KMSVDD: K-means Clustering을 이용한 Support Vector Data Description

KMSVDD: Support Vector Data Description using K-means Clustering

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

Abstract - 기존의 Support Vector Data Description (SVDD) 방법은 학습 데이터의 개수가 증가함에 따라 학습 시간이 지수 함수적으로 증가하므로, 대량의 데이터를 학습하는 데에는 한계가 있었다. 본 논문에서는 학습 속도를 빠르게 하기 위해 K-means clustering 알고리즘을 이용하는 SVDD 알고리즘을 제안하고자 한다. 제안된 알고리즘은 기존의 decomposition 방법과 유사하게 K-means clustering 알고리즘을 이용하여 학습 데이터 영역을 sub-grouping한 후 각각의 sub-group들을 개별적으로 학습함으로써 계산량 감소 효과를 얻는다. 이러한 sub-grouping 과정은 hypersphere를 이용하여 학습 데이터를 둘러싸는 SVDD의 학습 특성을 훼손시키지 않으면서 중심점으로 모여진 작은 영역의 학습 데이터를 학습하도록 함으로써, 기존의 SVDD와 비교하여 학습 정확도의 차이 없이 빠른 학습을 가능하게 한다. 다양한 데이터들을 이용한 모의실험을 통하여 그 효과를 검증하도록 한다.

Key Words : KMSVDD, SVDD(support vector data description), K-means clustering, decomposition method

1. 서 론

기계의 고장 진단이나 이미지 검색, 신분 증명과 같은 문제들은 기존의 패턴 분류나 회귀 문제와는 다른 접근 방법을 필요로 한다. 이러한 문제들은 one-class 분류 문제로 불리며 클래스의 경계를 묘사하는 작업을 통하여 클래스에 속한 개체들과 속하지 않은 개체들 사이의 구분을 가능하도록 한다. SVDD(support vector data description)[1]는 대표적인 one-class 분류 기법 중의 하나이며 클래스에 속하는 개체들을 특징(feature) 공간상에서 hypersphere에 속하도록 학습하여 데이터 공간에서 클래스 경계를 찾아낸다. 클래스 경계를 묘사하기 위해서는 SVM(support vector machine)[2] 에서의 경우와 동일하게 SV(support vector)를 학습해야 하므로, QP(quadratic programming) 문제를 해결하여야 하며 데이터 영역이 non-separable 문제를 해결하기 위하여 커널을 이용하여 학습한다.

일반적으로 QP 문제는 학습에 사용되는 데이터의 개수가 증가함에 따라 학습 시간이 길어지는 단점을 가지고 있다. 이를 해결하기 위하여 기존의 SVM에서는 SMO (sequential minimal optimization)[3] 와 같은 빠른 QP 학습 알고리즘을 이용하거나, chunking 알고리즘[3]이나 decomposition 방법[4][5]처럼 학습 데이터 영역을 나누어 각각의 subproblem을 학습하여 학습 시간을 단축하는 방법들이 이용되었다. 이러한 방법들은 SVM과 동일한 최적화 기법을 사용하는 SVDD에 적용될 수 있으며, 학습시간의 개선을 가능하게 해 준다.

본 논문에서는 K-means clustering 알고리즘을 이용하여 기존의 decomposition 방법과 유사하게 학습 데이터 영역을 다수의 sub-group으로 나누어 학습하여 학습 속도를 단축시키는 방법을 제안하고자 한다. SVDD는 특징공간 상에서 같은 클래스에 속하는 모든 데이터들을 가능한 하나의 hypersphere에 포함되도록 학습을 하며 이는 데이터 공간에서도 닫힌 공간의 클래스 영역으로 나타내게 된다. K-means 알고리즘을 이용하여 학습 데이터 영역을 나누게 되면, k개의 중심점들로 좀 더 군집되어 여러 개의 sub-group들이 생기게 된다. K-means 알고리즘의 이러한 특징은 가능한 모든 학습 데이터들을 하나의 hypersphere에 포함하도록 학습하려는 SVDD의 특성과 유사하다. 따라서 SVDD 학습 전에 K-means 알고리즘을 이용하여 학습 데이터 영역을 sub-grouping 하는 것은 학습 데이터 간의 유사성은 높으며 개체 수는 적은 여러 개의 subproblem으로 문제를 나누어 해결함으로써 학습 성능의 저하 없이 계산 시간이 단축되는 효과를 가져 온다.

본 논문에서는 제안한 알고리즘을 학습 데이터의 개수가 다른 여러 데이터들에 적용하여보고 기존의 SVDD와의 비교를 통하여 학습에 걸리는 시간과 정확도의 차이를 조사하도록 한다.

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

2. KMSVDD

2.1. SVDD

SVDD는 hyperplane을 가정하여 패턴을 분류할 수 있는

판단 경계면(decision boundary)을 학습하는 Schölkopf [1]의 방법과 유사하지만, 특징공간에서의 클래스의 경계를 hyperplane이 아닌, a를 중심으로 하고, R을 반지름으로 하여 가능한 모든 학습 영역을 포함하는 hypersphere를 가정하여 학습한다는 차이점을 가지고 있다.

SVM과 유사하게 hypersphere에 의해 결정되는 목적 함수를 정의하면 다음과 같다.

$$F(R,a)=R^2 \tag{1}$$

제약조건: $\|x_i-a\|^2 \le R^2, \quad \forall i$

여기서 hypersphere의 외곽 경계면 근처에 있는 학습 데이터에 대한 학습을 고려하기 위해, slack 변수 $\xi_i \ge 0$을 도입하여 위의 식을 다시 표현하면,

$$F(R,a)=R^2+C\sum_i \xi_i \tag{2}$$

제약조건: $\|x_i-a\|^2 \le R^2+\xi_i, \quad \xi_i \ge 0 \quad \forall i$

와 같이 된다. 위의 식에서 모수 C 는 구면체의 부피와 오차 사이의 절충 값을 조절한다.

목적함수와 제약조건을 라그랑지 승수법을 사용하여 표현하면,

$$\begin{aligned} L(R,a,\alpha_i,\gamma_i,\xi_i) &= R^2+C\sum_i \xi_i \\ &-\sum_i \alpha_i\{R^2+\xi_i-(\|x_i\|^2-2a\cdot x_i+\|a\|^2)\} \\ &-\sum_i \gamma_i\xi_i \end{aligned} \tag{3}$$

와 같다.

여기서 라그랑지 승수들은 $\alpha_i \ge 0, \gamma_i \ge 0$가 되며, L은 R, a, ξ_i에 관해서는 최소화 되어야하고, α_i, γ_i에 관해서는 최대화되어야 한다. L을 R, a, ξ_i 각각에 대해서 편미분한 결과를 0으로 놓은 후, 이를 QP을 이용하여 학습한다. 라그랑지 승수 α_i 가 학습 데이터 x_i에 대해 $0 < \alpha_i < C$를 만족 하는 데이터 x_i들을 support vector (SV)라 하고, $\alpha_i = C$ 인 데이터 x_i들을 바깥점(outlier)이라고 한다.

새로운 데이터 z에 대한 클래스를 판단하기 위해서는 sphere의 중심으로부터의 거리를 계산해야 한다.

$$\begin{aligned} \|z-a\|^2 &= K(z,z)-2\sum_i \alpha_i K(z,x_i) \\ &+\sum_{i,j} \alpha_i\alpha_j K(x_i,x_j) \le R^2 \end{aligned} \tag{4}$$

여기서 K는 non-separable 문제를 해결하기 위하여 도입된 커널 함수를 나타낸다. 계산 결과 거리가 반지름 R보다 작거나 같으면 데이터 z는 같은 클래스에 속하는 것으로 판단된다.

2.2. KMSVDD

본 논문에서 제안된 K-means clustering을 이용하는 SVDD 알고리즘은 그림 1과 같으며 다음과 같이 2 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: SVDD를 이용한 데이터 영역 묘사

각 클래스 별로 분할된 k개의 sub-group들에 대하여 SVDD를 이용하여 개별적인 학습을 수행한다. 학습 후에 각 클래스 별로 형성된 k개의 묘사 영역을 합쳐 해당 클래스의 데이터 영역으로 결정한다.

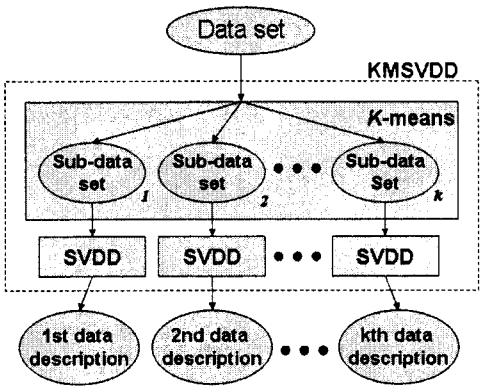

그림 1. KMSVDD 알고리즘

3. 결 과

본 논문에서 제안된 KMSVDD와 기존의 SVDD 알고리즘의 학습 시간 및 정확도 차이를 비교하기 위하여 다음과 같은 모의실험을 실시하였다.

	Banana shape data set				Donut shape data set			
	SVDD		KMSVDD		SVDD		KMSVDD	
Data #	Learning time (s)	SV #	Learning time (s)	SV #	Learning time (s)	SV #	Learning time (s)	SV #
100	1.79	30	0.49	34	1.83	33	0.54	47
200	12.34	36	0.79	37	12.02	38	0.87	60
300	46.33	42	1.19	41	44.80	40	1.40	62
400	123	55	1.90	45	116	42	2.06	67
500	265	56	2.96	47	253	43	3.15	67
600	529	64	4.02	47	493	43	4.18	70
700	918	74	5.55	48	882	44	5.73	73
800	1676	90	7.69	48	1473	45	7.98	72
900	2379	97	9.96	49	2237	45	10.34	74
1000	3493	101	13.39	51	3322	47	13.82	75

표 1. 각 데이터 별 학습 시간 및 선택되는 support vector의 개수 (k=3)

학습에 사용된 데이터는 바나나와 도넛 모양의 분포 형태를 가지며 각각 두개의 클래스를 가지는 학습 데이터를 사용하였다. 학습하여야 할 데이터의 개수를 달리 하면서 학습에 걸리는 시간과 학습된 데이터 경계 모양을 비교하여 보았다. 클래스의 경계를 결정하는 SV를 구하는 QP 알고리즘은 matlab 6.5의 quadprog를 두 알고리즘에 모두 동일하게 적용하였으며, P4 3.0GHz Ram 1G의 컴퓨터에서 모의실험을 실시하였다. K-means 알고리즘에 필요한 k개의 중심의 수는 각 클래스 별로 3개로 하였고, C는 0으로 설정하였다. Non-separable 한 문제 해결을 위하여 가우시안 커널을 사용하였으며, 커널에 사용한 width 모수는 2(바나나 모양), 3(도

넛 모양)으로 설정 하였다.

표 1과 그림 2, 4 은 두 가지 데이터 집합에 대한 학습시간을 나타낸다. 표와 그림에서 확인 할 수 있듯이 기존의 SVDD 알고리즘은 학습 데이터의 개수가 증가함에 따라 학습에 필요한 시간이 기하급수적으로 증가함에 반하여 제안된 알고리즘은 학습할 데이터 영역을 나누어 학습함으로써 학습시간을 단축시키는 효과를 가지고 있음을 확인할 수 있었다.

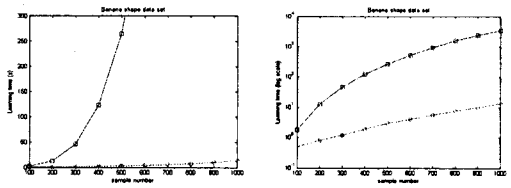

그림 2. 바나나 모양의 데이터 영역에 걸리는 학습 시간
(좌) 학습시간 (단위 초) (우) 학습시간의 로그스케일로 그린 경우

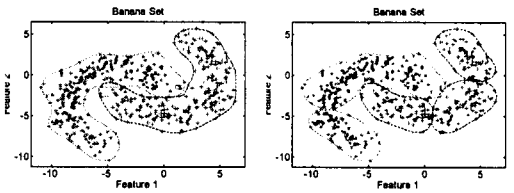

그림 3. 바나나 모양에 대한 데이터 영역 표시 (좌) SVDD (우) KMSVDD

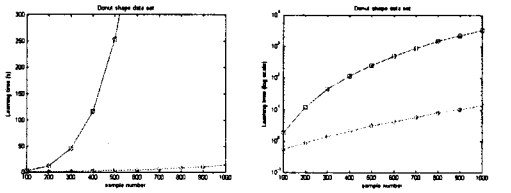

그림 4. 도넛 모양의 데이터 영역에 걸리는 학습 시간
(좌) 학습시간 (단위 초) (우) 학습시간의 로그스케일로 그린 경우

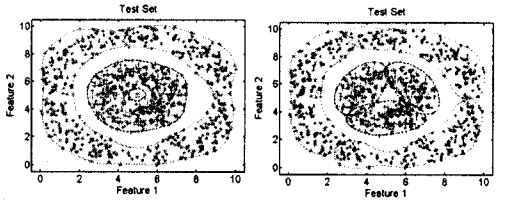

그림 5. 도넛 모양에 대한 데이터 영역 표시 (좌) SVDD (우) KMSVDD

그림 3과 5에서 볼 수 있듯이 학습 후 얻는 데이터 영역은 두 알고리즘 모두 비슷한 결과를 보여 주었다. KMSVDD의 경우 k-means 알고리즘에 의해 각 클래스 별로 k개의 데이터 영역 묘사들이 나타나며 SVDD는 하나의 데이터 영역 묘사만을 얻게 된다. 학습에 의해 결정되는 SV는 두 알고리즘 모두 비슷한 위치의 데이터를 선택하나, KMSVDD는 k개의 sub-group 간의 경계에 위치한 데이터들 중에서 SV가 추가적으로 선택됨으로써, 선택되는 SV의 개수가 증가한다. 그러나 이는 표 1에서 알 수 있듯이 k값에 따라 단순 증가하는 것은 아니며, 학습 데이터의 분포와 학습을 위해 설정되는 모수 값에 따라 영향을 받는다. 그림 3(좌), 그림 5(우)에서 나타나는 클래스 간 묘사 영역의 중첩은 다른 클래스에 대한 학습데이터를 학습에 고려하는 SVDD 알고리즘(SVDD with negative examples)[1]을 적용하면 개선 될 수 있다.

k값을 증가시키면 학습 시간을 더욱 단축 할 수 있으나 데이터 영역 묘사의 개수가 늘어나 새로운 데이터 입력에 대하여 클래스를 판단하는 시간이 오래 걸리게 되며 선택되는 SV 개수도 일반적으로 증가한다. 따라서 KMSVDD에서는 k의 선택이 학습에 중요한 인자가 되며 이의 선택은 데이터의 분포에 따라 달라져야 한다.

4. 결 론

본 논문에서는 one-class classifier의 대표적인 알고리즘인 SVDD에서 데이터의 개수가 증가함에 따라 QP 문제를 푸는 시간이 기하급수적으로 증가하는 문제를 해결하기 위하여, K-means clustering 알고리즘을 사용하여 좀 더 밀집된 여러 개의 소집단으로 데이터 영역을 나눈 후 각 부분 소집단 별로 QP문제를 해결하여 계산 시간을 단축 할 수 있는 알고리즘을 제안하였다. 모의실험 결과는 데이터 영역의 분해로 인한 학습 시간의 감소 효과를 명확하게 보여 주었다. 또한 두 알고리즘에 의해 얻은 데이터 영역을 비교하였을 경우 유사한 형태의 데이터 묘사 영역을 얻을 수 있었다.

향후 과제에서는 KMSVDD의 사용자가 선정하는 모수 값과 학습 시간 및 정확도에 대한 관계를 이론적으로 분석해 보고자 한다. KMSVDD를 사용하면 학습 영역을 좀 더 군집된 형태의 소집단으로 나눌 수 있으므로 학습에 사용되는 커널의 width 모수를 크게 선정 하더라도 가능한 모든 데이터 영역을 포함할 수 있을 것으로 생각된다. 또한 앞서 언급한 k값에 따른 학습 시간과 데이터 영역 묘사 식 변화에 대한 이론적 연구를 진행할 계획이다.

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] J.C.Platt, "Fast training of support vector machines using sequential minimal optimization", *Advances in Kernel Methods-Support Vector Learning*, MIT Press, 1998.

[4] T.Joachims, "Making large-scale SVM learning practical" *Advances in Kernel Methods-Support Vector Learning*, MIT Press, 1998.

[5] E.Osuna, R.Freund, and F.Girosi, "Training support vector machines", Conference on Computer Vision and Pattern Recognition, pp. 130-136 ,1997.

[6] Richard O. Duda, Peter E. Hart, David G. Stork, Pattern Classification, Wiley-Interscience, 2001.