평균 이동 알고리즘 기반의 지지 벡터 영역 표현 방법

Support Vector Data Description using Mean Shift Clustering

장형진*, 김표재**, 최정환***, 최진영****
(Hyung Jin Chang*, Pyo Jae Kim**, Jung Hwan Choi***, Jin Young Choi****)

Abstract - SVDD의 scale problem을 해결하기 위하여, 학습 데이터를 sub-grouping하여 group 단위로 SVDD를 통해 학습함으로써 학습 시간을 줄이는, K-means clustering을 이용한 SVDD 방법(KMSVDD)이 제안되었다. 하지만 KMSVDD는 K-means clustering 알고리즘의 본질상 최적의 K값을 정하기 힘들다는 문제와, 동일한 데이터를 학습할지라도 clustered group이 랜덤하게 형성되기 때문에 매번 학습의 결과가 달라지는 문제점이 있었다. 또한 데이터의 분포 상태와 관계없이 무조건 타원(elliptic) 형태의 K개의 cluster로 나누기 때문에 각각의 나눠진 cluster들은 데이터 분포에 대한 특징을 나타내기 힘들게 된다. 이러한 문제점을 해결하기 위하여 본 논문에서는 데이터 분포에서 mode를 먼저 찾은 후 이 mode를 기준으로 clustering하는 Mean Shift clustering 방법을 이용한 SVDD를 제안하고자 한다. 제안된 알고리즘은 KMSVDD와 비교해 데이터 학습 속도에서는 큰 차이가 없으면서도 데이터의 분포 상태를 고려한 형태로 clustering한 sub-group을 학습하므로 학습의 정확도가 일정하게 되며, 각각의 cluster는 데이터 분포의 특징을 포함하는 효과가 있다. 또한 Mean Shift Kernel의 bandwidth의 결정은 K-Means의 K와는 달리 어느 정도 여유를 갖고 결정되어도 학습 결과에는 차이가 없다. 다양한 데이터들을 이용한 모의실험을 통하여 위의 내용들을 검증하도록 한다.

Key Words : SVDD(support vector data description); KMSVDD; Mean Shift Clustering; K-means clustering

1. 서 론

특정 질병의 감염 여부나 공정상의 이상 원인을 판단할 때에는 여러 가지 측정 데이터들의 조합을 통하여 하나의 특정 집단(class)을 형성하게 되고 이 집단에 속하는지 여부를 판단하게 된다. 이러한 문제를 one-class 분류 문제라고 한다. 이러한 문제의 해결을 위해서는 다른 클래스와의 상호 관계에 의한 분류 경계를 학습하는 형태의 신경망이나 SVM이 아닌, 자신이 속한 클래스의 외곽 경계를 나타내는 형태의 SVDD(support vector data description)[1]가 적합하다. SVDD는 클래스에 속하는 개체들을 특징(feature) 공간상에서 분류해 내는 방법인 SVM(support vector machine)과 유사하게 개체들을 특징 공간상의 hypersphere에 속하도록 SV(support vector)를 학습하여 전체 데이터를 둘러싸는 형태로 클래스 경계를 표현하는 방법이다.

SV를 학습하는 과정에서 QP(quadratic programming) 문제를 해결해야 하는데, 일반적으로 QP 문제는 학습 데이터의 개수가 증가할수록 학습 시간이 지수 함수적으로 증가하는 문제점이 있다. 이러한 문제점을 해결하기 위해 KMSVDD(K-means support vector data description)[2]가 제안되었다. 이는 학습 데이터 영역을 나누어 각각의 subproblem을 학습하는 decomposition과 유사한 기법으로, K-means clustering 알고리즘을 이용하여 학습 데이터 영역을 K개의 sub-group으로 나누어 학습함으로써 학습 시간을 크게 단축시켰다. 하지만 KMSVDD는 K-means clustering의 본질적인 문제인 K값을 정하기 힘들다는 점과 동일한 데이터일지라도 매번 clustering의 결과가 조금씩 달라진 않다는 점, 또한 cluster는 데이터의 분포 상태와 무관하게 K개의 타원 형태로 형성된다는 문제점이 있다.

본 논문에서는 K-means clustering 대신 Mean Shift 알고리즘[3]을 이용하여 clustering[4][5]을 수행하는 방법을 제안하고자 한다. Mean Shift 알고리즘은 확률 밀도 분포에서 mode를 찾는 방법으로 찾은 mode를 기준으로 clustering을 하므로 cluster의 개수는 mode의 개수와 같게 된다. 물론 Mean Shift 역시 K-means의 K와 같이 사용자가 정해 주어야 하는 계수인 kernel bandwidth가 있지만, K와는 달리 clustering 되는 결과에 영향이 적다. 또한 KMSVDD의 학습 시간과 비교하여 큰 증가 없이 유사한 학습 성능으로, 시행 횟수에 관계없이 clustering 되는 모양이 균일할 뿐만 아니라 데이터의 분포 상태를 최대한 표현하는 형태로 나타나게 된다.

본 논문에서는 제안한 알고리즘을 계산 속도와 학습 결과를 다양한 형태의 학습 데이터를 이용하여 KMSVDD의 성능과 비교하며, 실제 데이터 분포를 보존하며 clustering되는 정도를 parzen window를 통해 얻어낸 data 분포와 비교하여 조사하도록 한다. 아울러 다양한 크기의 Mean Shift kernel bandwidth에 대한 학습 결과를 비교한다.

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

2. MSC-SVDD

2.1. SVDD

SVDD는 특징 공간상의 a를 중심으로 하고, R을 반지름으로 하여 가능한 모든 학습 영역을 포함하는 hypersphere를 가정하여 학습한다. Hypersphere에 의해 결정되는 목적 함수를 slack 변수 $\xi_i \geq 0$을 도입하여 정의하면 다음과 같다.

$$F(R,a) = R^2 + C\sum_i \xi_i \quad (1)$$

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

위의 식에서 모수 C 는 구면체의 부피와 오차사이의 절충값을 조절한다. 목적함수와 제약조건을 라그랑지 승수법을 사용하여 표현하면다음과 같다.

$$L(R,a,\alpha_i,\gamma_i,\xi_i) = R^2 + C\sum_i \xi_i \quad (2)$$
$$-\sum_i \alpha_i \{R^2 + \xi_i - (\| x_i \|^2 - 2a \cdot x_i + \| a \|^2)\} - \sum_i \gamma_i \xi_i$$

여기서 라그랑지 승수들은 $\alpha_i \geq 0, \gamma_i \geq 0$가 되며, L은 R, a, ξ_i에 관해서는 최소화되어야하고, 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.

$$\| z - a \|^2 = K(z,z) - 2\sum_i \alpha_i K(z,x_i) \quad (3)$$
$$+ \sum_{i,j} \alpha_i \alpha_j K(x_i,x_j) \leq R^2$$

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. Mean Shift Clustering

2.2.1. Mean Shift Algorithm

Mean Shift Algorithm은 density gradient를 추정을 위한 non-parametric 기법으로 Fukunaga와 Hostetler[5]에 의해 제안되었다. n-dimension의 공간 X에서 중심이 x인 특정 kernel 안의 집합을 S라 할 때 x∈X의 sample mean m(x)는 다음과 같다.

$$m(x) = \frac{\sum_{s \in S} K(s-x)s}{\sum_{s \in S} K(s-x)} \quad (4)$$

이러한 m(x)과, window의 중심 x의 차를 Mean Shift vector라고 하며 이러한 window의 중심 x가 sample mean으로 반복적으로 이동하는 것이 Mean Shift 알고리즘이다. 즉 Mean Shift는 interest region인 특정영역(window)의 중심이 center of mass로 이동히는 vector를 가지면서 window를 이동시켜나간다.

이러한 알고리즘을 확장해서 살펴보면, 반지름이 r인 구 S_x와 특성벡터 y를 표현하는 mean x로 표현되며 다음과 같다.

$$\mu(x) = \frac{\int_{y \in S_r} P(y)(y-x)dy}{\int_{y \in S_r} P(y)dy} z \quad (5)$$

여기서 p(*)는 p차원 특성 vector의 확률밀도함수이다. Mean Shift 알고리즘에서의 Mean Shift vector $\mu(x)$는 확률밀도 ∇p(x)의 기울기에 비례하고, p(x)에 반비례한다.

$$\mu(x) = c\frac{\nabla p(x)}{p(x)} \quad (6)$$

Mean Shift vector의 이동은 local maximum density에 수렴 할 때까지 반복해 나가며, vector의 방향은 density gradient가 증가하는 곳을 향한다.

이러한 반복을 통해 수렴된 local maximum density값을 mode라고 하며, mode는 기울기 ∇f(x)=0인 지점을 의미한다.

2.2.2. Mean Shift Clustering

Mean shift를 통해 수렴된 각각의 local maximum density 값인 mode를 중심으로 clustering하는 것이 바로 mean shift clustering이다.

동일한 mode를 중심으로 갖는 sample point들의 집합을 basin이라고 하며, 이러한 basin에 의해서 clustering영역이 결정되어진다. 따라서 clustering 개수는 mode의 개수에 따라 자동적으로 결정되어진다.[5]

2.2.3. MSC-SVDD 알고리즘

본 논문에서 제안된 Mean Shift Clustering을 이용한 SVDD 알고리즘은 KMSVDD 알고리즘과 유사한 형태로 다음과 같이 2단계로 나눌 수 있다.

1 단계: Mean Shift Clustering을 이용한 sub-grouping

학습하고자 하는 데이터 분포에서 Mean Shift 알고리즘을 이용하여 mode를 찾은 후, 이 mode를 중심으로 하는 clustering을 통해 학습 영역을 sub-group들로 분할한다.

2 단계: SVDD를 이용한 데이터 영역 묘사

각각의 sub-group들에 대하여 개별적인 학습을 수행한다.

3. 실험 결과

본 논문에서 제안된 MSC-SVDD와 기존의 KMSVDD 알고리즘의 학습 시간 및 데이터 분포 모양 차이를 비교하기 위하여 모의실험을 실시하였다. 학습에 사용된 데이터는 다양한 분포 형태를 가지며 각각 두개의 클래스를 가지는 학습 데이터를 사용하였다. 클래스의 경계를 결정하는 SV를 구하는 QP 알고리즘은 matlab 6.5의 quadprog를 두 알고리즘에 모두 동일하게 적용하였으며, P4 3.0GHz Ram 1G의 컴퓨터에서 모의실험을 실시하였다. 양쪽 모두 동일한 계수 값들을 갖는 동일한 형태의 SVDD을 사용하였다.

3.1. 학습 시간 단축

표 1과 그림 1은 세 가지 알고리즘을 사용한 학습시간을 나타낸다. Clustering을 사용하지 않는 SVDD 알고리즘은 학습 데이터의 개수가 증가함에 따라 학습에 필요한 시간이 기하급수적으로 증가함에 비하여 KMSVDD와 MSC-SVDD는 학습시간을 크게 단축시킴을 확인할 수 있었다. MSC-SVDD 알고리즘을 썼을 때 학습 시간을 단축시키는 성능은

KMSVDD를 비교했을 때 큰 차이가 없음을 확인할 수 있다.

Data #	Learning Time (sec)		
	KMSVDD	MSC-SVDD	SVDD
300	1.47	2.26	52.99
400	1.82	2.99	133.49
500	3.04	5.11	276
600	3.36	8.52	568.13
700	4.54	8.41	975.3
800	5.20	10.39	1650

표 1. 각 데이터 별 학습 시간

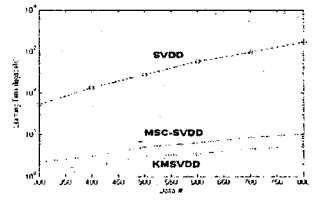

그림 1. 학습 시간의 로그 스케일로 나타냄

3.2. 데이터 분포 특징 표현

K-means Clustering은 K개의 mean을 중심으로 elliptical한 형태로 clustering을 하므로 데이터의 분포 상태와는 무관하게 clustering을 하게 된다. 이와는 달리 Mean Shift clustering은 분포 특징을 고려하여 mode를 중심으로 clustering하게 된다.

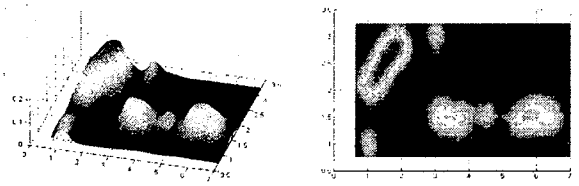

그림 2. Parzen Window를 이용하여 구한 데이터 분포 형태
(좌) 옆에서 본 분포 형태 (우) 위에서 본 분포 형태

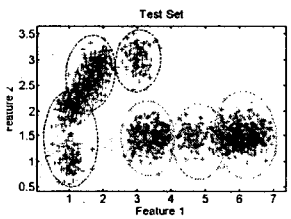

그림 3. KMSVDD 이용 학습

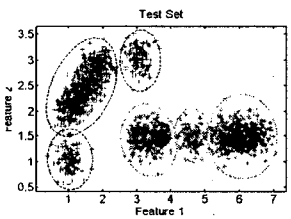

그림 4. MSC-SVDD 이용 학습

그림2는 Parzen Window 방법을 이용하여 실제 데이터 분포 상태를 나타낸 것이다. 그림3는 KMSVDD를 이용하여 학습한 형태이고, 그림4는 MSC-SVDD를 이용하여 학습한 형태이다. 위의 그림들을 보면, KMSVDD를 이용하여 학습할 경우 왼쪽 클래스의 기다란 분포와 오른쪽 클래스의 가운데 있는 작은 분포를 양쪽의 데이터 분포에서 침범하여 표현됨을 알 수 있다. 이는 유사한 크기의 타원 형태로 clustering이 이뤄지는 K-means clustering의 특징에 의한 것으로써 학습된 형태는 데이터 분포의 특징을 잃어버리게 된다. 이와는 달리 Mean Shift Clustering을 이용한 SVDD의 결과는 왼쪽 클래스의 기다란 분포와 오른쪽 클래스의 가운데 있는 작은 분포를 모두 잘 보존한 형태로 학습이 이루어진다.

3.3. Mean Shift Window Bandwidth에 따른 학습 형태

Mean Shift window의 bandwidth는 K-means의 K와 같이 사용자가 정해 주어야 하는 계수지만, 그림5에서 볼 수 있듯이 bandwidth가 5에서 9까지는 동일한 학습 결과를 나타내고 11일 경우 또한 거의 유사한 결과를 얻을 수 있음을 통해 이 값은 K와는 달리 학습 결과에 크게 영향을 주지는 않음을 확인할 수 있다.

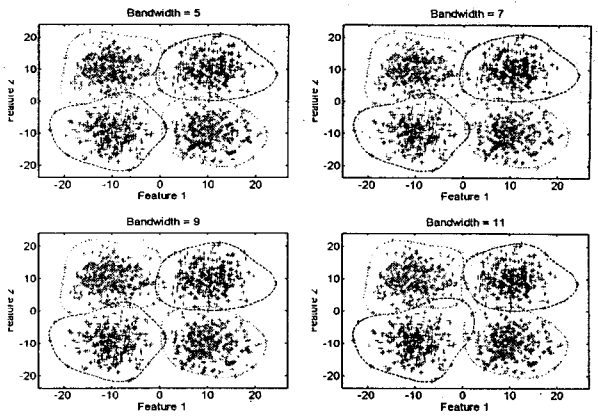

그림 5. Mean Shift window의 bandwidth를 변화시켜가며 학습한 결과 비교
(좌상) BW=5, (좌우) BW=7, (좌하) BW=9, (우하) BW=11

5. 결 론

본 논문에서는 SVDD를 사용하여 많은 수의 데이터를 학습하는데 있어서, 학습 시간을 단축하면서도 데이터 분포 상태의 특징을 보존하기 위한 방법을 제시하였다. 기존에 학습 시간 단축을 위한 알고리즘으로 KMSVDD가 제안되었으나, 이는 데이터 분포와 무관하게 임의로 K개의 소집단으로 나누므로 학습 결과는 데이터 분포 특징을 잃게 되는 문제가 있었다. Mean Shift Clustering 알고리즘을 사용하여 데이터 분포 밀도상의 mode를 중심으로 이루어져 있는 데이터 분포의 특징을 유지하며 여러 개의 소집단으로 데이터 영역을 나눈 후, 각 소집단 별로 SVDD 학습을 하여 계산 시간을 단축할 수 있는 알고리즘을 제안하였다. 모의실험 결과는 MSC-SVDD의 학습 시간 감소 효과가 KMSVDD와 비교하여 큰 차이가 없음을 보여주었고, 데이터 분포의 특징을 보존하며 학습이 이뤄짐을 확인할 수 있었다. 또한 K값이 학습 결과에 큰 영향을 주는 것과는 달리 Mean Shift의 window bandwidth의 선택은 큰 영향을 주지 않아 선택의 폭이 큰 이점이 있음을 확인할 수 있었다.

향후 과제에서는 데이터 분포의 특징을 이용한 특징 추출 및 학습 방법을 연구해 보고자 한다.

참 고 문 헌

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

[2] 김표재, 장형진, 송동성, 최진영, "KMSVDD: K-Means Clustering을 이용한 Support Vector Data Description," Information and Control Symposium, 2006.

[3] K. Fukunaga and L.D. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," IEEE Trans. Information Theory, vol. 21, pp. 32-40, 1975.

[4] Y. Cheng, "Mean shift, mode seeking, and clustering," IEEE Trans. Pattern Anal. Machine Intell., vol. 17, 790-799, 1995.

[5] D. Comaniciu, and P. Meer, "Mean shift: A robust approach toward feature space analysis," IEEE Trans. Pattern Anal. Machine Intell., pages 24(5):603-619, 2002.