마진 벡터를 이용한 앙상블 SVM의 학습

박상호, 김태순, 박선, 강윤희, 이주홍
인하대학교 컴퓨터공학과
e-mail: parksangho@datamining.inha.ac.kr

Ensemble SVM Learning Using Margin Vector

Sang-Ho Park, Tae-Soon Kim, Sun Park,
Yun-Hee Kang, Ju-hong Lee
Dept. of Computer Science and Engineering, Inha University

요 약

SVM은 일반화된 높은 분류 정확률을 보인다. 그러나, SVM은 데이터의 양이 커질수록 높은 시간 공간적 복잡성 때문에 근사화 알고리즘(Approximation Algorithm)을 이용한다. 이러한 접근 방법은 실제 구현에서 높은 시간 공간적 복잡성을 요구하여 분류 정확률과 효율성을 낮아지게 한다. 따라서, 본 논문은 SVM을 앙상블 구조로 구성하여 분류 정확률을 높이고, 분류자의 최적 하이퍼플랜(Optimal Hyperplane)결정을 위해 마진 벡터만을 이용하여 시간 공간적 문제를 해결하였다. 실험 결과, 본 논문에서 제시한 방법이 단일 SVM을 이용한 방법보다 높은 분류 정확률과 높은 효율성을 가짐을 보여준다...

1. 서론

SVM(Support Vector Machine)은 최소의 일반화 에러를 가진 최적의 분류 평면을 결정하는 기법이다 [3]. 이러한 낮은 일반화 분류에러를 가지는 SVM은 다양한 데이터의 분류(Classification)에 매우 효과적이다. 하지만, SVM의 학습은 높은 시간 공간적 복잡도 때문에 근사화 알고리즘(Approximation Algorithm)을 이용하는 단점이 있다. 이러한 접근 방법은 실제 구현에서 분류 정확률과 효율성을 낮아지게 한다. 이러한 문제점을 해결하기 위해서 본 논문은 SVM을 앙상블 구조로 구성하고, 전처리(Preprocessing)단계에서 마진 벡터(Margin Vector)들을 추출하여 앙상블 구조의 SVM들을 학습시킨다.

본 논문에서는 단일 SVM과 마진벡터를 이용한 앙상블 구조의 SVM의 시간 공간적 복잡도와 분류 정확도를 비교한다.

본 논문은 2장에서는 관련 연구에 대하여 알아보고, 3장에서는 마진 벡터를 이용한 앙상블 SVM에 대하여 설명하고, 4장에서는 실험, 5장에서는 결론을 살펴본다.

2. 에이전트 개발도구의 요구사항

2.1 SVM

본 논문에서 사용하는 SVM은 Vapnik[2]에 의해서 1995년 이원문제를 해결하기 위해서 제안된 알고리즘이다. 먼저, SVM은 데이터집합의 입력이 X_i 이고, 출력이 y_i인 학습데이터의 집합을 D라고 하면 SVM은 매핑 함수(Φ)에 의해서 입력공간의 데이터 (X_i)를 특징 공간으로 매핑(Mapping)한다.

$$W^T\Phi(X_i)+b\leq 1(y_i=-1\text{인 경우}) \quad (1)$$

$$W^T\Phi(X_i)+b\geq 1(y_i=+1\text{인 경우}) \quad (2)$$

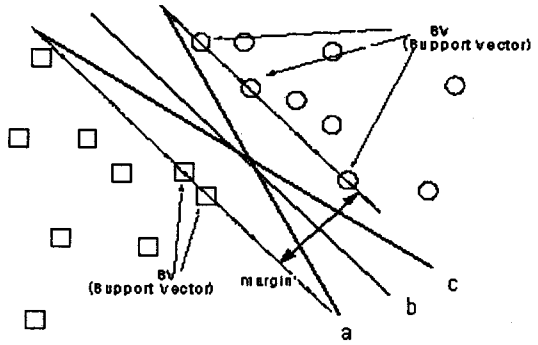

그림1. 2차 공간에서의 SVM 결정 경계

특징 공간에서 데이터가 선형적으로 분리 가능하면 학습 데이터 집합의 모든 요소들에 대하여 식(1)과 (2)를 만족하는 벡터 W와 스칼라 b가 존재하여 두 클래스를 분리할 수 있는 하이퍼플랜(Hyperplane, $W^T\Phi(X)+b$)들을 만들 수 있다. 이들 중에서 양과 음의 구역에 있는 데이터들의 분리를 가장 명확하게 하는 것을 최적의 하이퍼플랜으로 결정한다. 예를 들면, 그림1의 하이퍼플랜 a, b, c중에서 각각의 하이퍼플랜을 기준으로 양과 음의 두 영역에 가장 근접한 데이터들의 거리(Margin: $\frac{1}{||W||^2}$)를 최대화시키는 b가 최적의 하이퍼플랜으로 결정된다.

즉, 선형으로 분리 가능한 경우에는 최적의 하이퍼플랜을 위한 W를 $||W||^2$의 최소화 과정을 통해 결정하고, 식 (3)의 선형결정함수를 이용하여 최적의 선형함수를 결정한다.

$$f(x) = sign(\sum_{i=1}^{l} y_i \alpha_i \cdot (X \cdot X_i) + b) \quad (3)$$

선형적으로 분리할 수 없는 경우에는 입력공간을 분리하는 비선형 결정면을 이용한다. 하지만 비선형 결정면의 식을 분석적으로 계산해낸다는 것은 어려운 일이므로, 다항식, RBF(Radial Basis Function), 다층 퍼셉트론(Multi-Layer Perceptron)등의 커널함수를 사용하여 입력 벡터X를 고차원 특징공간으로 매핑한 후, 선형의 경계선을 찾는 문제로 전환한다. 이처럼 커널함수를 사용하면은 입력벡터를 특징공간으로 투영시킴으로써 내적에 대한 계산만을 하므로 계산이 간편해진다. 결국, 입력공간에서 식(4)의 비선형 결정함수를 이용하여 최적의 선형 함수를 결정한다.

$$f(x) = sign(\sum_{i=1}^{l} y_i \alpha_i \cdot K(X, X_i) + b) \quad (4)$$

y_i는 학습데이터의 레이블, α_i는 랑그랑즈 승수, K( ) 는 커널함수, X는 입력데이터, X_i는 SV(Support Vector), b는 bias이다.

2.2 앙상블

앙상블(Ensemble)[2]이란 서로 다른 여러 개의 분류기들의 출력을 통합하여 최종 분류하는 일종의 복수 분류기 시스템(Multiple Classifier System:MSC)을 말한다.

Hansen[1]은 하나의 분류기만을 이용한 경우보다 앙상블 구조를 가진 분류기들을 이용한 경우가 분류 정확률이 높다는 것을 보였다.

분류 성능을 향상시키기 위한 기본적 앙상블 알고리즘은 다음과 같다. 먼저, 적은 양의 학습데이터를 가지고 학습되어진 분류자(Weak Learner)들을 생성하고 학습시킨다. 학습 앙상블 알고리즘을 이용하여 예측을 복잡한 예측으로 통합시킨다. 다수결의 원칙은 통합 전략 중에서 가장 간단한 방법이다. 앙상블을 구성하는 분류자들의 출력이 레이블된 형태로 이루어질때에 각 클래스마다 모든 분류자의 레이블에 대한 Vote의 총합을 비교하여 최대의 Vote를 가진 클래스를 최종 분류 결정으로 정하는 원칙이다.

3. 마진 벡터를 이용한 앙상블 SVM

본 논문은 SVM의 효율성을 높이기 위해 전처리 단계에서 각 분류기의 학습을 위해 복원 추출된 데이터들의 마진 벡터(Margin Vector)들을 추출하여 앙상블 구조의 SVM의 학습에 이용한다.

3.1 마진 벡터의 추출

마진 벡터는 두 개의 클래스 데이터들이 존재할 때에, 각각의 클래스 중심점과 데이터들 사이의 거리의 비가 임계값 이상의 값을 가지는 데이터를 말한다. SVM은 관찰자(Supervised)의 입장에서 학습, 테스트된다. 따라서, 모든 데이터는 +1혹은 -1 레이블(Label)을 가지며, 각 레이블에 의해 두 개의 클래스(X, Y)로 구분할 수 있다. 분리된 두 개의 클래스의 데이터들은 마진벡터들의 추출을 위해서 중심점(Center:m)을 찾는다.

선형인 경우 레이블된 두 가지의 패턴들의 중심점(m_x , m_y)을 다음과 같이 정의한다.

$$m = \frac{1}{n}\sum_{i=1}^{n} X_i \quad (5)$$

또한, 두개의 예제들 사이의 특징(Feature)들의 차이를 의미하는 거리(Distance)는 다음과 같이 정의한다.

$$d(X_1, X_2) = \| X_1 - X_2 \|_2 = \sqrt{\sum_{i=1}^{N} (X_1^i - X_2^i)^2} \quad (6)$$

위의 거리를 이용하여 중심거리(Center-Distance)는 두 개의 클래스의 중심을 각각 m_x, m_y라고 하면 X와 m_x의 거리(즉,d_{xx}), y와 m_y의 거리(d_{yy}), X와 m_y의 거리(d_{xy}), y와 m_x의 거리(d_{yx})와 같이 4개의 중심거리가 생긴다.

이때에 처음 두개는 자기중심거리(Self-center distance)라고 하고, 나머지 두개는 상호중심거리(Mutual-center distance)라고 한다.

두개의 클래스 패턴이 주어졌을 경우에 각 클래스의 중심거리비율은 다음과 같이 표현된다.

$$Ratio\ X=d_{xx}/d_{xy} \quad (7)$$

$$Ratio\ Y=d_{yy}/d_{yx} \quad (8)$$

위의 Ratio를 가지고 두 개의 클래스 x, y에 대해 주어진 임계값(Threshold:r_x, r_y)이상의 데이터를 마진 벡터로 결정한다.

비선형인 경우에는 매핑함수(Φ)에 의하여 학습 데이터를 특징공간으로 매핑한다. 특징 공간에서 중심(Center:m_Φ)는 다음과 같이 정의된다.

$$m_\Phi = \frac{1}{n}\sum_{i=1}^{n}\Phi(X_i) \quad (9)$$

(n은 학습 데이터의 수이다.)

커널 함수를 사전에 알지 못하는 경우에 특징 공간에서 한 패턴의 자기중심거리는 다음과 같다.

$$d_{xx}^{H}(X,m_\Phi)=\sqrt{K(X,X)-\frac{2}{n}\sum_{i=1}^{n}K(X,X_i)+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}K(X_i,X_j)} \quad (10)$$

또한, 상호중심거리는 다음과 같다.

$$d_{xy}^{H}(X,m_\Phi')=\sqrt{K(X,X)-\frac{2}{n}\sum_{i=1}^{n}K(X,Y_i)+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}K(Y_i,Y_j)} \quad (11)$$

위의 두 식을 이용하여 특징 공간에서 학습 데이터의 중심거리 비율은 다음과 같이 계산되어진다.

$$Ratio_{\Phi(X)} = d_{XX}^{H} / d_{XY}^{H} \quad (12)$$

$$Ration_{\Phi(X')} = d_{YY}^{H} / d_{YX}^{H} \quad (13)$$

비선형인 경우에도 위의 중심거리 비율을 임계값과 비교하여 임계값이상의 값을 가지는 데이터를 마진 벡터로 결정한다.

3.2 마진 벡터를 이용한 앙상블 SVM

본 논문은 SVM을 앙상블 구조로 구성한다. 학습과정에서 앙상블 구조의 SVM들은 마진 벡터만을 이용하여 여러 개의 분류자를 학습시킨다. 이때에, 여러 개의 분류자 학습을 위해서 전체 학습데이터에서 분류자의 수만큼 랜덤하게 복원 추출(Raplacement Selection)하여, 각 분류자들의 학습을 위한 마진벡터들을 추출한다. 추출된 마진 벡터들을 가지고 각 분류자들을 학습시킨다. 예를 들어, 전체 학습데이터가 T라고 했을 경우에, 먼저 T에서 n개의 데이터를 추출한다. 추출된 n개의 데이터를 가지고 MV_1(Margin Vector_1)을 결정한다. MV_1을 가지고 SVM_1을 학습시킨다. 이와 같은 과정을 분류자의 수만큼 반복하여 앙상블 SVM들을 학습시킨다. 그림2는 마진벡터를 이용한 앙상블 SVM 학습.테스트 과정을 보여준다.

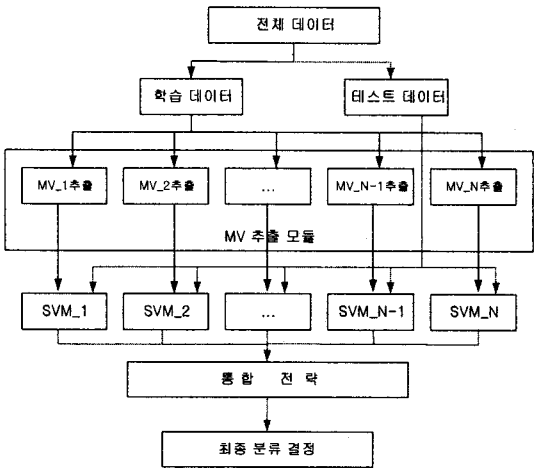

그림2. 앙상블 SVM의 학습/테스트

테스트 단계에서는, 위에서 학습되어진 N개의 분류기를 가지고 똑같은 테스트 데이터에 대하여 분류작업을 한다. 분류되어진 결과는 N×M의 매트릭스를 만든다. 여기에서 N은 각각의 분류자를 의미하고, M은 분류될 클래스를 의미한다. 이렇게 N×M매트릭스 각각의 M에는 분류 결정된 클래스를 '1'로 ,그 외의 분류 결정은 '0'으로 Voting한다. 각 클래스 마다 Vote의 수를 다른 클래스의 Vote와 비교한다. 비교되어진 Vote를 가장 많이 가진 클래스를 다수결의 원칙에 의해서 최종 분류가 이루어진다.

	클래스1	클래스2
분류자_1	00110010011...	11001101100...
분류자_2	11110010110...	00001101001...
분류자_3	11011010101...	00100101010...
분류합	22231030222...	11102303111...
분류결과	11110010111...	00001101000...

표1. 분류 결과를 위한 NM매트릭스

표1은 클래스1과 클래스2로 각각의 데이터에 대하여 3개의 분류자를 이용하여 테스트한 결과를 나타내는 NM 매트릭스이다. 각각의 비트는 하나의 패턴

을 가진 데이터에 대한 분류결과를 나타낸다. 만일, 패턴 Xi가 클래스1에 분류되었다면은 클래스1에 "1"로 Voting될 것이다. 그렇지 않다면은 "0"으로 Voting된다. 따라서, Mj의 위치에 있는 k개의 비트열은 k개의 패턴에 대한 분류결과의 Voting을 나타낸다.

4. 실험

본 논문은 펜티엄Ⅳ 1.5GHz, RAM 256M상의 윈도우2000상에서 C언어로 구현하였다. 실험 자료는 UCI repository중에서 680개의 실험 고객 데이터를 가지고 실험하였다. 학습과 테스트의 비율은 7대3으로 하였다. 커널함수는 RBF(Radial Basis Function)을 사용하였고, 복원추출 비율은 생략(Missing)될 학습 데이터의 경우의 수를 줄이기 위해서 80%로 실험하였다. 표2는 각 분류자를 위해 추출된 마진벡터의 수를 보여준다.

	마진 벡터의 수(양)	마진 벡터의 수(음)	전체 학습 데이터의 수
SVM_1	166(190)	150(185)	380
SVM_2	148(190)	107(185)	380
SVM_3	156(190)	144(185)	380
SVM_4	160(190)	162(185)	380
SVM_5	123(190)	155(185)	380

표2. 각 분류자들을 위한 마진 벡터

분류 정확률은 오분류(Misclassification)된 경우의 수를 이용한 식(14)의 분류 에러(Classification Error)식을 이용하여 계산되어진다.

$$\text{분류에러} = \frac{\text{오분류된경우의수}}{\text{모든경우의수}} \qquad (14)$$

학습된 분류자를 이용하여 테스트 데이터에 대한 실험결과는 아래와 같다. S_SVM은 단일 SVM을 이용한 분류 정확률을 ME_SVM은 마진 벡터만를 이용한 앙상블 구조의 SVM의 분류 정확률을 나타낸다.

5. 결론

위의 실험결과, 표2에서 학습을 위해 추출된 마진 벡터의 수는 각 분류자마다 다르다. 이것은 추출된 데이터의 군집의 차이를 나타내는 것으로 각 클래스의 중심에 데이터들이 집중되어 있는 경우 추출된 마진 벡터의 수는 줄어들고, 반대의 경우는 데이터들이 이산되어 있는 경우를 나타낸다. 그림3에서 단일 SVM를 사용한 방법의 에러율은 우량 클래스의 경우 23.48%,불량 클래스는 21.84%를 보였다. 이에 비해서 본 논문에서 제시한 방법의 에러율은 우량 클래스의 경우 22.72%, 불량 클래스는 21.08%로 우량과 불량 클래스에 대하여 분류의 정확률이 각각 0.76%씩 높았다.

향후의 연구계획으로 SVM이 3원 분류이상에서 분류 정확도가 현저히 떨어지는 단점을 극복할 수 있는 분류기법에 대하여 연구할 계획이다.

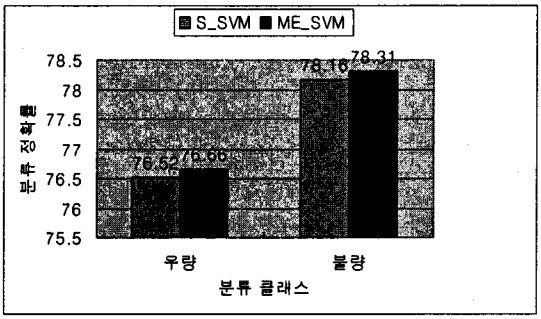

그림3. S_SVM과 ME_SVM를 이용한 분류 정확률

참고문헌

[1] Atiya, Amir, F., "Backkrupty Prediction For Credit Risk Using Neural Networks : a suvey and new results," IEEE transactions on neural networks, Vol.12, No.4, pp.929-935,2001.

[2] Thomas G. Dietterich, "Machine Learning Research: Four Current Directions". The AI Magazine, vol 18, no.4, 97-136,1998.

[3] Vapnik, V., The Nature of Statistical Learning Theory. Springer-Verlag, 1995.