

혼합정수 선형계획법 기반의 비선형 패턴 분류 기법

김광수 · 류홍서[†]

고려대학교 산업시스템정보공학과

An MILP Approach to a Nonlinear Pattern Classification of Data

Kwangsoo Kim · Hong Seo Ryou

Department of Industrial Systems and Information Engineering, Korea University

In this paper, we deal with the separation of data by concurrently determined, piecewise nonlinear discriminant functions. Toward the end, we develop a new l_1 -distance norm error metric and cast the problem as a mixed 0-1 integer and linear programming (MILP) model. Given a finite number of discriminant functions as an input, the proposed model considers the synergy as well as the individual role of the functions involved and implements a simplest nonlinear decision surface that best separates the data on hand. Hence, exploiting powerful MILP solvers, the model efficiently analyzes any given data set for its piecewise nonlinear separability. The classification of four sets of artificial data demonstrates the aforementioned strength of the proposed model. Classification results on five machine learning benchmark databases prove that the data separation via the proposed MILP model is an effective supervised learning methodology that compares quite favorably to well-established learning methodologies.

Keywords: machine learning, data mining, mixed integer and linear programming, classification

1. Introduction

패턴 분류(pattern classification)는 두 개 이상의 집합으로 구성된 데이터에서 각 집합의 차별적인 특성을 발견하고 추출하는 데이터마이닝의 기본적인 문제이다. 패턴 분류는 데이터를 구성하는 집합의 개수에 따라 이진 분류(binary classification)와 다범주 분류(multicategory classification)로 나눌 수 있으며, 질병 진단 및 예측(Mangasarian *et al.*, 1995; Wolberg and Mangasarian, 1990), 얼굴 인식(Osuna *et al.*, 1997), 신용 평가(Carter and Catlett, 1997) 등 다양한 분야에 적용될 수 있다.

본 논문에서는 데이터를 구성하는 집합의 개수가 두 개인 이진 분류를 패턴 분류와 동일한 문제로 간주하고 이진 분류를 위한 혼합정수 선형계획법(Mixed Integer and Linear Programming : MILP) 기반의 비선형 패턴 분류 기법을 제시할 것이다. 이는 대부분의 패턴 분류 문제들이 이진 분류 문제에 속하며, 다범주 분류 문제의 경우도 이진 분류의 연속적인 적

용으로 해결될 수 있기 때문이다(Ullman, 1973).

수리계획법(mathematical programming)에 기반을 둔 패턴 분류 기법들 중에서 가장 기본적인 방법은 한 개의 선형 곡면(linear surface)을 이용하여 데이터의 두 집합을 분류(classification)하는 것이다. 이때, 대부분의 선행 연구에서는 선형계획법(Linear Programming : LP)의 해법을 통해 두 집합을 분류하는 선형 곡면을 구축하였다(Bennett, 1993; Bennett and Mangasarian, 1992; Mangasarian, 1993). 하지만, 대부분의 현실 문제들이 한 개의 선형 곡면만으로는 분리(separation)할 수 없는 경우에 속한다는 점에서, 선형 의사결정 곡면(linear decision surface)을 이용한 패턴 분류는 매우 제한적일 수밖에 없다. 따라서 비선형 곡면을 이용한 패턴 분류 기법을 고려하게 되는데, Mangasarian(1965)은 비선형 곡면 중에서 그 곡면의 파라미터(parameter)들이 선형을 이루는 경우에는 선형계획법의 해법을 통해 비선형 의사결정 곡면을 구축할 수 있음을 제시하였다.

이 논문은 2005년도 정부재원(교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음(KRF-2005-003-D00445).

[†] 연락처자 : 류홍서 교수, 136-713 서울시 성북구 안암동 5가 고려대학교 산업시스템정보공학과, Tel : 02-3290-3394,

E-mail : hrsyoo@korea.ac.kr

2005년 12월 접수, 2005년 12월 수정본 접수, 2005년 12월 게재 확정.

Multisurface Method(MSM)는 선형 곡면 혹은 비선형 곡면 한 개를 이용하여 두 집합을 분리하려는 기존 연구와는 달리 두 개 이상의 선형 곡면을 사용하는 방법이다(Mangasarian, 1968). MSM은 데이터의 두 집합이 완전히 분리될 때까지 LP 모델을 반복적으로 적용하여 평행한 선형 곡면들을 순차적으로 구하고 그것들을 조합하여 조각별 비선형 곡면(piecewise nonlinear surface)을 얻는 방법이다. 이러한 MSM은 선형 분리가 불가능한 복잡한 데이터들도 선형계획법의 해법만을 이용해서 100% 분리할 수 있도록 한다는 점에서 기존의 방법들에 비해 장점이 있지만, 순차적으로 구해지는 평행한 선형 곡면들이 탐욕 전략(greedy rule)을 따라 구해지기 때문에 과도하게 복잡한 의사결정 곡면을 구축하여 일반화 능력(generalization ability)이 떨어진다는 단점도 가지고 있다(Nakayama and Kagaku, 1998).

이상에서 언급한 선형 연구들은 선형계획법 기반의 패턴 분류 기법으로서 한 번에 한 개의 혹은 한 쌍의 평행한 곡면을 이용할 수밖에 없다는 한계를 지니고 있다. 즉 두 개 이상의 독립적인 곡면들을 동시에 사용할 수 없다는 공통점이 있다. 선형 연구들이 이러한 공통점을 갖는 이유는 두 개 이상의 곡면을 동시에 사용하여 의사결정 곡면을 구축하는 문제가 비볼록 최적화(nonconvex optimization) 해법이 요구되는 NP-난도문제에 속하기 때문이었다(Falk and Lopez-Cardona, 1997; Megiddo, 1998). 그런 의미에서 두 개의 선형 곡면을 동시에 이용한 Bennett and Mangasarian(1994)의 겹선형 분리(bilinear separation)는 기존의 선형계획법을 이용한 탐욕 접근법(greedy approach)과는 근본적으로 다르다. 이 연구에서 그들은 겹선형 분리 문제를 세 종류의 겹선형계획법(Bilinear Programming : BP) 기반의 모델로 구축하였고, 그로부터 두 개의 선형 곡면으로 구성되는 의사결정 곡면을 구하였다. 하지만 겹선형계획법은 특별히 어려운 문제로 알려진 곱셈 계획법(Multiplicative Programming : MP)에 속하여 해법이 어렵기 때문에(Ryoo and Sahinidis, 2003), 그들은 겹선형계획법 모델에 대한 해법으로 일종의 휴리스틱인 Frank and Wolfe(1956)의 알고리즘을 이용하였을 뿐 전역 최적해법을 이용하지는 않았다.

이상에서 언급된 패턴 분류 기법들은 사용자가 분류자(classifier)의 모형을 사전에 정해주고 학습을 통해 그 모형에 적합한 분류자를 구하는 방식을 취하고 있다. 반대로 사용자가 아닌 학습 기계가 분류자의 복잡성(complexity)을 결정하는 학습 방식이 있는데, Support Vector Machine(SVM)(Bradley and Mangasarian, 2000; Mangasarian and Musicant, 2000; Mangasarian, 2000)과, 불 논리(Boolean logic)에 기반한 Logical Analysis of Data(LAD)(Boros *et al.*, 2000; Ryoo and Jang, 2005)가 이의 대표적인 예이다.

SVM은 핵(kernel)을 이용하여 문제를 고차원 공간으로 변형하고 그것을 선형계획법의 해법 또는 전통적인 볼록 최적화 기법을 이용해 풀어내어 원래 문제가 정의된 공간에서 임의의 비선형 형태를 가지는 분류자를 구축할 수 있다는 장점을 가진다. 그리고 LAD는 데이터의 이분화(data binarization), 특

징 벡터 추출(feature selection), 그리고 패턴 인식(pattern construction)의 세 단계로 이루어지는 기계학습 방식으로 그 사용이 편하다는 장점이 있다. 하지만 순차적으로 패턴을 생성해 나가는 LAD는 MSM과 마찬가지로 패턴들 사이의 상호 작용을 고려하지 못하기 때문에 탐욕적 기법에 속한다. 또한 SVM과 LAD는 100% 적합(fitting)을 이루는 임의의 분류자를 구축하는 것을 전제로 하기 때문에 과대적합(overfitting) 문제가 발생하기 쉽다는 단점이 있다.

본 논문에서는 탐욕적 기법의 한계 및 과대적합 문제점을 해결하기 위해 k 개의 비선형 곡면을 동시에 이용하여 의사결정 곡면을 구축하는 패턴 분류 기법을 개발하였다. 이를 위해 Mangasarian(1965)의 연구를 바탕으로 패턴 분류에 사용되는 곡면을 단순한 선형 곡면에서 비선형 곡면으로 확장시켰으며, k 개의 비선형 곡면이 동시에 사용되는 경우를 위해 l_1 -norm 에러 측정 지표를 새롭게 설계하였고, 마지막으로 이를 이용하여 새로운 MILP 모델을 개발하였다. 이 MILP 모델을 이용한 기계학습은 여러 개의 비선형 곡면을 동시에 이용하되 각 곡면의 상호 작용을 고려할 수 있기 때문에 데이터에 대한 분리 능력(separability)이 높으면서도 간결하고 정확한 의사결정 곡면을 구축할 수 있다. 더욱이 주어진 이 모델은 사용자에게 사전에 지정된 비선형 곡면을 이용하여 의사결정 곡면을 구축하기 때문에 학습 기계가 분류자를 임의로 구축하는 방법에 비해 과대적합 문제점이 발생할 소지가 적다. 이상의 특성으로 인하여 본 연구는 선형 의사결정 곡면을 이용할 때의 낮은 분리능력의 문제점, MSM 이외의 탐욕적 기법이 가지는 문제점, LAD 및 SVM이 가지는 과대적합 문제점을 동시에 해결하였다는 점에서 큰 의미를 가진다. 또한 MILP 관련 최적화 이론, 기법, 알고리즘, 효율적인 메타 휴리스틱 및 상용 소프트웨어들을 MILP 모델의 해법에 적용할 수 있기 때문에 본 논문에서 개발된 MILP 기반의 패턴 분류 기법은 기존 MP 모델을 이용한 기계 학습과 달리 실질적인 학습 방법론이다.

본 논문의 구성은 다음과 같다. 다음 장에서는 k -조각별 비선형 분류를 위한 MILP 모델(이하 k PNC 모델)을 개발하고, 3장과 4장에서는 k PNC 모델을 이용한 패턴 분류 기법과 다른 기계학습 기법들(MSM, LAD, SVM)을 비교하였다. 이를 위해 MSM, LAD, SVM, k PNC를 2차원 상에서 인위적으로 만든 문제들과(3장), 기계학습 데이터베이스(Murphy and Aha, 1994)에서 선택한 다섯 개의 표준(benchmark) 데이터베이스(4장)에 적용하여 학습 및 테스트 정확도를 도출하였다. 그리고 마지막 5장에서는 이러한 결과들을 바탕으로 k PNC 모델을 이용한 기계학습의 우수성을 정리하고 향후 연구 과제에 대해 제시하였다.

2. Model development

패턴 분류는 n 차원 공간위에 정의된 서로 다른 두 집합 A, B 에 대해서 $f(\omega, A_i) > \gamma (\forall A_i \in A)$ 와 $f(\omega, B_j) < \gamma (\forall B_j \in B)$ 를 만족

시키는 임의의 의사결정 곡면 $f(\omega, \mathbf{x}) = \gamma$, ($\mathbf{x} \in \mathbb{R}^n, \gamma \in \mathbb{R}^1$)을 찾는 문제이다. 이때, 사용되는 함수 f 가 \mathbf{x} 에 관한 선형 함수이면 선형 패턴 분류, 비선형 함수이면 비선형 패턴 분류라 한다. 하지만, 패턴 분류에서 구해야 하는 의사결정 곡면 $f(\omega, \mathbf{x}) = \gamma$ 은 \mathbf{x} 에 영향을 받지 않고 ω 에 의해 결정되기 때문에 비선형 패턴 분류 문제가 선형 패턴 분류 문제와 항상 다른 종류의 문제가 되는 것은 아니다. 예를 들어 다음의 두 함수 $f^1(\omega, \mathbf{x}) = \omega_1 x_1 + \omega_2 x_2$ 와 $f^2(\omega, \mathbf{x}) = \omega_1 x_1^2 + \omega_2 x_1 x_2$ 을 생각해보자. f^1, f^2 는 각각 선형 함수와 비선형 함수이지만 \mathbf{x} 값이 주어지면 두 함수 모두 ω 에 대한 선형 함수인 $f(\omega, \mathbf{c}) = \omega_1 c_1 + \omega_2 c_2$ 가 된다. 따라서 어떤 함수 $f(\omega, \mathbf{x})$ 가 함수의 파라미터 ω 에 대하여 선형이면 이러한 함수를 이용한 패턴 분류 문제는 \mathbf{x} 에 대한 함수의 종류와 상관없이 선형계획법 문제가 된다(Mangasarian, 1965).

본 논문에서는 ω 가 선형을 이루는 곡면들을 편의상 $f(\omega_L, \mathbf{x})$ 로 나타내고 이러한 곡면들을 동시에 k 개 이용한 비선형 패턴 분류 모델을 제시할 것이다. 모델 개발에 앞서 k 개의 곡면을 동시에 이용하여 얻을 수 있는 의사결정 곡면들의 예를 <Figure 1>에서 제시하였다.

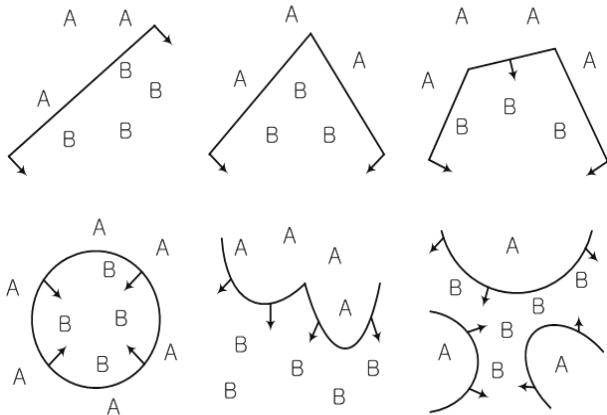


Figure 1. k-piecewise nonlinear classifiers.

<Figure 1>에서는 $k=1, 2, 3$ 인 경우에 선형 곡면만을 이용한 의사결정 곡면들(<Figure 1>의 위)과 이차 곡면만을 이용한 의사결정 곡면들(<Figure 1>의 아래)이 제시되어 있다. 이 그림을 통해서 선형 곡면만을 이용하는 경우에 구축되는 의사결정 곡면들이 이차 곡면을 사용하여 구축되는 의사결정 곡면들에 비해 상당히 단순하고 제한적임을 알 수 있으며 더불어 비선형 패턴 분류 모델의 우수성 및 가치를 이해할 수 있다.

본 논문에서는 <Figure 1>에서 볼 수 있는 것처럼, k 개의 독립된 의사결정 곡면에 대해서 $f_k(\omega_L, \mathbf{x}) < \gamma_k$ (화살표 방향)을 만족하는 부분들의 교집합을 집합 B 의 영역으로, $f_k(\omega_L, \mathbf{x}) > \gamma_k$ 을 만족하는 부분들의 합집합을 집합 A 의 영역으로 정의하였다. 그리고 이렇게 A 와 B 의 영역을 구분할 수 있도록 k 개의 독립된 의사결정 곡면으로부터 조각별 비선형 곡

면을 구축하는 수리 모델을 개발하였다.

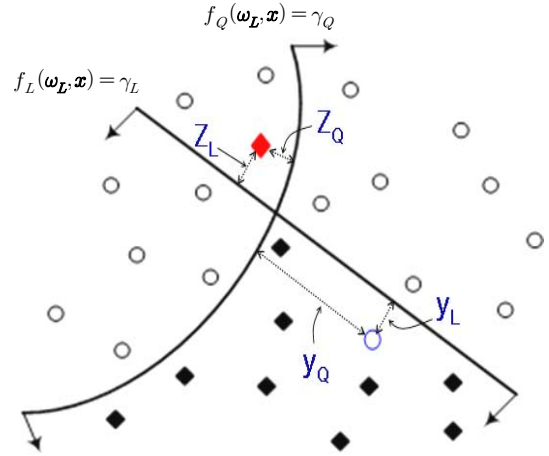


Figure 2. misclassified points with respect to one linear and one quadratic discriminant functions.

모델 개발에 대한 이해를 돕기 위해 <Figure 2>를 제시하였다. <Figure 2>는 $f(\omega_L, \mathbf{x})$ 중에서 가장 단순한 형태인 선형 곡면 $f_L(\omega_L, \mathbf{x})$ 과 이차 곡면 $f_Q(\omega_Q, \mathbf{x})$ 을 사용하는 경우에 대한 예인데, 이 그림에서 동그라미와 마름모는 각각 집합 A 와 집합 B 의 원소를 나타내며, 곡면의 화살표는 $f(\omega_L, \mathbf{x}) < \gamma$ 부분을 표시한다. 집합 A, B 의 영역에 대한 정의에 따라 <Figure 2>에서 집합 B 의 원소들은 $f_L(\omega_L, B_j) < \gamma_L$ 과 $f_Q(\omega_Q, B_j) < \gamma_Q$ 을 동시에 만족해야 하며, 집합 A 의 원소들은 $f_L(\omega_L, A_i) > \gamma_L$ 혹은 $f_Q(\omega_Q, A_i) > \gamma_Q$ 을 만족시켜야 한다. 하지만 <Figure 2>에서는 이러한 조건을 만족시키지 못하는 원소가 두 개 존재하는데, 이러한 에러를 줄이기 위해 A 의 원소에 대하여는 y_L 과 y_Q 중 적어도 하나가 0이 되고 B 의 원소에 대하여는 z_L 과 z_Q 가 모두 0이 되도록 하는 것이 필요함을 관찰을 통해 알 수 있다.

이러한 관찰을 바탕으로 하여 k 개의 비선형 곡면 $f(\omega_L, \mathbf{x})$ 을 동시에 이용하는 다음의 (MP) 모델로 일반화 할 수 있다.

$$\begin{aligned} \min_{\omega, \gamma, y, z} \quad & \sum_{i \in A} \prod_{k \in K} y_{ik} + \sum_{j \in B} \sum_{k \in K} z_{jk} \\ \text{s.t.} \quad & f_k(\omega_L, A_i) - \gamma_k + y_{ik} \geq \epsilon, \quad i \in A, k \in K \quad (1) \\ & f_k(\omega_L, B_j) - \gamma_k - z_{jk} \leq -\epsilon, \quad j \in B, k \in K \quad (2) \\ & y_{ik} \geq 0, \quad i \in A, k \in K \quad (3) \\ & z_{jk} \geq 0, \quad j \in B, k \in K \quad (4) \end{aligned}$$

[기호 정의]

- A_i : A 의 i 번째 원소
- B_j : B 의 j 번째 원소
- $f_k(\omega_L, \mathbf{x}) = \gamma_k$: K 의 k 번째 의사결정 곡면

[변수 및 파라미터 정의]

- y_{ik} : $|\min\{f_k(\omega_L, A_i) - \gamma_k, 0\}|$

- z_{jk} : $|\max\{f_k(\omega_L, B_j) - \gamma_k, 0\}|$
- ϵ : 엄격한 분리를 위한 작은 실수 값

위의 식 (1)은 집합 A 의 원소가 $f_k(\omega_L, \mathbf{x}) > \gamma_k$ 을 만족해야 하고, 그렇지 않을 경우 에러 비용(y_{ik})이 발생함을 나타낸다. 이와 반대로 식 (2)는 집합 B 의 원소가 $f_k(\omega_L, \mathbf{x}) < \gamma_k$ 을 만족해야 하고, 그렇지 않을 경우 에러 비용(z_{jk})이 발생함을 나타낸다. 이렇게 측정된 에러 비용을 가지고 A 와 관련된 y_{ik} 의 곱과 B_j 와 관련된 z_{jk} 의 합을 최소화 하도록 목적함수가 설계되었다. 이 목적함수를 통해서 임의의 A_i 는 y_{ik} 의 값 중 하나가 0이 되면 다른 y_{ik} 의 값에 상관없이 A_i 와 관련된 전체 에러 비용이 0이 됨을 알 수 있다.

하지만 (Ryoo and Sahinidis, 2003)에서 나타난 바와 같이 위의 MP 분류 모델은 그 해를 구하기가 어려우며, 더욱이 관련된 알고리즘과 소프트웨어들이 부족하다는 단점이 있다 따라서 이러한 MP 모델의 단점을 피하고 새로운 모델을 개발하기 위해 본 논문에서는 새로운 에러 측정 지표를 설계하였다.

<Figure 2>를 다시 살펴보면 집합 A 의 원소는 에러 지표 y_L 과 y_Q 중 하나만 0이 되면 정확하게 분류되는 것임을 알 수 있다. 이것을 k 개의 곡면을 사용하는 경우로 일반화하면 집합 A 의 임의의 원소 A_i 를 정확하게 분류하기 위해서는 그 원소와 관련된 에러 지표들($y_{i1}, y_{i2}, y_{i3}, \dots, y_{ik}$) 중에서 가장 작은 값을 0이 되도록 최소화 시키면 된다는 것이다.

따라서 새로운 에러 측정 지표 $r_i := \min_{k \in K} \{y_{ik}\}$ 을 이용한 (kPNC-01) 모델은 다음과 같다.

$$\begin{aligned} \min_{\omega, \gamma, \mathbf{y}, \mathbf{z}, \mathbf{r}} \quad & \sum_{i \in A} r_i + \sum_{j \in B} \sum_{k \in K} z_{jk} \\ \text{s.t.} \quad & (1), (2), (3), (4) \\ & r_i \geq 0, \quad i \in A \end{aligned} \quad (5)$$

위의 (kPNC-01) 모델은 0-1 정수변수의 도입과 Ryoo and Sahinidis(2001)의 볼록화(convexification) 기법으로 선형 함수만을 포함하는 다음의 (kPNC) 모델로 변형될 수 있다.

$$\begin{aligned} \min_{\omega, \gamma, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{x}} \quad & \sum_{i \in A} \sum_{k \in K} u_{ik} + \sum_{j \in B} \sum_{k \in K} z_{jk} \\ \text{s.t.} \quad & (1), (2), (3), (4) \\ & u_{ik} - y_{ik} - Mx_{ik} + M \geq 0, \quad i \in A, k \in K \\ & \sum_{k \in K} x_{ik} \geq 1, \quad i \in A \\ & u_{ik} \geq 0, \quad i \in A, k \in K \\ & x_{ik} \in \{0, 1\}, \quad i \in A, k \in K \end{aligned}$$

Theorem 1. (kPNC) 모델은 (kPNC-01) 모델과 동일하다.

proof. 『 먼저 $(\omega^*, \gamma^*, \mathbf{y}^*, \mathbf{z}^*, \mathbf{u}^*, \mathbf{x}^*)$ 을 (kPNC) 모델의 최적해라 가정하자. 이 최적해의 $(\omega^*, \gamma^*, \mathbf{y}^*, \mathbf{z}^*, \mathbf{u}^*, \mathbf{x}^*)$ 과 동일하거나, 새로운 최적해를 구할 수 있는데 그 방법은 다

음과 같다. 주어진 $(\omega^*, \gamma^*, \mathbf{y}^*, \mathbf{z}^*)$ 에서 각 A_i 의 $(y_{i1}, y_{i2}, y_{i3}, \dots, y_{ik})^*$ 을 조사하여 $y_{il} \leq y_{ik}$ ($k \in K \setminus \{l\}$)을 만족하는 y_{il} 을 찾고, $x_{il} = 1$, $x_{ik} = 0$ ($k \in K \setminus \{l\}$)이라 놓자. 이렇게 생성된 $(\omega^*, \gamma^*, \mathbf{y}^*, \mathbf{z}^*, \mathbf{x}^{**})$ 을 (kPNC) 모델에 적용하면 최소화 문제의 특성으로 인해 $u_{il} = y_{il}$, $u_{ik} = 0$ ($k \in K \setminus \{l\}$)을 만족하는 $(\omega^*, \gamma^*, \mathbf{y}^*, \mathbf{z}^*, \mathbf{u}^{**}, \mathbf{x}^{**})$ 을 얻을 수 있으며 이것은 (kPNC) 모델의 최적해가 된다. 동시에 $(\omega^*, \gamma^*, \mathbf{y}^*, \mathbf{z}^*, \mathbf{u}^{**}, \mathbf{x}^{**})$ 은 식 (1), 식 (2), 식 (3), 식 (4)를 만족하며, 각 A_i 대해서 $\sum_{k \in K} u_{ik} = u_{il} = y_{il} = r_i$ 을 만족시키므로 (kPNC-01) 모델의 최적해가 된다.

다음으로 $(\omega^*, \gamma^*, \mathbf{y}^*, \mathbf{z}^*, \mathbf{r}^*)$ 을 (kPNC-01) 모델의 최적해라 가정하자. 이 최적해의 $(\omega^*, \gamma^*, \mathbf{y}^*, \mathbf{z}^*)$ 을 이용하여 앞의 방식과 동일한 방법으로 $(\mathbf{u}^{***}, \mathbf{x}^{***})$ 을 구할 수 있는데, 이 $(\mathbf{u}^{***}, \mathbf{x}^{***})$ 은 $(\omega^*, \gamma^*, \mathbf{y}^*, \mathbf{z}^*)$ 과 함께 사용되어 (kPNC) 모델의 최적해가 된다.』

3. Illustrative pattern classification

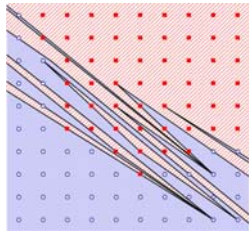
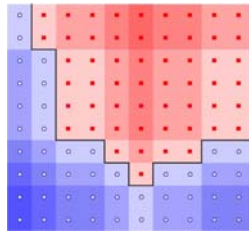
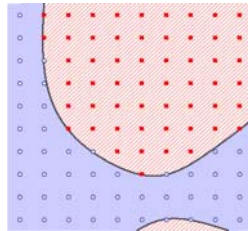
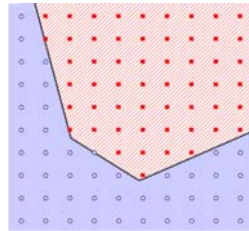
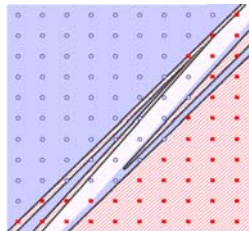
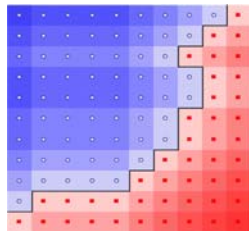
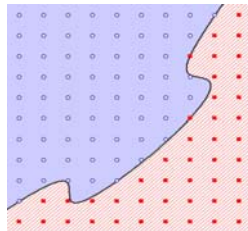
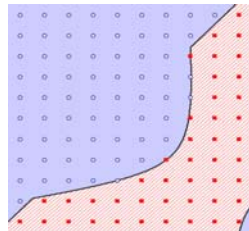
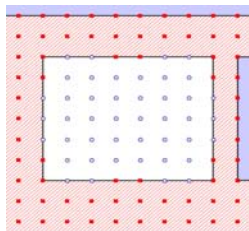
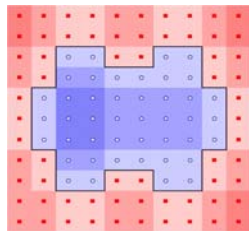
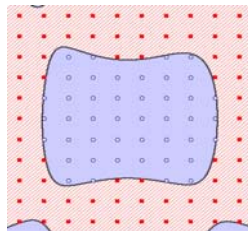
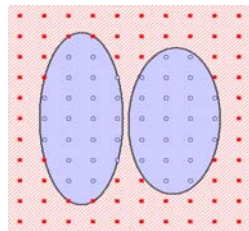
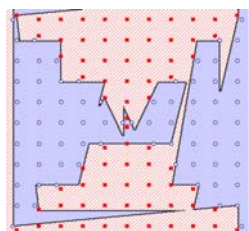
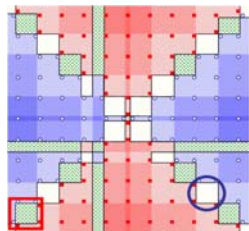
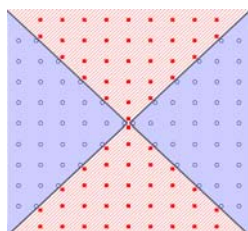
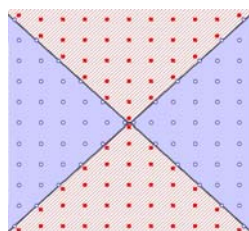
3장에서 kPNC 모델을 이용한 기계 학습의 장점을 보이기 위해 패턴 분류의 다른 기계 학습 기법인 MSM, LAD, SVM과 비교하였다. MSM은 (Nakayama and Kagaku, 1998)에서 제시된 LP 모델 및 기법을 이용하였고, LAD는 (Ryoo and Jang, 2005)의 MILP 모델을 이용한 패턴 분류 기법을 적용하였으며, SVM은 Mangasarian(2000)이 제시한 LP 모델을 사용하였다.

<Table 1>은 MSM, LAD, SVM, kPNC를 인위적으로 만든 2차원 문제들에 적용한 결과이다. 이 결과를 통해 각 기법들이 구축하는 의사결정 곡면을 직접 비교할 수 있다.

<Table 1>에서 kPNC는 MSM, LAD와 달리 곡면 2~3 개만으로 구성되는 간결한 분류자를 구축하여 주어진 문제를 완벽히 분리하고 있다. 이에 반해 MSM은 Mangasarian(1968)이 제시한 퇴화 절차(degenerate procedure)를 적용하지 않을 경우, 주어진 문제의 두 집합을 완벽히 분리하지 못하며(Ex2: 87.3% 분리, Ex3: 49.1% 분리), 순차적으로 평행한 선형 곡면을 얻어서 분류자를 구축하는 탐욕적 기법을 쓰기 때문에 과도하게 복잡한 의사결정 곡면을 형성한다(Ex1, Ex4). 특히 Ex4는 kPNC를 이용한 기계 학습의 장점을 잘 보여주는 예로서, MSM에 의해서는 25개의 선형 곡면으로 분리되지만, kPNC에 의해서는 이차 곡면 2개만으로 분리가 된다.

LAD는 각각의 항목(attribute)에 대해서 절단점(cutpoint)을 생성하고 이러한 절단점을 조합하여 초직사각형(hyper-rectangular) 형태의 패턴들을 생성하게 되는데, 이로 인해서 Ex1, Ex2, Ex4와 같은 문제에서 kPNC가 대각선을 통해 쉽게 분류할 수 있는 것도 패턴 여러 개를 이용해야 분류할 수 있다는 단점을 가지게 된다. 또한 LAD는 학습(training)을 통해 주어진 문제를 완전히 분리한다 할지라도 kPNC나 SVM처럼 전체 공간을 이분하지 않기 때문에 생기는 문제를 지니게 된다(예를 들어 새로운 데이터의 집합을 판단하지 못하는 경우(indecisive decision region)와, 집합 A 와 집합 B 의 패턴들이 서로 상반되

Table 1. Comparison of 4 machine learning approaches with 4 examples

	MSM	LAD	SVM	PNC
Ex1	 <p>19 linear surfaces 100% separation</p>	 <p>4 positive patterns 7 negative patterns</p>	 <p>polynomial kernel, $\nu=10^4, \lambda=1, \rho=0,$ $\mu=-1, \text{degree}=3$</p>	 <p>3 linear surfaces</p>
Ex2	 <p>8 linear surfaces with indecisive region (87.3% separation)</p>	 <p>7 positive patterns 8 negative patterns</p>	 <p>sinusoidal kernel, $\nu=10^4, \lambda=1, \rho=1,$ $\mu=-1, \text{degree}=3$</p>	 <p>1 linear surface and 1 quadratic surface</p>
Ex3	 <p>8 linear surfaces with indecisive region (49.1% separation)</p>	 <p>4 positive patterns 10 negative patterns</p>	 <p>sinusoidal kernel $\nu=10^5, \lambda=50/3 \pi, \rho=2 \pi,$ $\mu=1, \text{degree}=6$</p>	 <p>2 quadratic surfaces</p>
Ex4	 <p>25 linear surfaces 100% separation</p>	 <p>8 positive patterns 8 negative patterns ○ indecisive region □ conflicting region</p>	 <p>polynomial kernel, $\nu=10^4, \lambda=1, \rho=0,$ $\mu=-1, \text{degree}=2$</p>	 <p>2 quadratic surfaces</p>

는 정보를 제공하는 경우(conflicting decision region)도 발생한다(Ex4).

위의 문제에서 SVM이 구축한 의사결정 곡면은 kPNC의 그것과 가장 유사하다. 하지만 주어진 데이터에 잡음(noise)이 포함되어 있을 경우, SVM은 데이터의 100% 분리를 기본으로 함으로 인하여 과대적합 문제를 야기하게 된다. 이 점은 다음 장에서 잡음 데이터 셋에 대한 SVM의 분류결과를 분석하여 설명하기로 한다.

4. Numerical experiments & discussions

본 논문에서 개발된 kPNC 모델을 이용한 기계학습의 학습 능력을 테스트하고 MSM, LAD, SVM과의 비교를 위해서 기계학습 연구에 통용되는 다섯 개의 현실 데이터베이스를 Murphy and Aha(1994)로부터 다운로드 받았다. 그리고 다섯 개의 현실 데이터베이스를 실험에 적용하기 위해 각 데이터 집합으로부터 잃어버린 항목이 있는 데이터는 삭제하였다.

<Table 2>는 다섯 개의 데이터베이스에 대한 정보를 정리한 것이다.

Table 2. Five databases studied

Database	number of attributes	number of observations	
		total	class A / class B
Australian credit card	15	653	296 approval / 357 denied
Boston housing	13	506	260 income \geq \$21K / 246 else
Cleveland heart disease	13	297	137 disease / 160 no disease
Pima Indian diabetes	8	768	268 diabetes / 500 no disease
Wisconsin breast cancer	9	683	239 malignant / 444 benign

우리는 학습 및 테스트를 위하여, 각 데이터베이스에 속한 전체 원소들 중에서 $P\%$ ($P \in \{50, 80\}$)를 무작위로 선택하여 학습 문제를 만들고 나머지 데이터를 가지고 테스트 문제를 구성하였다. 이렇게 만들어진 학습 및 테스트 문제에 MSM, LAD, SVM, k PNC 기법을 적용하여 정확도를 구하였으며, 이러한 과정을 30번 반복하여 정확도에 대한 평균 및 표준편차를 구하였다.

우리는 이 과정에서 LP 및 MILP의 해법을 위해 ILOG CPLEX 9.0(2003)을 이용하였다. k PNC의 경우, MILP의 해법에서 분지한계기법의 노드 수를 최대 10,000개로 제한하고, 가장 좋은 해를 이용하여 PNC 분류자를 구축하였다.

<Table 3>은 5개의 표준 데이터베이스로부터 만든 문제 집합에 k PNC 모델을 적용한 결과이다. 이 표에서 k 는 k PNC 모델에 사용된 꼭면의 개수, type은 사용된 꼭면의 종류를 나타낸다. 이 <Table 3>의 결과에서 볼 수 있는 것은 k PNC 모델이 제

Table 3. PNC classifier results with five benchmark databases (L : linear surface, Q : quadratic surface)

Database	Class A	k	type	PNC classifiers' accuracy with					
				50% training		80% training		100% training	
				training	testing	training	testing	training	
Australian Credit Card	either	1	L	86.7 \pm 1.2	86.0 \pm 1.7	86.7 \pm 0.6	85.8 \pm 2.5	86.5	
	approved	2	LL	89.6 \pm 1.7	85.2 \pm 1.5	87.8 \pm 1.0	85.4 \pm 2.6	87.3	
		3	LLL	89.4 \pm 1.6	85.3 \pm 1.4	88.2 \pm 1.1	85.7 \pm 2.6	88.2	
		1	Q	100.0 \pm 0.0	71.8 \pm 3.3	98.0 \pm 1.2	77.9 \pm 3.2	94.8	
		2	LQ	97.4 \pm 4.8	72.4 \pm 5.0	98.5 \pm 3.3	77.5 \pm 4.0	95.1	
		2	QQ	100.0 \pm 0.0	67.1 \pm 4.5	100.0 \pm 0.0	71.4 \pm 4.3	100.0	
	denied	2	LL	90.0 \pm 1.9	84.4 \pm 1.4	87.6 \pm 1.1	85.4 \pm 2.7	87.7	
		3	LLL	89.5 \pm 1.9	84.5 \pm 2.0	88.0 \pm 1.5	85.2 \pm 2.5	86.8	
		1	Q	100.0 \pm 0.0	71.8 \pm 3.2	98.0 \pm 1.2	77.8 \pm 3.2	94.8	
		2	LQ	99.2 \pm 3.1	72.0 \pm 3.3	95.8 \pm 6.9	78.1 \pm 3.8	97.5	
2		QQ	100.0 \pm 0.0	70.0 \pm 6.0	100.0 \pm 0.0	71.2 \pm 4.4	100.0		
Cleveland Heart Disease	either	1	L	87.2 \pm 2.2	81.8 \pm 2.2	85.3 \pm 0.9	83.4 \pm 3.3	84.2	
	disease	2	LL	97.4 \pm 2.8	76.4 \pm 2.8	88.1 \pm 1.0	81.7 \pm 3.9	87.9	
		3	LLL	98.7 \pm 1.9	75.9 \pm 2.8	89.9 \pm 1.7	80.1 \pm 4.1	86.5	
		1	Q	100.0 \pm 0.0	66.5 \pm 5.0	100.0 \pm 0.0	68.4 \pm 5.6	100.0	
	no disease	2	LL	94.6 \pm 2.9	76.9 \pm 3.2	88.7 \pm 1.3	80.9 \pm 4.4	88.2	
		3	LLL	98.4 \pm 1.9	75.4 \pm 3.5	90.1 \pm 1.1	80.1 \pm 3.8	88.2	
		1	Q	100.0 \pm 0.0	66.8 \pm 4.4	100.0 \pm 0.0	69.1 \pm 5.1	100.0	
	Boston Housing	either	1	L	88.9 \pm 1.8	85.9 \pm 1.5	88.2 \pm 0.9	85.3 \pm 3.3	87.2
		income \geq \$21K	2	LL	93.2 \pm 1.8	84.9 \pm 1.9	89.9 \pm 1.1	86.2 \pm 2.8	89.3
			3	LLL	95.8 \pm 2.0	84.5 \pm 2.0	91.2 \pm 1.2	86.0 \pm 3.9	89.1
1			Q	100.0 \pm 0.0	74.6 \pm 3.5	100.0 \pm 0.0	80.1 \pm 3.5	100.0	
< \$21K		2	LL	95.3 \pm 1.8	86.1 \pm 2.0	91.9 \pm 1.3	86.7 \pm 3.4	91.7	
		3	LLL	97.6 \pm 1.5	84.9 \pm 2.7	92.6 \pm 1.4	86.6 \pm 2.6	93.1	
		1	Q	100.0 \pm 0.0	74.4 \pm 3.8	100.0 \pm 0.0	80.3 \pm 4.8	100.0	
Wisconsin Breast Cancer		either	1	L	97.7 \pm 0.6	96.3 \pm 0.8	97.5 \pm 0.4	95.8 \pm 1.4	97.4
		disease	2	LL	99.2 \pm 0.5	95.3 \pm 1.4	98.7 \pm 0.4	96.0 \pm 1.5	98.2
			3	LLL	99.8 \pm 0.3	93.9 \pm 1.4	99.3 \pm 0.3	95.3 \pm 2.0	98.8
	1		Q	100.0 \pm 0.0	91.7 \pm 1.8	100.0 \pm 0.0	91.5 \pm 2.1	100.0	
	no disease	2	LL	99.9 \pm 0.3	94.7 \pm 1.2	98.5 \pm 0.6	95.1 \pm 1.8	97.8	
		3	LLL	100.0 \pm 0.0	95.1 \pm 1.3	98.9 \pm 0.7	94.9 \pm 1.7	97.5	
		1	Q	100.0 \pm 0.0	91.6 \pm 1.8	100.0 \pm 0.0	91.1 \pm 2.6	100.0	
	Pima Indian Diabetes	either	1	L	78.1 \pm 1.2	76.5 \pm 1.6	77.9 \pm 0.9	77.2 \pm 3.3	77.3
		disease	2	LL	79.2 \pm 1.4	76.3 \pm 1.6	78.3 \pm 1.0	76.8 \pm 3.3	77.7
			3	LLL	78.2 \pm 2.2	74.4 \pm 2.4	76.9 \pm 1.3	75.3 \pm 3.2	77.3
1			Q	82.8 \pm 1.5	74.1 \pm 1.8	80.6 \pm 0.7	75.7 \pm 2.3	80.5	
2			LQ	83.6 \pm 1.8	73.5 \pm 1.7	80.0 \pm 1.2	75.1 \pm 2.4	79.2	
2			QQ	88.7 \pm 2.0	72.0 \pm 2.5	84.2 \pm 1.1	74.6 \pm 2.8	85.5	
no disease		2	LL	79.0 \pm 1.3	76.0 \pm 1.7	78.3 \pm 0.8	76.8 \pm 3.1	78.3	
		3	LLL	78.9 \pm 2.0	74.9 \pm 2.6	77.7 \pm 1.2	76.7 \pm 3.2	77.5	
		1	Q	82.8 \pm 1.5	74.1 \pm 1.8	80.6 \pm 0.7	75.7 \pm 2.3	80.5	
		2	LQ	85.9 \pm 1.6	72.6 \pm 2.3	81.9 \pm 1.4	74.8 \pm 2.5	80.1	
	2	QQ	90.2 \pm 2.4	70.8 \pm 2.2	84.3 \pm 1.1	74.0 \pm 2.6	82.4		

공하는 분류자의 우수한 분리 능력이다. 비선형 곡면을 이용할 수 있는 *k*PNC 모델은 5개의 데이터베이스 중에서 3개를 (Cleveland heart disease, Boston housing, Wisconsin breast cancer) 단지 한 개의 이차 곡면으로 100% 분리하였으며, Australian credit card는 이차 곡면 2개를 이용하여 100% 분리하였다. 반면에 선형 곡면만을 이용하는 경우는 주어진 5개의 데이터베이스 중에서 단 1개도 100% 분리하지 못하였다. 이는 앞의 <Figure 1>에서 제시하였던 것처럼 비선형 곡면을 이용한 의사결정 곡면이 선형 곡면만을 이용한 의사결정 곡면에 비해 훨씬 우수한 분리 능력을 가지고 있음을 뒷받침하는 증거이며, 동시에 다양한 비선형 곡면을 적용할 수 있는 본 모델의 활용 가능성을 충분히 예측할 수 있도록 한다.

<Table 4, 5>에서는 MSM, LAD, SVM, *k*PNC의 가장 좋은 테스트 결과와 그 때의 학습 결과를 요약하였다. <Table 3>과 <Table 4, 5>의 결과에서 주의할 것은 테스트를 위해 사용된 다섯 개의 현실 데이터가 과대적합 문제를 야기하는 잡음을 포함하고 있다는 것이다. 이것은 <Table 3>에서 제시된 *k*PNC의 결과와 <Table 4, 5>에서 제시된 MSM, LAD의 결과에 의해 뒷받침된다. <Table 3>에서는 *k*PNC의 학습 정확도가 높아짐에 따라 테스트 정확도가 일반적으로 낮아지며, <Table 4, 5>에서는 MSM과 LAD가 주어진 학습 집합을 완벽하게 분리함에도 불구하고 두 기법의 일반화 능력이 학습 에러를 포함하고 있는 *k*PNC보다 낮게 나타난다. 이와 같이 현실 데이터에는 일반적으로 잡음이 포함되는데, 이것은 데이터의 100% 적합을 전제로 하고 있는 MSM, LAD, SVM 기법을 이용한 패턴 분류에서 특별히 과대적합 문제를 조심해야 하는 이유가 된다. 이에 반하여 사용자에게 의해 사용될 함수의 종류와 개수가 정해지는 *k*PNC는 상대적으로 과대적합의 문제점을 피하기 쉬우며 이

것은 <Table 4, 5>에서 나타난 테스트 결과를 통해 알 수 있다.

<Table 4, 5>에서 제시된 SVM의 결과는, 세 가지 핵(neural, polynomial, sinusoidal)에 대하여 파라미터 값들의 다양한 조합(neural : 6가지, polynomial : 18가지, sinusoidal : 15가지)을 적용하였을 때 얻은 가장 좋은 테스트 결과이다. 이 결과에서 우리는 SVM이 다양한 조합의 파라미터를 이용하여 결과를 얻었음에도 불구하고, LAD에 비해 우수한 테스트 결과를 보여주지 못하고 있음을 주목할 수 있다. 이 결과는 위의 데이터에 잡음이 포함되어 있다는 점을 감안할 때, SVM이 잡음에 민감한 기법이라는 것과 LAD가 위의 세 학습기법 중 잡음에 적게 영향을 받는 기법이라는 것으로 해석할 수 있다. 마지막으로 MSM은 탐욕적 기법의 한계로 인하여 과도하게 복잡한 의사결정 곡면을 구축하는 문제점과 데이터의 100% 적합으로 인한 과대적합 문제점을 가지고 있는데, 이로 인해 LAD 및 *k*PNC에 비해 일반화능력이 떨어지는 것 같다.

5. Conclusions

본 논문에서는 *k*-조각별 비선형 분류를 위하여 MILP 모델인 *k*PNC 모델을 개발하였다. *k*PNC 모델은 비선형 함수를 동시에 *k*개 사용하되, 각 함수의 상호 작용을 고려할 수 있기 때문에 분리 능력이 높으면서도 가장 정확한 의사결정 곡면을 제공한다. 이러한 *k*PNC 모델의 우수성을 살펴보기 위해 MSM, LAD, SVM과 비교하였는데, *k*PNC 모델을 이용한 새로운 기법은 다른 기법들보다 단순하고 정확한 분류자를 구축하였으며 가장 우수한 일반화 능력을 나타내었다.

앞으로 수행되어야 할 연구는 *k*PNC 모델의 해법 개발에 관

Table 4. Comparison of 4 machine learning approaches with 50% training

Database	MSM		LAD		SVM		PNC	
	train	test	train	test	train	test	train	test
A. C. card	100.0±0.0	77.8±3.8	100.0±0.0	81.7±2.3	93.1±1.3	81.3±2.5	87.2±2.2	86.0±1.7
C. H. disease	100.0±0.0	76.1±2.7	100.0±0.0	75.5±3.2	89.3±2.4	76.7±3.0	86.7±1.2	81.8±2.2
B. housing	100.0±0.0	83.0±2.9	100.0±0.0	81.1±1.9	91.6±1.3	81.9±2.0	95.3±1.8	86.1±2.0
W. B. cancer	100.0±0.0	94.9±1.1	100.0±0.0	94.6±1.3	97.9±0.6	97.0±0.8	97.7±0.6	96.3±0.8
P. I. diabetes	100.0±0.0	68.2±2.7	100.0±0.0	68.3±2.3	72.6±1.8	64.0±1.3	78.1±1.2	76.5±1.6

Table 5. Comparison of 4 machine learning approaches with 80% training

Database	MSM		LAD		SVM		PNC	
	train	test	train	test	train	test	train	test
A. C. card	99.8±0.9	77.6±4.5	100.0±0.0	82.1±3.2	89.8±0.7	84.5±3.2	86.7±0.6	85.8±2.5
C. H. disease	100.0±0.0	77.0±5.2	100.0±0.0	78.9±5.2	93.4±1.4	78.7±4.6	85.3±0.9	83.4±3.3
B. housing	100.0±0.0	83.1±3.7	100.0±0.0	83.4±3.5	90.0±1.1	84.2±3.3	91.9±1.3	86.7±3.4
W. B. cancer	100.0±0.0	94.2±2.1	100.0±0.0	94.2±1.2	97.9±0.3	96.7±1.3	98.7±0.4	96.0±1.5
P. I. diabetes	100.0±0.0	67.5±3.6	100.0±0.0	68.9±4.1	70.1±1.0	65.5±2.8	77.9±0.9	77.2±3.3

한 것이다. 현재 k PNC 모델의 해법으로 MILP 관련 전역 최적화 기법인 분지한계기법을 사용하고 있는데, LP 완화모델이 좋은 한계를 제공하지 못하는 어려움이 있다 따라서 이에 대한 해결을 위해 절단면 기법, 분할 기법에 관한 연구를 수행할 것이며, 데이터의 감소를 위한 효율적인 군집화 기법 및 지지 벡터 추출 기법 개발에 대한 연구가 진행될 것이다.

References

- Al-Khayyal, F. A. and Falk, J. E. (1983), Jointly constrained biconvex programming, *Mathematics of Operations Research*, **8**(2), 273-286.
- Bennett, K. P. (1993), Decision tree construction via linear programming, In *Proceedings of the 9th Conference on Artificial Intelligence for Applications*, Orlando, Florida, 212-218.
- Bennett, K. P. and Mangasarian, O. L. (1992), Robust linear programming discrimination of two linearly inseparable sets, *Optimization Methods and Software*, **1**, 23-34.
- Bennett, K. P. and Mangasarian, O. L. (1994), Bilinear separation of two sets in n -space, *Computational Optimization and Applications*, **2**, 207-227.
- Boros, E., Hammer, P. L., Ibaraki, T., Kogan, A., Mayoraz, E., and Muchnik, I. (2000), An implementation of logical analysis of data, *IEEE Transactions on Knowledge and Data Engineering*, **12**(2), 292-306.
- Bradley, P. S. and Mangasarian, O. L. (2000), Massive data discrimination via linear support vector machines, *Optimization Methods and Software*, **13**, 1-10.
- Carter, C. and Catlett, J. (1997), Assessing credit card applications using machine learning, *IEEE Expert*, Fall, 71-79.
- Falk, J. E. and Lopez-Cardona, E. (1997), The surgical separation of sets, *Journal of Global Optimization*, **11**, 433-462.
- Frank, M. and Wolfe, P. (1956), An algorithm for quadratic programming, *Naval Research Logistics Quarterly*, **3**, 95-110.
- ILOG CPLEX 9.0 User's Manual*, ILOG CPLEX Division, Incline, Nevada, October 2003.
- Mangasarian, O. L. (1965), Linear and nonlinear separation of patterns by linear programming, *Operations Research*, **13**, 444-452.
- Mangasarian, O. L. (1968), Multisurface method of pattern separation, *IEEE Transactions on Information Theory*, **14**(6), 801-807.
- Mangasarian, O. L. (1993), Mathematical programming in neural networks, *ORSA Journal on Computing*, **5**, 349-360.
- Mangasarian, O. L. (2000), Generalized support vector machines, In A. Smola, P. Bartlett, B. Schölkopf and D. Schuurmans, editors, *Advances in Large margin classifiers*, MIT Press, 135-146.
- Mangasarian, O. L. and Musicant, D. R. (2000), Data discrimination via nonlinear generalized support vector machines, In M. C. Ferris, O. L. Mangasarian, and J.-S. Pang, editors, *Complementarity : Applications, Algorithms and Extensions*, chapter 1. Kluwer Academic Publishers.
- Mangasarian, O. L., Setiono, R., and Wolberg, W. H. (1990), Pattern recognition via linear programming : Theory and application to medical diagnosis, *Large-Scale Numerical Optimization*, 22-31.
- Mangasarian, O. L., Street, W. N., and Wolberg, W. H. (1995), Breast cancer diagnosis and prognosis via linear programming, *Operations Research*, **43**(4), 570-577.
- Megiddo, N. (1988), On the complexity of polyhedral separability, *Discrete and Computational Geometry*, **3**, 325-337.
- Murphy, P. M. and Aha, D. W. (1994), UCI repository of machine learning databases, Department of Computer Science, University of California at Irvine, CA.
- Nakayama, H. and Kagaku, N. (1998), Pattern classification by linear goal programming and its extensions, *Journal of Global Optimization*, **12**, 111-126.
- Osuna, E., Freund, R., and Girosi, F. (1997), Training support vector machines : an application to face detection, In *IEEE Conference on Computer Vision and Pattern Recognition*, Puerto Rico, 130-136.
- Ryoo, H. S. and Jang, I.-Y. (2005), MILP approach to pattern generation in logical analysis of data, *Machine Learning*, submitted.
- Ryoo, H. S. and Sahinidis, N. V. (2001), Analysis of bounds for multilinear functions, *Journal of Global Optimization*, **19**(4), 403-424.
- Ryoo, H. S. and Sahinidis, N. V. (2003), Global optimization of multiplicative programs, *Journal of Global Optimization*, **26**, 387-418.
- Ullman, J. R. (1973), *Pattern recognition techniques*, Crane, London.
- Wolberg, W. H. and Mangasarian, O. L. (1990), Multisurface method of pattern separation for medical diagnosis applied to breast cytology, *Proceedings of the National Academy of Sciences*, **87**, 9193-9196.