Negative data를 고려한 K-means Support Vector Data Description

K-means Support Vector Data Description concerning Negative data

송동성*, 김표재**, 장형진***, 최진영****
(Dong Sung Song*, Pyo Jae Kim**, Hyung Jin Chang***, Jin Young Choi****)

Abstract - SVDD는 one-class 분류기법이지만, 다중 클래스 분류에도 적용될 수 있다. 이 때 타 클래스의 data가 고려대상 클래스의 학습된 경계안에 들어오지 않도록 하기 위하여 negative data를 고려한 SVDD방법이 사용되어 왔다. 그러나 이 방법은, 고려해야 하는 데이터 수가 늘어남에 따라 학습에 소요되는 시간이 증가하게 되는 문제점을 가지고 있다. 본 논문에서는 negative data를 고려한 학습 시, SVDD대신 KMSVDD를 사용하고 negative data일 가능성이 없는 영역에 놓인 데이터를 제외하는 기법을 사용함으로써 학습시간의 증가를 완화하는 방법을 제안하고자 한다. 이를 통해서 대상 클래스에 속하지 않은 모든 data를 negative data로 고려하여 학습을 진행할 때 보다 빠른 시간에 유사한 결과를 얻을 수 있다. 몇 가지 모의실험을 통하여 그 효과를 검증하도록 한다.

Key Words : SVDD, KMSVDD, Negative data, Clustering

1. 서 론

SVDD(support vector data description)[1]는 one-class 분류 문제를 푸는 대표적인 방법 중 하나이다. 이는 클래스 경계를 묘사하는 작업을 통해 주어진 data가 클래스 속한 것인지 클래스에 속하지 않은 것인지를 구분한다. 클래스의 경계 묘사에 있어서는 SVM(support vector machine)[2]과 유사하게 SV(support vector)를 학습해야 하므로 QP(quadratic programming)문제를 풀어야 한다. QP문제는 학습에 사용되는 데이터의 개수가 증가함에 따라 학습 시간이 길어지는 문제를 가지고 있다. 이는 대용량의 데이터를 처리해야 하는 문자인식이나 얼굴인식과 같은 문제에 SVDD를 적용하기 어렵게 하는 요인이 된다. KMSVDD(K-means support vector data description)는 이러한 대용량 데이터 처리 문제를 다루는 여러 방법 중의 하나로, K-means clustering방식을 이용하여 분할된 몇 개의 보다 작은 크기를 가진 data set에 대해 SVDD를 적용하는 방식이다. 하나의 큰 문제를 풀지 않고, 문제를 쪼개어 여러 개의 작은 문제를 푸는 것과 같은 이 방법은 실제 적용 사례를 살펴 볼 때 기존의 SVDD에 비해 학습시간을 단축 시키는 성능의 개선이 있음을 알 수 있다[3].

SVDD는 one-class 분류 기법 이지만, 다중 클래스 분류에도 응용이 가능하다. 예를 들면, data set에 안에 여러 개의 클래스가 있을 때, 각각 대상 클래스만을 고려하여 학습을 한 뒤 얻어진 data description 경계를 통합하는 작업을 통해 여러 개의 클래스를 지닌 분류 문제에도 활용할 수 있다. 그러나 다른 클래스의 data를 고려하지 않고서 학습을 하게 되면, 학습된 클래스 경계의 내부로 다른 클래스의 data가 들어올 수도 있다. 이러한 data를 negative data 라고 하며, 이 경우 학습된 클래스 경계가 정확한 클래스 분류에 한계를 지니고 있게 된다. SVDD는 이러한 negative data를 고려한 학습 방법을 제시하고 있는 소수의 학습 방법 중 하나이다. 하지만 이 방법은 고려해야 할 data 가 많아짐에 따라 학습시간이 급격히 증가하는 문제를 가지고 있다.

본 논문에서는 KMSVDD의 SVDD학습을 negative data를 고려한 SVDD학습으로 대체하여, 다중클래스 분류 문제에서 negative data를 고려한 SVDD 학습을 수행 할 때 학습시간을 줄이고자 한다. 이는 학습에 포함되는 negative data의 수를 줄이기 위해, 학습 후 클래스 경계 안으로 들어올 가능성이 없는 negative data를 선별하여 제외한다. 이 선별 작업은 K-means clustering 단계와 cluster별 SVDD 학습 단계 사이에서 수행하도록 한다.

본 논문에서는 몇 가지 의미 있는 data set에 대하여 제안된 알고리즘을 포함하는 KMSVDD 학습 방법과 negative data를 고려한 SVDD 학습방법을 적용하여 보고 학습 시간과 성능의 차이를 비교하도록 한다.

저자 소개
* 송동성: 서울대학교 전기 · 컴퓨터공학부 석사과정, ASRI
** 김표재: 서울대학교 전기 · 컴퓨터공학부 박사과정, ASRI
***장형진: 서울대학교 전기 · 컴퓨터공학부 통합과정, ASRI
**** 최진영: 서울대학교 전기 · 컴퓨터공학부 교수, ASRI

2. KMSVDD concerning Negative data

2.1 SVDD concerning negative data

Negative data를 고려한 data description은 negative data를 hypersphere 바깥쪽에, target data는 안쪽에 위치하도록

한다는 점에서 일반적인 Support Vector Classifier와 다르다. Support Vector Classifier는 여러 개의 클래스들을 구분할 수 있지만 어느 클래스에도 속하지 않는 바깥점(outlier)들은 잡아내지 못한다.

먼저 Negative data를 고려한 SVDD는 target에 대하여 $y_i = 1$, outlier에 대하여 $y_l = -1$ 라고 정의한다.

SVDD에서와 유사하게 최소화해야 할 목적함수를 정의하면 다음과 같다. (이때 i 는 target data, l 은 negative data)

$$F(R, a, \xi_i, \xi_l) = R^2 + C_1 \sum_i \xi_i + C_2 \sum_l \xi_l \quad (1)$$

제약조건 : $\| x_i - a \|^2 \le R^2 + \xi_i$, $\| x_l - a \|^2 \ge R^2 - \xi_l$

$\xi_i \ge 0$, $\xi_l \ge 0$ $\forall i, l$

와 같이 된다. 위의 식에서 모수 C_1, C_2 는 구면체의 부피와 오차사이의 절충 값을 조절한다.

목적함수와 제약조건을 라그랑지 승수법을 써서 표현하면,

$$L(R, a, \alpha_i, \alpha_l, \gamma_i, \gamma_l, \xi_i, \xi_l) = R^2 + C_1 \sum_i \xi_i + C_2 \sum_l \xi_l - \sum_i \gamma_i \xi_i - \sum_l \gamma_l \xi_l \quad (2)$$
$$- \sum_i \alpha_i \{ R^2 + \xi_i - (x_i - a)^2 \} - \sum_l \alpha_l \{ (x_l - a)^2 - R^2 + \xi_l \}$$

와 같다.

여기서 라그랑지 승수들은 $\alpha_i \ge 0$, $\alpha_l \ge 0$, $\gamma_i \ge 0$, $\gamma_l \ge 0$가 되며, L은 R, a, ξ_i, ξ_l에 관해서는 최소화 되어야 하고, $\alpha_i, \alpha_l, \gamma_i, \gamma_l$에 관해서는 최대화되어야 한다. L을 R, a, ξ_i, ξ_l 각각에 대해서 편미분한 결과를 0으로 놓은 후, 이를 QP을 이용하여 학습한다.

$$L = \sum_i \alpha_i (x_i \cdot x_i) - \sum_{i,l} \alpha_i \alpha_l (x_l \cdot x_i) - \sum_{i,l} \alpha_i \alpha_l (x_i \cdot x_l) \quad (3)$$
$$+ 2 \sum_{i,j} \alpha_i \alpha_j (x_i \cdot x_j) - \sum_{l,m} \alpha_l \alpha_m (x_l \cdot x_m)$$

이 때, $\alpha_i' = y_i \alpha_i$ 라고 하면, 식 (3)은 negative data를 고려하지 않은 normal SVDD에서의 식와 유사해 진다.

$$L = \sum_i \alpha_i' (x_i \cdot x_i) - \sum_{i,j} \alpha_i' \alpha_j' (x_i \cdot x_j) \quad (4)$$

Negative data를 고려하여 새롭게 진행된 학습에서는, 기존 학습에서 data description 경계 내·외부에 있던 negative data들이 대부분 새롭게 형성된 data description 경계위에 마치 SV처럼 위치하게 된다[1]. 이 때 Data description 경계의 모양은 negative data에 관한 모수(C_2)를 어떻게 설정하느냐에 따라 절대적으로 negative data를 배제하기도 하고 어느 수준까지 포함하기도 하는 식으로 형성된다.

또한 negative data를 고려한 학습 시, data description 경계가 negative data를 배제하지 못하거나 target data만을 고려했을 경우보다 광범위하게 형성될 경우, inner product $(x_i \cdot x_j)$에 커널함수를 적용하여 data를 밀접하게 둘러싸는 경계를 형성하도록 조정한다. 이는 커널함수의 width를 적절히 결정하는 것을 통해 가능하다.

2.2 KMSVDD concerning negative data

Negative data를 고려한 SVDD 학습시, 학습에 필요한 negative data들에 대한 정의가 없다면 자신의 클래스를 제외한 모든 data를 negative data로 고려해야 한다. 이는 SVDD 학습의 특성상 학습 data set의 크기가 커짐에 따라 급격히 학습시간이 증가하는 문제를 가지고 있다.

이에 기존의 KMSVDD 알고리즘을 바탕으로 하여 학습해야 하는 negative data의 수를 한정할 수 있는 SVDD 학습 알고리즘을 제안하고자 한다. 본 논문에서 제안된 알고리즘은 다음과 같이 3 단계로 요약할 수 있다.

1 단계: K-means 알고리즘을 이용한 데이터 영역 나눔

학습하고자 하는 데이터 영역에 k개의 중심을 가정한다. K-means 알고리즘을 이용하여 학습 영역을 k개의 sub-group들로 분할한다.

2 단계: 학습에 사용될 negative data 선별

▶ Step 1 : negative data를 포함하는 sub-group 결정

각 클래스 별로 분할된 k개의 sub-group들에 대하여 각각자신의 group안에 속한 data의 중심점(평균점, 이하 M_{it})과 중심으로부터 가장 멀리 떨어진 data와의 거리(이하 MAX_{dist})를 구한다. 예를 들어 그림 1에서 클래스가 2개(A, B라 하자)이고, 각 클래스마다 sub-group이 3개(1, 2, 3)라고 가정하면 총 6개의 sub-group은 각각 자신의 M_{it}, MAX_{dist}를 가진다.

이제 A class의 1(이하 A-1) 에 대하여 B class의 sub-group 1, 2, 3 중 negative data로 고려하지 않아도 될 만큼 멀리 떨어져 분포된 것을 찾아낸다. 예를 들어 A-1 와 B-1의 중심점간의 거리를 구하고 이 거리가 A-1의 MAX_{dist}와 B-1의 MAX_{dist}의 합 보다 크면 두 sub-group은 서로의 data 분포에 대한 영향을 주지 않는 관계로 판단한다. 이러한 두 sub-group 사이에서는 한 sub-group의 SVDD 학습 시 다른 쪽의 data에 대하여 고려하지 않아도 된다.

▶ Step 2 : sub-group 내부에서의 negative data 선별

Step 1에서 결정된 sub-group data 중 일부만을 선택하여 negative data를 고려한 학습에 포함하도록 하자. 형성될 data description 경계는 중심으로부터 가장 멀리 떨어져 있는 data사이의 거리(MAX_{dist}) 보다 더 광범위 하게 형성될 수 없다. 따라서 Step 1에서 결정된 sub-group data들 중 data 중심으로부터 MAX_{dist}사이에 포함되는 negative data만을 골라내어 학습에 사용할 negative data로 결정한다.

3 단계: SVDD concerning negative data를 이용한 데이터 영역 묘사

각 클래스 별로 분할된 k개의 sub-group들에 대하여 2단계에서 한정한 negative data 들을 포함한 SVDD 학습을 수행한다. 학습 후에 각 클래스 별로 형성된 k개의 묘사 영역을 합쳐 해당 클래스의 데이터 영역으로 결정한다.

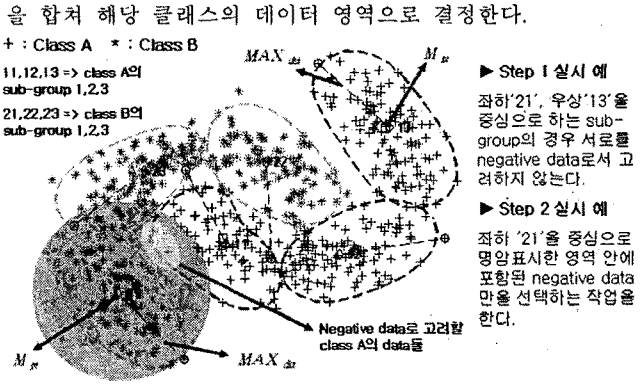

그림 1. negative data 선별 알고리즘

3. 결 과

본 논문에서 제안하는 선별된 negative data를 고려한 KMSVDD와 negative data를 고려한 SVDD알고리즘과의 학습 시간 및 성능 비교를 위해 다음 모의실험을 실시하였다.

학습에 사용된 데이터는 바나나 모양의 분포 형태를 가지며 두개의 클래스를 가지는 학습 데이터를 사용하였다. 학습 대상 데이터의 수를 변화시키면서 학습에 걸리는 시간과 학습된 데이터 경계 모양을 비교하여 보았다. 학습 시간을 조사할 때 사용한 QP알고리즘은 matlab 7.1의 quadprog를 두 알고리즘에 동일하게 사용하였으며, Intel Pentium Mobile processor 1.73GHz, Ram 512M의 컴퓨터에서 모의실험을 실시하였다. K-means 알고리즘에 이용되는 k개의 중심의 수는 각 클래스 당 3개로 설정하였고, 모수값에 해당하는 rejection ratio는 target data의 경우 0.05, negative data의 경우 0.025로 설정하였다. Non-separable 한 문제 해결을 위하여 가우시안 커널을 사용하였으며, 커널 width는 5로 설정 하였다.

그림 2는 negative data를 고려하지 않고 진행한 KMSVDD와 본 논문에서 제안된 알고리즘을 사용하여 negative data를 처리한 KMSVDD 두 경우에 대하여 형성된 data description 영역을 나타낸다. Negative data를 고려하지 않은 경우에 비해 negative data를 고려한 경우의 data description 경계는, negative data를 배제하기 위하여 좀 더 밀접하게 대상 클래스영역을 둘러싸고 있음을 알 수 있다.

그림 3은 상대 클래스 data전체를 negative data로 고려한 SVDD와 본 논문에서 제안된 알고리즘을 사용하여 negative data를 고려한 KMSVDD에 대하여 형성된 data descripition 영역를 나타낸다. 두 경우에, 형성된 data description 경계는 전체적으로 유사하나 약간의 차이를 보인다. 제안된 알고리즘을 이용한 경우, sub-group별로 negative data를 철저히 배제하고 있기 때문에 이와 같은 차이가 발생하였다.

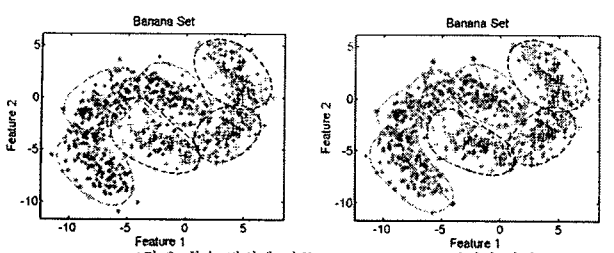

그림 2. 학습 방식에 따른 data description 경계의 차이

(좌) KMSVDD (우) KMSVDD with negative data

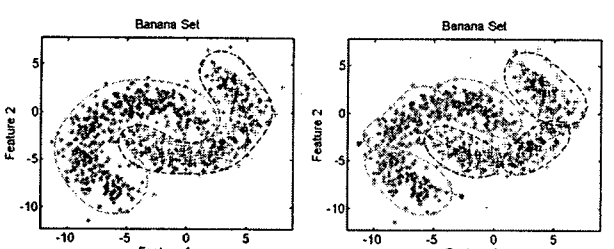

그림 3. 학습 방식에 따른 data description 경계의 차이

(좌) SVDD with negative data (우) KMSVDD with negative data

표 1과 그림 4는 data 수에 따른 학습시간의 차이를 비교한 실험의 결과이다. 바나나 모양 분포의 data 수를 늘려가면서, 본 논문에서 제안된 알고리즘을 사용한 경우와 기존 방식들(SVDD, KMSVDD)간의 학습시간을 비교하였다.

	Banana shape data set		
	KMSVDD with selectied negative data	KMSVDD with negative data	SVDD with negative data
Data (#)	Learning time(s)	Learning time(s)	Learning time(s)
100	3.36	24.99	8.82
200	10.17	662.34	258.30
300	67.96	2870.18	1125.01
400	80.45	18627.11	3200.18

표 1. 데이터 수에 따른 학습 시간 (k-3)

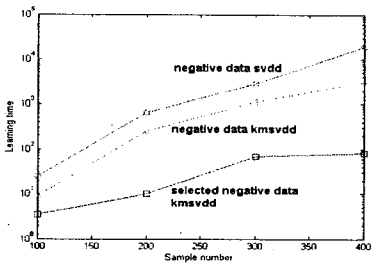

그림 4. 바나나 모양 데이터의 학습시간 (학습시간을 로그스케일로 그린 경우)

Negative data를 사용하는 SVDD학습의 경우 데이터 수가 늘어남에 따라 학습시간이 기하 급수적으로 증가하는 것을 확인 할 수 있다. KMSVDD를 이용하였지만 대상 클래스의 data를 제외한 나머지를 data를 모두 negative data로 고려한 경우에도 데이터 수의 증가에 따라 학습시간이 기하 급수적으로 늘어났다. 그에 비해 제안된 알고리즘을 사용한 경우 데이터 수의 증가에 따른 학습시간의 증가가 크지 않았다..

4. 결 론

본 논문에서는, 다중 클래스 분류 문제에 적용된 SVDD에서 클래스 간의 묘사 경계가 중첩되는 문제를 풀기 위하여 negative data를 포함한 학습을 할 때, 학습시간의 급격한 증가가 발생하는 문제를 개선하기 위한 연구를 진행하였다. 데이터 개수가 증가할 때 SVDD의 학습시간이 기하급수적으로 증가하는 문제를 해결하기 위하여 제안되었던 KMSVDD에 기반하여, negative data를 고려한 학습에 사용될 data를 선별하는 알고리즘을 제안하였다. 모의 실험의 결과를 통해 제안된 알고리즘이 기존의 negative data를 고려한 SVDD 학습과 비교하여, 형성된 data description 경계는 유사하지만 학습시간을 줄이는 것을 확인 할 수 있다.

추후 진행 과제로는 negative data를 고려한 다중 클래스 분류 문제에 SVDD 또는 KMSVDD를 적용 할 때, 주어진 target data를 학습 하는데 필요한 모수 C_1과 negative data를 학습 하는데 필요한 모수 C_2의 적절한 설정에 대한 연구가 있다.

참 고 문 헌

[1] David M.J. Tax, "Support Vector Data Description", Machine Learning, vol, 54, pp. 45-66, 2004.

[2] Vapnik, V. *Statistical Learning Theory*, Wiley New York, 1998.

[3] 김표재, 장형진, 송동성, 최진영 "KMSVDD: K-Means Clustering을 이용한 Support Vector Data Description" 정보 및 제어 심포지움, 2006