한국지능시스템학회 논문지 2010, Vol. 20, No. 3, pp. 375-379

유전자 알고리즘을 이용한 강인한 Support vector machine 설계

Design of Robust Support Vector Machine Using Genetic Algorithm

이희성 · 홍성준 · 이병윤 · 김은태*

Heesung Lee, Sungjun Hong, Byungyun Lee and Euntai Kim

연세대학교 전기전자공학부

요 약

Support vector machine (SVM)은 튼튼한 이론적 배경을 가지고 있고 구조적 위험을 성공적으로 최소화하기 때문에 추천가 시스템과 같은 다양한 패턴 인식 분야에서 사용되고 있다. 하지만 SVM이 초평면을 결정할 때 이상점들은 margin 손실들을 가지고 있기 때문에 이들은 초평면을 결정하는데 매우 중요한 역할을 하고 있다. 그 이유로 SVM은 이상점들에게 매우 민감한 문제점을 갖는다. 강인한 SVM을 위해 우리는 이상점들의 margin 손실의 최대치를 제한하지만 이것은 non-convex 최적화 문제를 포함한다. 따라서 본 논문에서는 non-convex 최적화 문제에 적합한 유전자 알고리즘을 이용하여 강인한 SVM을 설계하는 방법을 제안한다. 제안하는 알고리즘의 우수성을 보여주기 위하여 UCI repository에서 선택된 여러 데이터베이스들을 이용한 실험을 수행하였다.

키워드 : SVM, Robust SVM, 유전자 알고리즘, 이상점, UCI repository

Abstract

The support vector machine (SVM) has been widely used in variety pattern recognition problems applicable to recommendation systems due to its strong theoretical foundation and excellent empirical successes. However, SVM is sensitive to the presence of outliers since outlier points can have the largest margin loss and play a critical role in determining the decision hyperplane. For robust SVM, we limit the maximum value of margin loss which includes the non-convex optimization problem. Therefore, we proposed the design method of robust SVM using genetic algorithm (GA) which can solve the non-convex optimization problem. To demonstrate the performance of the proposed method, we perform experiments on various databases selected in UCI repository.

Key Words : SVM, Robust SVM, GA, outlier, UCI repository

1. 서 론

패턴 인식 시스템은 최근 관심을 끌고 있는 추천가 시스템이나 바이오 인식 시스템과 같이 응용분야 다양하고 활용도가 방대하기 때문에, 많은 인식 이론 및 기술들이 여러 연구자, 공학자들에 의해 개발되고 있다. 특히 패턴을 고차원 특징 공간으로 사상시킬 수 있다는 점과 대역적으로 최적의 식별이 가능하다는 점 때문에 Support vector machine (SVM)은 최근 들어 많이 연구되고 있다. SVM의 가중치는 선형 부등 조건을 가진 2차 계획법(Quadratic Programming) 문제를 해결함으로써 얻어지게 된다. 또한 전통적인 대부분의 패턴 인식 시스템들은 학습 집단을 이용하여 학습 오류를 최소화하는 경험적 위험 최소화 방법을 기초로 하고 있지만, SVM은 분류 오류 확률을 최소화 하는 구조적인 위험 최소화 방법에 기초하고 있다[1].

하지만 일반적으로 SVM은 이상점(outlier)들에 민감하다[2]. SVM이 초평면(hyperplane)을 결정할 때 이상점들은 최대 margin 손실을 갖기 때문에 초평면은 이상점들의 영향을 많이 받게 되고 그 결과 SVM의 성능이 하락한다. 따라서 이상점들에 강인한 SVM들이 많이 개발되고 있다[2, 3]. 강인한 SVM들은 이상점들의 영향을 줄이기 위해 손실함수를 변화시킨다. 하지만, 변화된 손실 함수들은 non-convex 형태이기 때문에, 기존의 방법들은 이를 해결하기 위해 일반적인 convex 최적화 방법을 사용하지 못하고 대신 복잡한 다른 최적화 방법들을 이용하고 있다. 하지만 이들은 계산량이 크고 국지적 최적점에 도달할 가능성이 크다.

SVM의 가중치들은 패턴들의 특징 수에 비례한다. 특히 최근에는 복잡하고 크기가 매우 큰 데이터들을 처리해야 되기 때문에 이에 따라 결정해야 되는 SVM의 가중치들도 많

접수일자 : 2010년 2월 23일
완료일자 : 2010년 5월 30일
* 교신 저자
본 연구는 한국콘텐츠진흥원의 2010년도 문화콘텐츠산업기술지원사업의 "웹 미디어의 IPTV 서비스 활용을 위한 콘텐츠 동적 결합과 콘텐츠 생성 기술" 과제 연구 결과로 수행된 결과입니다.

아진다. 또한 복잡한 데이터에는 이상점들이 많이 포함될 가능성이 크다. 따라서 이상점들이 존재하는 상태에서 SVM의 많은 가중치들을 효율적으로 결정해야 한다. 하지만 기존의 방법들은 계산량이 크고 복잡한 최적화 방법을 이용하고 있기 때문에 효율적인 연산을 못하고 있다.

자연 선택과 자연 발생의 과정을 기초로 다수의 개체를 동시에 진화시켜 가면서 최적의 해를 찾는 유전자 알고리즘은 많은 최적화 문제에서 사용되고 있다[4]. 특히 유전자 알고리즘은 non-convex 최적화 문제를 효율적으로 해결할 수 있다[5]. 본 논문에서는 이상점들의 영향을 적게 받는 강인한 SVM을 효율적으로 설계하기 위하여, non-convex 최적화 문제를 유전자 알고리즘을 이용하여 해결한다.

본 논문의 구성은 다음과 같다. 2장에서는 유전자 알고리즘과 SVM을 설명하고, 3장에서는 유전자 알고리즘을 이용하여 이상점들에 강인한 SVM을 설계하는 방법을 제안한다. 4장에서는 제안된 시스템의 효용성을 보이기 위한 실험과 그의 고찰을 한 뒤 마지막으로, 5장에서는 결론과 추후 과제에 대한 설명을 한다.

2. 배경지식

2.1 Genetic Algorithm

유전자 알고리즘은 진화 알고리즘의 한 부류이다. 유전자 알고리즘은 생명체의 유전 및 진화 과정을 전산학적으로 모델링(modeling)한 기계학습의 방법으로, 탐색해야 할 공간이 매우 넓은 경우 유용하게 사용되는 탐색 및 최적화 기법이다. 유전자 알고리즘의 가장 큰 특징은 잠재적 해인 염색체(chromosome)들의 집단을 이루어 만들어진 해의 집단(population)을 운용한다는 것이다. 각 염색체는 적자생존의 법칙에 의하여 상대적으로 우수한 것이 살아남을 확률이 크며, 또한 유전 연산자에 의하여 진화과정을 거치게 된다. 적합도가 높은 개체의 집합이 선택(selection)되어 다음 세대의 자손을 생성하는 부모가 되며, 자손은 교차(crossover), 돌연변이(mutation)의 유전 연산자를 통해 생성된다[4]. 일반적인 유전자 알고리즘의 절차는 다음과 같다.

```
procedure GA
  t=0;
  initialize P(t);
  evaluate P(t);
  while termination condition not satisfied do
    t=t+1;
    select P(t) from P(t-1);
    recombine and mutate P(t);
    evaluate P(t);
  end
end
```

초기 염색체들의 집단인 P(0)를 생성한다. 초기 집단은 해 공간 내에서 무작위로 분포되도록 선택되거나 경험적인 방법으로 선택된다. 그리고 각 염색체들의 적합도를 평가한 다음 평가된 적합도에 비례하여 선택된 염색체들을 P(t)에 복사한다. P(t)안의 염색체들에게 유전자 연산을 적용시킨 후 P(t)를 재생성한다. 그리고 그 결과를 바탕으로 자식 세대 P(t+1)을 생성한다[6]. 본 논문에서는 SVM의 최적화를 위해 실수 코딩 유전자 알고리즘(real-coded genetic algorithm)[7]을 사용하였다. 염색체가 0과 1만 갖는 이진 코딩 유전자 알고리즘과는 달리 실수 코딩 유전자 알고리즘은 정밀도를 높이거나 탐색구간을 확대하기 위해 염색체의 길이를 증가할 필요가 없다. 또한 부호화 과정과 복호화 과정이 없으므로 염색체의 유전자들이 SVM의 가중치와 bias값들을 일대일의 실수 값으로 표현할 수 있고 상당히 넓은 영역을 표현할 수 있다.

2.2 Support Vector Machine

기본적으로 SVM은 입력 패턴들의 교사학습방법을 통하여 학습 데이터들을 두 클래스로 분류한다. 학습 데이터들의 클래스는 초평면에 의해 결정되고 다음과 같은 방법으로 SVM은 초평면을 계산한다. 학습 데이터가 다음과 같이 구성되어 있다고 할 때,

$$\{x_1, x_2, ..., x_N\} (x_i \in R^n) \tag{1}$$

각 학습 데이터들은 각 클래스 레이블을 표시하는 목표변수 y_i를 갖는다. 입력 공간(input space)에서 서로 다른 클래스를 분류하는 초평면을 찾는 것은 매우 제한적이기 때문에 입력 공간을 더 높은 차원의 특징 공간(feature space)으로 사상시키고 SVM은 이 특징 공간에서 다음과 같은 최적의 초평면을 찾는다.

$$W \bullet \psi(x) + w_0 = 0 \tag{2}$$

여기에서 $\psi(\bullet)$는 입력 공간에서 특징 공간으로의 사상함수이고 W와 w_0는 초평면의 가중치와 bias이다. 이 논문에서 사상함수로 우리는 다음과 같은 polynomial mapping을 사용하였다.

$$\begin{aligned} z_i &= \psi(x_j) \\ &= [x_{j1}, ..., x_{jn}, x_{j1}^2, ..., x_{jn}^2, \\ &\quad \sum_{k=1}^{n} x_{j1}x_{jk}, ..., \sum_{k=1}^{n} x_{jk}x_{j1}] \end{aligned} \tag{3}$$

여기서 $z_i \in R^M$이다. 특징 공간으로 사상 된 데이터들이 선형분류가 되지 않을 경우 여유 변수(slack variable) ξ_i를 도입하고 초평면은 다음 식의 해로 결정할 수 있다.

$$\begin{aligned} &\min \frac{1}{2} W^T W + C\sum_{i=1}^{N} \xi_i \\ &s.t.\ y_i(W \bullet \Psi(x_i) + w_0) \geq 1 - \xi_i \\ &\qquad\qquad \text{for } i = 1, ..., N \end{aligned} \tag{4}$$

여기서 $W \in R^M$, $\xi_i \geq 0$이고 C는 마진 최대화와 분류 규칙 위반 사이의 균형을 맞추는 regularization parameter이다[8]. 일반적인 SVM에서 여유 변수의 값은 초평면과 데이터와의 거리에 비례하는데 이상점들은 보통 이 거리가 매우 크기 때문에 일반적인 SVM에서는 이상점들이 support vector들로 간주되고 SVM들은 이상점들에 매우 민감하게 된다[2, 3].

3. 유전자 알고리즘을 이용한 강인한 Support Vector Machine 설계

일반적인 SVM은 어떤 학습 데이터의 여유 변수의 값이

아무리 크더라도 그 값을 초평면을 결정하는 데 그대로 사용한다. 이상점들은 여유 변수의 값이 매우 크기 때문에 일반적인 SVM들은 이상점의 영향을 크게 받는다. 본 논문에서는 이상점들이 갖는 여유 변수의 최대값을 제한하는 다음과 같은 새로운 손실 함수 $K(\xi_i)$를 도입하여 이상점들의 영향을 줄인다.

$$\min \frac{1}{2} W^T W + C\sum_{i=1}^{N} K(\xi_i)$$
$$s.t.\ y_i(W \bullet \Psi(x_i) + w_0) \geq 1 - K(\xi_i) \quad \text{for } i = 1,...,N \tag{5}$$

여기서 $K(\xi_i) = \begin{cases} \xi_i, & \xi_i < \alpha \\ \alpha, & \xi_i \geq \alpha \end{cases}$ 이고, α는 여유 변수가 가질 수 있는 최대값이다. $K(\xi_i)$는 그림 1과 같은 형태를 지닌다.

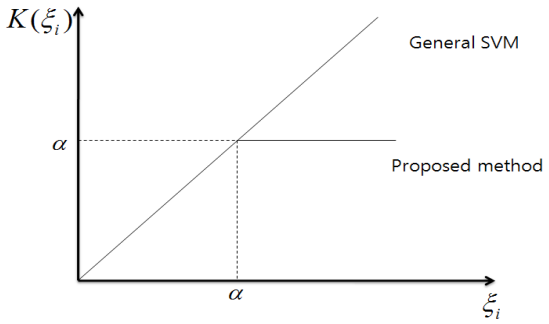


그림 1. 제안하는 손실 함수.
Fig. 1. Proposed loss function.

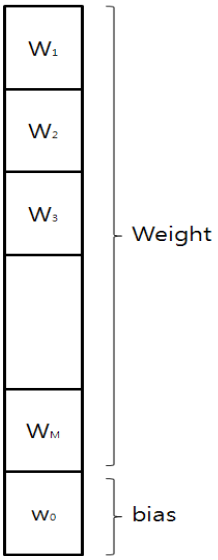


그림 2. 염색체의 구성.
Fig. 2. Structure of the chromosome.

그림 1에서 볼 수 있듯이, 일반적인 SVM에서 사용하는 손실 함수와는 달리 $K(\xi_i)$는 non-convex형태이기 때문에 일반적인 convex 최적화 방법으로는 식 (5)를 통해 초평면의 가중치 W와 bias w_0를 구할 수 없다. 따라서 우리는 non-convex 최적화에서도 효율적인 유전자 알고리즘을 사용하여 초평면을 계산한다. 제안하는 시스템의 염색체의 구성은 그림 2와 같다.

염색체의 실수 코딩 유전자들은 초평면의 가중치와 bias에 일대일 대응되어 효율적으로 식 (5)의 최적화 문제를 해결한다. 교차 연산자로 산술 교차(arithmetic crossover)와 휴리스틱 교차(heuristic crossover)를 사용하였고 돌연변이로는 균등 돌연변이(uniform mutation)와 경계 돌연변이(boundary mutation)[5]를 각각 사용하였다. 유전자 파라미터는 표 1과 같이 사용하였다.

표 1. 유전자 알고리즘의 파라미터들.
Table 1. Parameters for Genetic Algorithm.

Parameter	Value
Maximum generation number	500
Crossover rate	0.6
Mutation rate	0.05
Population size	2000

4. 실 험

제안하는 알고리즘을 평가하기 위하여 Pima, German, Tic-Tac-Toe 데이터베이스들을 UCI Machine learning repository[9]에서 선택하였다. UCI repository는 패턴인식 분야에서 알고리즘들의 성능을 평가하기 위하여 많이 사용되고 있다. 사용하는 데이터베이스를 표 2에 요약하였다.

표 2. 사용된 Database들.
Table 2. Used Databases.

종 류	Pima	German	Tic-Tac-Toe
Number of instances	768	1000	958
Number of features	8	24	9
Number of classes	2	2	2

우리는 각 데이터베이스를 60대 40의 비율로 학습 데이터와 테스트 데이터로 나누었다. 일반적인 SVM은 식 (4)를, 제안하는 SVM은 식 (5)를 각각 최적화하는 초평면을 찾기 위해 학습 데이터를 사용하였고, 알고리즘들의 평가를 위해 테스트 데이터를 사용하였다. 세 개의 데이터베이스들을 이용한 일반적인 SVM과 제안하는 SVM의 결과를 표 3에 도시하였다.

표 3. 제안된 방법의 인식 결과.
Table 3. Results of the proposed method.

Databases	General SVM	Proposed method	Diff
Pima	65.36	64.71	0.65
German	69.00	70.00	1.00
Tic-Tac-Toe	65.29	65.29	0
Average	66.55	66.67	0.55

그리고 일반적인 SVM의 인식률을 1로 했을 때 제안하는 SVM의 상대적인 인식률을 그림 3에 보였다.

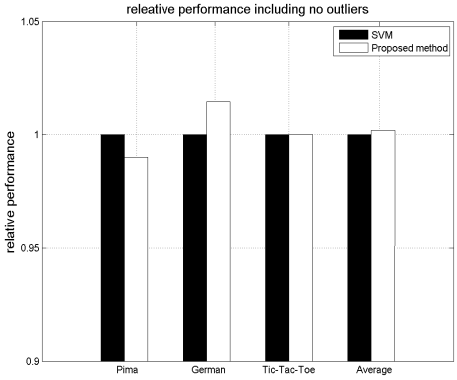


그림 3. 제안된 방법의 상대적인 인식 결과.
Fig. 3. Relative results of the proposed method.

제안하는 알고리즘의 이상점에 대한 강인함을 보이기 위해 우리는 각 클래스에서 데이터들의 10%를 임의로 선정하여 그 샘플들의 클래스의 레이블을 바꾸었다. 따라서 임의로 선택된 데이터들은 이상점의 역할을 한다. 수정된 데이터베이스들을 이용하여 일반적인 SVM과 제안하는 알고리즘의 결과를 비교하였고 그 결과를 표 4와 그림 4에 도시하였다.

표 4. 10% 이상점이 포함되었을 때 제안된 방법의 인식 결과.
Table 4. Results of the proposed method including 10% outliers.

Databases	General SVM	Proposed method	Diff
Pima	35.29	60.78	25.49
German	67.00	70.00	3.00
Tic-Tac-Toe	34.71	65.29	30.58
Average	45.67	65.36	19.69

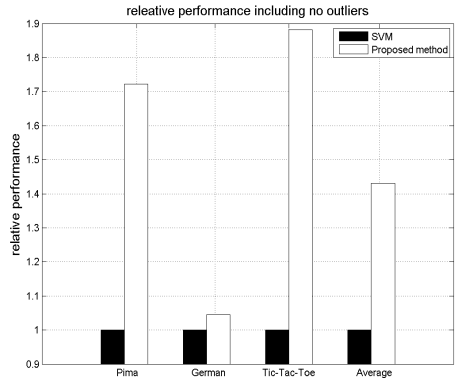


그림 4. 10% 이상점이 포함되었을 때 제안된 방법의 상대적 인식 결과.
Fig. 4. relative results of the proposed method including 10% outliers.

표 3과 그림 3을 통해 일반적인 SVM과 제안하는 알고리즘은 이상점이 없을 경우에는 비슷한 성능을 보임을 알 수 있다. 하지만 이상점이 있을 경우에는 제안하는 SVM이 매우 좋은 성능을 갖고 있음을 표 4와 그림 4를 통해 알 수 있다. 따라서 일반적인 SVM에 비해 제안하는 SVM은 이상점들에 매우 강인함을 알 수 있다.

5. 결 론

구조적 위험을 성공적으로 최소화하는 SVM은 다양한 패턴인식 분야에서 사용되고 있다. 하지만 이상점들은 margin 손실들을 가지고 있어 SVM이 초평면을 결정하는데 중요한 역할을 한다. 그 이유로 일반적인 SVM들은 이상점들에 매우 민감하다. 강인한 SVM을 위해 우리는 이상점들의 margin 손실의 최대치를 제한하는 함수를 이용하였다. 하지만 이 함수는 non-convex 형태이기 때문에 본 논문에서는 non-convex 최적화 문제에 적합한 유전자 알고리즘을 이용하여 강인한 SVM을 설계하는 방법을 제안하였다. 제안하는 알고리즘과 일반적인 SVM을 UCI repository에서 선택된 여러 데이터베이스들에 적용시킨 결과 제안하는 알고리즘의 우수성을 볼 수 있었다. 향후 본 논문에서 제안한 알고리즘을 추천가 시스템에 실제 적용해볼 계획이다.

참 고 문 헌

[1] 조규행, 박윤식, 이계환, 장준혁, “효과적인 특징 벡터를 도입한 Support Vector Machine기반 음성 검출기,” *Telecommunications Reviews,* vol. 18, no. 2, pp. 362-370, 2008.

[2] L. Wang, H. Jia, and J. Li, “Training robust support vector machine with smooth Ramp loss in the primal space,” *Neurocomputing*, vol. 71, pp. 3020-3025, 2008.

[3] L. Xu, K. Crammer, and D. Schuurmans, “Robust support vector machine training via convex outlier ablation,” in Proc. *the 21st National Conference on Artificial Intelligence*, pp. 536-542, 2006.

[4] H. Lee, E. Kim, and M. Park, “A genetic feature weighting scheme for pattern recognition,” *Integrated Computer-Aided Engineering*, vol. 14, no. 2, pp. 161-171, 2007.

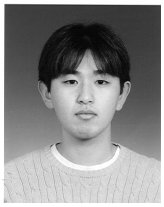
[5] C. R. Houck, J. A. Joines, and M. G. Kay, “A genetic algorithm for function optimization: A Matlab implementation,” NCSU-IE Technical Report 95-09, North Carolina State University, 1995.

[6] 이희성, 이제헌, 김은태 “GAVaPS를 이용한 다수 K-Nearest Neighbor classifier들의 feature 선택,” *한국지능시스템학회 논문지*, vol. 18, no. 6, pp. 871-875, 2008.

[7] 방연근, 이철희, “강화된 유전알고리즘을 이용한 이중 동조 기반 퍼지 예측시스템 설계 및 응용,” *전기학회 논문지*, vol. 59, no. 1, pp. 184-191, 2010.

[8] 손태식, 서정우, 서정택, 문종섭, 최홍민, “Support Vector Machine 기반 TCP/IP 헤더의 은닉채널 탐지에 관한 연구,” *정보보호학회 논문지*, vol. 14, no. 1, pp. 35-45, 2004.
[9] P. M. Murphy and D. W. Aha, “UCI Repository for Machine Learning Databases,” *Technical report, Dept. of Information and Computer Science,* Univ. of California, Irvine, Calif., 1994.

저 자 소 개



이희성 (Heesung Lee)
2003년 : 연세대학교 전기전자공학부 졸업 (공학사)
2005년 : 연세대학교 전기전자공학부 석사과정 졸업(공학석사)
2005년～현재 : 동 대학원 전기전자공학과 박사과정

관심분야 : Computational intelligence, 로봇 비전, 패턴 인식
E-mail : 4u2u@yonsei.ac.kr

홍성준 (Sungjun Hong)
2005년 : 연세대학교 전기전자공학부 졸업 (공학사)
2005년～현재 : 동 대학원 전기전자공학과 석박사통합과정

관심분야 : 기계 학습, 생체 인식
E-mail : imjune@yonsei.ac.kr

이병윤 (Byungyun Lee)
2007년 : 연세대학교 전기전자공학부 졸업 (공학사)
2009년～현재 : 동 대학원 전기전자공학과 석사과정

관심분야 : 신호 처리, 생체 인식
E-mail : yuni4u@yonsei.ac.kr

김은태 (Euntai Kim)
1992년 : 연세대학교 전자공학과 졸업 (공학사)
1994년 : 연세대학교 전자공학과 석사과정 졸업(공학석사)
1999년 : 연세대학교 전자공학과 박사과정 졸업(공학박사)

1999년 3월～2002년 2월 : 국립한경대학교 제어계측공학과 조교수
2002년 3월～현재 : 연세대학교 전기전자공학부 부교수
2003년 : University of Alberta, visiting researcher
1998년～현재 : IEEE TFS, IEEE SMC, IEEE CAS, FSS 등에서 심의위원 활동 중
2003년 : 대한 전자공학회 해동상 수상

관심분야 : Computational intelligence, 지능형 로봇
Phone : +82-2-2123-2863
E-mail : etkim@yonsei.ac.kr