy Matrix Approximation for Machine Learning," In Proceedings of the Seventeenth International Conference on Machine Learning, 2000.

[10] V. N. Vapnik, "The Nature of Statistical Learning Theory," New York: Springer-Verlag, 1995.

[11] V. N. Vapnik, "Statistical Learning Theory," New York: Wiley, 1998.

[12] G. Wahba, "Support Vector Machine, Reproducing Kernel Hilbert Spaces and the Randomized," GACV. Technical Report 984rr, Department of Statistics, University of Wisconsin, Madison, 1998.

[13] J. Zhu, T. Hastie, "Kernel Logistic Regression and the Import Vector Machine," NIPS2001 Conference, 2001

[14] http://www.ics.uci.edu/~mlearn/MLSummary.html

[15] http://www.r-project.org/

저 자 소 개

최 준 혁(정회원)

1990년 경기대학교 전자계산학과 졸업(이학사), 1995년 인하대학교 대학원 전자계산공학과 졸업(공학석사), 2000년 인하대학교 대학원 전자계산공학과 졸업(공학박사), 1997년 - 현재 김포대학 컴퓨터계열 조교수, <주관심분야 : 정보검색, 데이터마이닝, 신경망, 유전자 알고리듬 등>

임 기 욱(정회원)

1977년 2월 인하대학교 공과대학 전자공학과 졸업, 1987년 2월 한양대학교 전자계산학 석사, 1994년 8월 인하대학교 전자계산학 박사, 1977년 2월 - 1983년 2월 한국전자기술연구소 선임연구원, 1983년 3월 - 1988년 7월 한국전자통신연구소 시스템소프트웨어 연구실장, 1988년 8월 - 1989년 8월 미 캘리포니아 주립대학(Irvine) 방문연구원, 1989년 10월 - 1996년 12월 한국전자통신연구원 시스템연구부장, 주전산기(타이컴) III, IV개발 사업책임자, 1997년 1월 - 1999년 12월 정보통신연구진흥원 정보기술전문위원, 2001년 7월 - 2003년 2월 한국전자통신연구원 컴퓨터소프트웨어 연구소장, 2000년 3월 - 현재 선문대학교 지식정보산업공학과 교수, <주관심분야 : 실시간데이타베이스시스템, 운영체제, 시스템구조>

이 정 현(정회원)

1977년 인하대학교 전자공학과, 1980년 인하대학교 대학원 전자공학과(공학석사), 1988년 인하대학교 대학원 전자공학과(공학박사), 1979년~1981년 한국전자기술연구소 시스템연구원, 1984년~1989년 경기대학교 교수, 1989년~현재 인하대학교 전자전기컴퓨터공학부 컴퓨터공학전공 교수, <주관심분야 : 자연어처리, HCI, 정보검색, 음성인식, 음성합성, 컴퓨터구조, 홈네트워킹>