Proceedings of KFIS 2001 Fall Conference, 2001. 12. 1

커널 이완절차에 의한 커널 공간의 저밀도 표현 학습

Sparse Representation Learning of Kernel Space Using the Kernel Relaxation Procedure

류재홍*, 정종철
여수대학교 컴퓨터공학과

Jae Hung Yoo*, Jong Cheol Jeong,
Dept, of Computer Engineering, Yosu National University
*(jhy@ce.yosu.ac.kr)

ABSTRACT

In this paper, a new learning methodology for Kernel Methods is suggested that results in a sparse representation of kernel space from the training patterns for classification problems.
Among the traditional algorithms of linear discriminant function(perceptron, relaxation, LMS(least mean squared), pseudoinverse), this paper shows that the relaxation procedure can obtain the maximum margin separating hyperplane of linearly separable pattern classification problem as SVM(Support Vector Machine) classifier does.
The original relaxation method gives only the necessary condition of SV patterns. We suggest the sufficient condition to identify the SV patterns in the learning epochs.
Experiment results show the new methods have the higher or equivalent performance compared to the conventional approach.

Keywords : Sparse Representation, Kernel Space, Kernel Discriminant, Kernel Hyperplane, Relaxation Procedures

Ⅰ. 서 론

본 논문은 패턴 분류 문제에 대한 기법 중 SVM(Support Vector Machine)과 RBF(Radial Basis Function) 신경회로망 등 커널 방법(Kernel Methods)으로 패턴 가중치를 학습하는 과정에서 가중치가 영(zero)인 훈련 패턴이 되도록 많이 생성되도록 하는 것이 목표인 저밀도 표현(Sparse Representation)의 새로운 학습 방법을 소개한다.
기존의 SVM은 2진 분류 문제에 대하여 입력 공간에서 선형 분류 함수의 분류 초평면에 대한 최대 분류 여유(Maximum Margin of Separation)와 지원 벡터(Support Vector)를 정의하고 입력 패턴 가중치 벡터의 크기를 최소화하는 제한적 최적화 문제(Constrained Optimization Problem)를 제안하였다[4]. 이것은 Lagrange 제2공간(Dual Space)에서 QP(Quadratic Problem)가 된다. SVM은 QP의 해로써 훈련 패턴 가중치를 결정하는 일괄 처리 방식(Batch Mode Processing)이다.
하지만 QP는 연산과 메모리 복잡도가 크다는 단점을 갖고 있다. 이러한 문제점을 해결하기 위해서 Chunking, Active Set 등 QP 분해방법(Decomposition Method)과 최대 강하/ 상승(Steepest Decent/Ascent) 방법, Relaxation 등

각 훈련 패턴에 대하여 학습하는 방법이 소개되었다[5,6,7,8,9].

특히 ROMMA[9]는 본 논문과 사용하는 기존의 Relaxation과 가장 유사한 방법이다. 최대 분류 여유를 찾기 위하여 학습하는 것은 두 방법이 동등하다. ROMMA는 가중치 공간(Weight Space)에서 현재의 가중치 벡터에 의한 반공간(Halfspace) 영역을 정의하고 최대 분류 여유에 상응하는 최소 가중치 벡터 학습에 이용하였으나 수렴시 가중치 수정횟수는 기존의 Relaxation 학습 방법[1]과 동일한 상한선(Upper Bound)을 갖는다. 따라서 전통적인 Relaxation 학습 방법은 간단히 구현할 수 있는데 복잡한 ROMMA를 선호가 이유가 전혀 없다. ROMMA의 저자가 Duda의 고전[1]에 소개된 Relaxation 학습 방법을 인용하지 못한 것은 유감이고 바퀴의 재 발명(Reinvention of the Wheel) 성격이 짙다.

기존의 선형 판별 함수(LDF - Linear Discriminant Function) 학습 방법[1] 중 (Perceptron, Relaxation, LMS(Least Mean Squared), Pseudoinverse)에서 SVM의 최대 분류 여유(Maximum Margin of Separation)를 갖는 초평면을 Relaxation, Perceptron with Margin 등 여유 분류기(Margin Classifier)는 찾을 수 있다는 것을 본 논문에서 처음으로 밝힌다. SVM의 공헌도는 제2 공간(Dual Space) 또는 커널 공간에서 제대로 분류되는 훈련 패턴의 가중치를 영으로 둠으로써 성김 표현 또는 저 밀도 표현을 가능하게 한 것이다.

기존의 Relaxation 방법은 SV 패턴을 학습하는데 필요 조건(Necessary Condition)을 만족시킨다. 본 논문에서는 SV 패턴을 찾는 것에 대한 충분 조건(Sufficient Condition)을 제안하여 수정된 Relaxation 학습법이 최대 분류 여유를 갖는 초평면을 찾을 수 있도록 한다.

본 논문은 커널 공간에서 kernel 행렬을 패턴 행렬로 해석하는데 SVM의 QP의 Lagrange 승수 벡터(multiplier vector)는 패턴 가중치가 된다. 따라서 선형 판별 함수의 모든 학습 방법들을 그대로 커널 공간에 적용할 수 있다. 일반적으로 논의되는 특징 공간(Feature Space)[2]은 내적 커널(Inner Product Kernel) 형성에 대한 유효성 해석에 필요한 것이고 학습에는 본 논문에서 소개하는 커널 공간, 커널 초평면, 커널 판별함수 등의 개념이 중요한 것이다.

II. 본 론

먼저 SVM과 선형판별 함수 학습 방법에 대해 검토하고 본 논문에서 제안하는 수정된 Relaxation 학습에 의한 최대 분류 여유와 SV 패턴 학습법을 다음에 논의하고, 커널 공간에서의 저밀도 표현 학습과 실험 결과를 차례대로 소개한다. 정확성을 기하기 위하여 필요한 용어를 정의한다. 주어진 N개의 훈련 패턴은 m차원 행 벡터 $\mathbf{x}_i, (i=1,...,N)$와 패턴 행렬 $\mathbf{X}=[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$을 정의한다. 이진 분류 문제에서 목표치는 $d_i \in \{+1, -1\}$, $i=1,...,N$이고 목표치 벡터는 $\mathbf{d} = [d_1, d_2, ..., d_N]^T$가 된다. 선형 판별 함수 또는 선형 문턱 함수(Linear Threshold Function)는 다음과 같다.

$$g(\mathbf{x}) = \mathbf{w}^t \mathbf{x} + w_0 \quad (1)$$

여기서 $\mathbf{w}$는 m차원 가중치 벡터이고 w_0는 바이어스(bias term) 또는 문턱 값(threshold value)이다. 주어진 입력 패턴을 확장하면(augmented) m+1차원 동형 좌표 공간(Homeogenous Coordinate Space)의 $\mathbf{y} = [\mathbf{x}, 1]^T$가 된다. 확장된 패턴 행렬은 $\mathbf{Y}=[\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N]$가 된다. 가중치 벡터는 $\mathbf{a} = [\mathbf{w}, w_0]^T$가 된다. 따라서 선형 판별식은 다음과 같다.

$$g_a(\mathbf{y}) = \mathbf{a}^t \mathbf{y} \quad (2)$$

선형 내적 커널과 그의 확장형은 각각 $\mathbf{X}^T\mathbf{X}$와 $\mathbf{Y}^T\mathbf{Y}$가 된다.

2.1 SVM 분류기(Classifier)

선형 판별식에서 분리 초평면은 다음 식을 만족한다.

$$d_i\, g(\mathbf{x}_i) \geq b, \quad i=1,\ldots,N \quad (3)$$

패턴으로부터 분리 초평면까지의 거리는 다음과 같다.

$$\frac{d_i\, g(\mathbf{x}_i)}{\|\mathbf{w}\|} \geq r, \quad i=1,\ldots,N \quad (4)$$

식 3과 4에서 $b = r \|\mathbf{w}\|$ 놓으면, 식 1,과 4의 해는 가중치 벡터 $\mathbf{w}$를 최소화한다.

$$L(\mathbf{w}, \boldsymbol{\alpha}) = \frac{\|\boldsymbol{w}\|^2}{2} - \sum_{i=1}^{N} \alpha_i [d_i\, \mathbf{w}^t\, \mathbf{x}_i - b] \quad (5)$$

제2공간에서는 QP가 다음과 같다.

$$L(\boldsymbol{\alpha}) = b\sum_{i=1}^{N}\alpha_i - \sum_{i,j=1}^{N}\alpha_i\alpha_j d_i d_j\, \mathbf{x}_i^t\, \mathbf{x}_j \quad (6)$$

다음의 제한 조건(Constraints)을 갖는다.

$$\sum_{i=1}^{N}\alpha_i d_i = 0, \quad \alpha_i \geq 0, i=1,\ldots,N \quad (7)$$

바이어스 항을 함께 최소화하기 위해 확장공간에서 가중치 벡터 $\boldsymbol{a}$를 최소화하면[1]

$$\frac{d_i\, g(\mathbf{x}_i)}{\|\mathbf{w}\|} \geq \frac{d_i\, g_a(\mathbf{y}_i)}{\|\mathbf{y}\|} \geq r_a, \quad i=1,\ldots,N \quad (8)$$

식 3을 변형하면

$$d_i[d_i * b - g_a(\mathbf{y}_i)] \le 0, \quad i=1,\ldots,N \tag{9}$$

식 4는 다음과 같이 변한다.

$$L(\mathbf{a}, \alpha) = \frac{||\alpha||^2}{2} - \sum_{i=1}^{N} \alpha_i d_i [d_i * b - \mathbf{a}^t \mathbf{y}_i] \tag{10}$$

여기서 $\alpha_i d_i$를 α_i로 놓으면

$$L(\alpha) = b \sum_{i=1}^{N} d_i \alpha_i - \sum_{i,j=1}^{N} \alpha_i \alpha_j \mathbf{y}_i^t \mathbf{y}_j \tag{11}$$

또는

$$L(\alpha) = b\, \mathbf{d}^t \alpha - \alpha^t Y^t Y \alpha \tag{12}$$

다음의 제한 조건(Constraints)을 갖는다.

$$\alpha_i d_i \ge 0, i=1,\ldots,N \tag{13}$$

즉 식 7에서 등식이 제거된 것이다.
식 9이후에서 b=1로 놓고 오차 항을 넣고 부등호를 등호로 치환하면 회귀문제에도 공히 적용할 수 있으나 LMS 학습법이 적용된다..

2.2 선형 판별함수 학습법

Percetron은 잘못 분류된 훈련 패턴에 대하여 가중치를 학습한다.

$$\mathbf{a}(k+1) = \mathbf{a}(k) + d_k \mathbf{y}(k) \quad until\ d_k g_a(\mathbf{y}_k) \ge 0, k=1,\ldots,N \tag{14}$$

Percetron with Margin은 여유 이하로 학습된 훈련 패턴에 대하여 가중치를 학습한다.

$$\mathbf{a}(k+1) = \mathbf{a}(k) + \eta(k) d_k \mathbf{y}(k) \quad until\ d_k g_a(\mathbf{y}_k) \ge b > 0, k=1,\ldots,N \tag{15}$$

Relaxation도 여유 이하로 학습된 훈련 패턴에 대하여 가중치를 학습한다.

$$\mathbf{a}(k+1) = \mathbf{a}(k) + \eta(k) \frac{[d_k * b - \mathbf{a}^t \mathbf{y}(k)]}{||\mathbf{y}(k)||^2} \mathbf{y}(k) \quad until\ d_k g_a(\mathbf{y}_k) \ge \gamma > 0, k=1,\ldots,N \tag{16}$$

Percetron은 선형 분리 가능한 문제에 대하여 수렴시 가중치 수정횟수는 다음과 같은 상한선(Upper Bound)을 갖는다[1].

$$k_o = \frac{\beta^2 ||\mathbf{a}||^2}{\gamma^2} \tag{17}$$

여기서

$$\beta = \max_i ||\mathbf{y}_i||, \quad \gamma = \min_i d_i * \hat{\mathbf{a}}^t \mathbf{y}_i \tag{18}$$

이를 Percetron with Margin에 적용하면

$$k_o = \frac{\beta^2 ||\hat{\mathbf{a}}||^2}{b^2} \tag{19}$$

Relaxation 은 다음의 상한선을 갖는다.

$$k_o = \frac{\beta^2 ||\hat{\mathbf{a}}||^2}{(b-\gamma)^2} \tag{20}$$

따라서 Relaxation은 $b=\gamma$ 인 경우 수렴시간이 오래 걸림을 알 수 있다. 식 16과 20에서 γ을 조정하면($\gamma=0,\ 2b$) Percetron과 동등한 유한한 수렴 상한선을 얻을 수 있다.

2.3 Relaxation에 의한 SV 패턴학습법

식 3과 식 15, 16을 비교하면 SVM에서 요구하는 조건과 Percetron with Margin과 Relaxation에서의 학습 종결 조건이 동등함을 알 수 있다. 그러나 완전한 학습을 위하여 SV 패턴에 대한 세밀한 정의한 필요하다.

정의 1. : Support Vector 는 다음의 조건들을 만족하는 훈련 패턴이다.

$$\begin{aligned} &1.\ SVs = \{\mathbf{y}_i : d_i g_a(\mathbf{y}_i) = \gamma \pm \varepsilon > 0\}, \\ &\quad where\ \varepsilon\ is\ a\ positive\ contant \\ &2.\ Cardinality(SVs) = N_s \ge 2, \\ &3.\ |\textstyle\sum_i d_i| < Ns \end{aligned} \tag{21}$$

이것을 풀이하면 1번 조건은 기존의 여유 학습 방법에서 과도한 학습 결과에 만족하여 종결하는 것을 배제한다. 2, 3번 조건은 각 패턴 클래스에 적어도 하나 이상의 SV가 존재해야 한다는 것이다. 상기 3조건을 모두 만족해야 학습이 종료되는 것으로 여유학습 방법인 Relaxation과 Percetron with Margin등을 수정할 것을 본 논문에서 처음으로 제안하는 것이다.

2.3 커널 공간에서의 저밀도 표현과 SV 패턴 학습법

식 12를 α에 관하여 미분하면 다음과 같다.

$$\frac{\partial L(\alpha)}{\partial \alpha} = b\,\mathbf{d} - Y^t Y \alpha = 0 \tag{22}$$
$$\mathbf{d} = Y^t Y \alpha / b$$

여유 b를 1로 놓고 선형 내적 커널을 비선형 내적 커널을 치환하면 다음과 같다.

$$\mathbf{g}_k(\mathrm{Y}) = \mathrm{K}\ \boldsymbol{\alpha} = \mathbf{d} \qquad (23)$$

커널을 패턴 행렬로 커널의 각 행 또는 열은 패턴 벡터로 해석하면 SVM의 QP의 Lagrange 승수 벡터 $\boldsymbol{\alpha}$는 패턴 가중치가 된다.

$$\begin{aligned} g_k(\mathbf{y}_j) &= \boldsymbol{\alpha}^{\mathrm{t}}\,\mathrm{K}(:,j) \\ &= \mathrm{K}(j,:)\,\boldsymbol{\alpha} \end{aligned} \qquad (24)$$

커널 공간과 커널 판별 함수, 커널 초평면 등이 정의된다. 따라서 선형 판별 함수의 모든 학습 방법들을 그대로 커널 공간에 적용할 수 있다. 입력 공간의 판별함수 식 2를 벡터 함수로 정리하면 다음과 같다.

$$\mathbf{g}_a(\mathrm{Y}) = \mathrm{Y}^{\mathrm{t}}\,\mathbf{a} = \mathbf{d} \qquad (25)$$

입력 공간은 미지수의 개수가 m+1이고 방정식은 N 개인 과잉 결정 시스템(Over Determined System)이고(N >> m+1)이고 커널 공간은 미지수와 방정식 개수가 동일한 N개로 완전 결정 시스템(Exactly Determined System)이다. 따라서 커널공간에서는 거의 모든 경우 식 23의 해가 존재한다. 모든 가중치 요소가 비영(Nonzero)이고 모든 패턴이 SV 패턴이 될 수 있다. 저밀도 표현 학습을 위하여 가중치가 식 13을 만족하도록 한다.

커널 최대 강하/ 상승(Steepest Decent/Ascent) 방법(KSD)은 다음과 같다.

$$\nabla\boldsymbol{\alpha} = \eta(\mathbf{d} - \boldsymbol{\alpha}^{t}\mathrm{K}) \qquad (26)$$

각 패턴 $\mathbf{y}_j$ 에 대하여 α_j를 학습하는 KA와 SVMseq[7, 8]는 최대 강하 학습법이 아니고 좌표 강하(Coordinate Descent)방법이다[3].

$$\nabla\alpha_j = \eta\,(d_j - \boldsymbol{\alpha}^{t}\,\mathrm{K}(:,\mathrm{j})) \qquad (27)$$

이에 반하여 RBF 신경회로망의 델타 규칙은 커널 LMS (KLMS) 방법이다.

$$\nabla\boldsymbol{\alpha} = \eta\,(d_j - \boldsymbol{\alpha}^{t}\,\mathrm{K}(:,\mathrm{j}))\,\mathrm{K}(:,\mathrm{j}) \qquad (28)$$

커널 Percetron(KP)은 이미 퍼텐셜 함수(Potential Function) 학습법으로 소개되었다[1].

$$\nabla\boldsymbol{\alpha} = \eta\, d_j\,\mathrm{K}(:,\mathrm{j}) \qquad (29)$$

커널 공간에서 커널 초평면에 대한 저밀도 표현과 SV 패턴 학습을 위한 수정된 커널 Relaxation방법(KR) 은 다음과 같다.

$$\begin{aligned} &do \\ &\quad //relaxation \\ &\text{for } k=1,\ldots,N \\ &\text{if}(d_k\ g_k(\mathbf{y}_k) < \gamma-\varepsilon) \\ &\quad \boldsymbol{\alpha}(k+1) = \boldsymbol{\alpha}(k) + \\ &\eta(k)\frac{[d_k * b - \boldsymbol{\alpha}^{t}\,\mathrm{K}(:,k)]}{\|\mathrm{K}(:,k)\|^2}\,\mathrm{K}(:,k); \\ &else\text{if}(d_k\ g_k(\mathbf{y}_k) > \gamma+\varepsilon)\ \alpha_k = 0; \\ &end\ \text{for} \\ &\quad //SV\ maintenance \\ &\text{for } k=1,\ldots,N \\ &\quad update\ SVs\ condition\ \text{in}\ equation\ 21; \\ &end\ \text{for} \\ &until\ d_k\ g_k(\mathbf{y}_k) \geq \gamma > 0, k=1,\ldots,N \\ &\quad \text{and}\ equation\ 21\ holds\ true. \end{aligned} \qquad (30)$$

이상으로 Duda의 고전에서 제안한 퍼텐셜 함수 학습법들을 모두 완성하게 되었다[1].

2.4 실험 결과

첫 번째 실험은 표준 XOR 문제에서 41개의 데이터를 추가하여 기존의 4개의 표준 패턴만을 SV 패턴으로 선정하는지 조사하여 저밀도 학습이 가능한가를 KA와 수정된 KR QP결과를 참고대상(Benchmark)으로 하여 평가하였다. 그림 1, 2, 3을 보면 예상대로 KA는 19개의 SV패턴을 학습하였다. 수정된 KR과 QP는 예상된 최소 개수인 4개의 SV 패턴과 경계면을 동등하게 학습하였다.

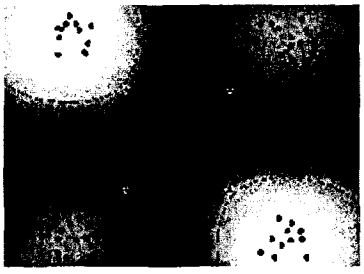

그림 1. 확장 XOR 문제에 대한 QP 학습결과

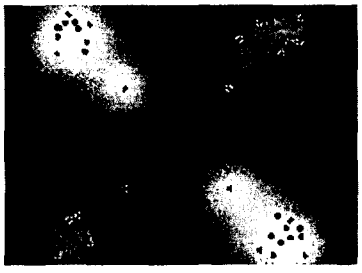

그림 2. 확장 XOR 문제에 대한 KA 학습결과

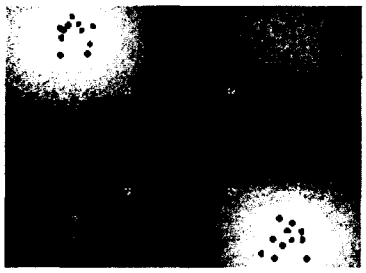

그림 3. 확장 XOR 문제에 대한 수정된 KR 학습결과

두 번째 실험은 일반화 능력 평가이다. 신경회로망 분류기 표준 평가 데이터인 SONAR 분류 문제를 갖고 실행하였다[9]. SONAR 데이터의 개수는 208개로 금속 원통의 지뢰와 암석에서 반사되는 SONAR 파형 패턴으로 차원은 60이다. 처음 104개의 데이터는 훈련데이터로 나머지 104개는 평가 데이터로 쓰였다. 표 1은 수정된 KR 방법과 KLMS, KA의 수행 능력을 보여주고 있다.

표 1. SONAR 분류 문제에 대한 수정된 KR, KLMS, KA 학습결과.

Algorithm	Epoch	Training /Test Data Performance	# of SV Patterns
KR(b=1, γ =2)	30	100.0/94.23	100
KR(b=1, γ =1.5)	80	100.0/94.23	100
KR(b=1, γ =1.0)	620	94.23/91.35	94
KR(b=1, γ =0.5)	740	89.42/84,62	89
KR(b=1, γ =0.0)	50	80.77/87.50	62
KLMS	20	100.0/94.23	101
KR	10	100.0/95.19	99

여기서 b와 γ는 식 16, 18 과 20에서 언급한 인수로서 KR의 수렴과 학습 능력에 많은 영향을 미치고 있음을 보여 준다. 수행능력에서는 3 방법 모두 94%이상이며 동등하다.

III. 결 론

본 논문에서 패턴 분류 문제에 대한 기법 중 커널 공간(Kernel Space) 방법으로 패턴 가중치를 학습하는 과정에서 가중치가 영(zero)인 훈련 패턴이 되도록 많이 생성되도록 하는 것이 목표인 저밀도 표현(Sparse Representation)의 새로운 학습 방법을 제안하였다.

감사의 글 : 본 논문은 과학기술부 · 한국과학재단 지정 여수대학교 설비자동화 및 정보시스템 연구개발센터의 연구비지원에 의한 것임.

IV. 참고문헌

[1] R. O. Duda et al. , *Pattern Classification and Scene Analysis,* John Wiley & Sons, Inc., 1973.

[2] S. Haykin, *Neural Networks - A Comprehensive Foundation,* 2nd Ed., Prentice-Hall, Inc., 1999.

[3] D. G. Luenberger, *Linear and Nonlinear Programming,* 2nd Ed. Addison-Wesley Publishing Company Inc., 1984.

[4] B. E. Boser, et al., "A Training Algorithm for Optimal Margin Classifiers," in Proc. of the 5th Annual Workshop on Computational Learning Theory 5, pp. 144-152, Pittsburgh, 1992.

[5] E. Osuna, et al., "Training Support Vector Machines : an Application to Face Detection." CVPR'97, 1997.

[6] J. C. Platt. "Fast Training of Support Vector Machines using Sequential Minimal Optimization," chapter 12. In Scholkopf et al. [32], 1998.

[7] T. T. Friess et al, " The Kernel-Adatron Algorithm : a Fast and Simple Learning Procedure for Support Vector Machines," Proc. 15th Intl. Conf. on Machine Learning, Morgan-Kaufman, 1998

[8] S. Vijayakumar et al, " Sequential Support Vector Classifiers and Regression", Int. Conf. Soft Computing, pp. 610-619 1999.

[9] Y. Li et al.. "The relaxed online maximum margin algorithm." In Advances in NIPS 13, 1999.