PoMEN 기반의 Latent One-Class SVM

이창기[O]
강원대학교, 컴퓨터과학과
leeck@kangwon.ac.kr

PoMEN based Latent One-Class SVM

Changki Lee[O]
Kangwon University, Computer Science department

요 약

One-class SVM은 데이터가 존재하는 영역을 추출하고, 이 영역을 서포트 벡터로 표현하며 표현된 영역 밖의 데이터들은 아웃라이어(outlier)로 간주된다. 본 논문에서는 데이터 포인트마다 숨겨진 변수(hidden variable) 혹은 토픽이 있다고 가정하고, 이를 반영하기 위해 PoMEN에 기반한 Latent One-class SVM을 제안한다. 실험결과 Latent One-class SVM이 대부분의 구간에서 One-class SVM 보다 성능이 높았으며, 특히 높은 정확율을 요구하는 경우에 더욱 효과적임을 알 수 있었다.

주제어: One-Class SVM, PoMEN, Latent One-Class SVM

1. 서론

One-class SVM은 비교사 학습(unsupervised learning)의 하나로 볼 수 있으며, 주어진 데이터를 잘 설명할 수 있는 최적의 서포트 벡터(support vector)를 구하는 기술이다 [1, 2]. One-class SVM은 데이터가 존재하는 영역을 추출하며, 이 영역을 서포트 벡터로 표현한다. 표현된 영역 밖의 데이터들은 아웃라이어(outlier)로 간주된다. One-class SVM은 신규성/이상 탐지(novelty/abnormal detection), 이미지 검색(image retrieval), 문서/텍스트 분류(document/text classification) 등에 성공적으로 적용되었다 [3, 4, 5, 6].

본 논문에서는 데이터 포인트마다 숨겨진 변수(hidden variable) 혹은 토픽이 있다고 가정하고, 이를 반영하기 위해 Partially observed Maximum Entropy Discrimination Markov Network (PoMEN)에 기반한 Latent One-class SVM을 제안한다.

본 논문은 다음과 같이 구성된다. 2장에서는 먼저 One-class SVM에 대해서 소개하고, 3장에서는 PoMEN에 대해서 살펴본다. 4장에서는 PoMEN에 기반한 Latent One-class SVM을 제안하고, 5장에서는 실제 데이터를 가지고 실험 및 결과를 분석한다. 6장에서는 결론 및 향후 연구 과제를 제시한다.

2. One-class SVM

본 장에서는 One-class SVM에 대해서 간략히 소개한다. 학습데이터를 $X=\{x_1, \ldots, x_n\}$, $x_i \in R^d$라 하고, Φ를 피처(feature) 매핑 함수라 하면, 피처 공간에서의 dot product는 커널을 통하여 $K(x,y)=\Phi(x)^T\Phi(y)$ 와 같이 계산된다. One-class SVM은 다음과 같이 정의된다 [1].

$$\min_{\mathbf{w},\rho} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i - \rho$$

$$s.t. \forall i, \mathbf{w}^T\Phi(\mathbf{x}_i) \geq \rho - \xi_i, \xi_i \geq 0, \quad (1)$$

위 식에서 $0 < \nu < 1$ 로써 outlier의 비율을 조절하는 파라미터이며 사용자가 지정한다. 조건부 최적화 문제인 수식 (1)을 풀기위해 라그랑지(Lagrange) 변수 α_i를 도입한 후 dual problem으로 변환하면 다음과 같다.

$$\max_{\alpha} -\sum_i\sum_j \alpha_i\alpha_j K(\mathbf{x}_i,\mathbf{x}_j)$$

$$s.t.\ 0 \leq \alpha_i \leq \frac{1}{\nu n}, \sum_i \alpha_i = 1. \quad (2)$$

또한 $\mathbf{w} = \sum_i \alpha_i \Phi(\mathbf{x}_i)$ 수식이 얻어지며, 이로부터 원점으로부터의 거리가 $\rho/\|\mathbf{w}\|$인 다음과 같은 결정(decision) 함수를 얻게 된다 (즉, f(x) >= 1 이면 +1, f(x) < 1 이면 -1 이 된다).

$$f(\mathbf{x}) = \mathbf{w}^T\Phi(\mathbf{x}) - \rho = \sum_i \alpha_i K(\mathbf{x}_i,\mathbf{x}) - \rho.$$

위 식에서, α_i값은 quadratic programming으로 결정할 수 있으며, ρ값은 서포트 벡터(support vector)인 x_s가 $f(x_s)=0$을 만족한다는 성질을 이용하여 구할 수 있다.

또 다른 형태의 One-class SVM인 Support Vector Data Description(SVDD)은 다음과 같이 정의된다 [2].

$$\min_{R,a} R^2 + C\sum_{i=1}^{n}\xi_i$$

$$s.t. \forall i, \|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2 \leq R^2 + \xi_i, \xi_i \geq 0, \quad (3)$$

위 식에서 a는 구의 중심이다. 조건부 최적화 문제인 수식 (3)을 풀기위해 라그랑지 변수 α_i를 도입한 후 dual problem으로 변환하면 다음과 같다.

$$\max_{\alpha} \sum_i \alpha_i K(\mathbf{x}_i, \mathbf{x}_i) - \sum_i \sum_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$$
$$s.t.\ 0 \le \alpha_i \le C, \sum_i \alpha_i = 1. \quad (4)$$

만약 커널 함수가 $K(\mathbf{x}_i, \mathbf{x}_i) = 1$을 만족하면 (예를 들어 RBF 커널) 수식 (4)는 수식 (2)와 같아진다. 즉, RBF 커널을 사용하면 SVDD와 One-class SVM이 같은 모델이 된다. 수식 (3)에서의 구의 중심은 다음과 같이 표현된다.

$$\boldsymbol{a} = \sum_i \alpha_i \Phi(\mathbf{x}_i)$$

α_i값은 quadratic programming으로 결정할 수 있으며, 수식 (3)에서의 구의 반지름 R값은 서포트 벡터인 x_s가 $\|\Phi(\mathbf{x}_s) - \boldsymbol{a}\|^2 = R^2$을 만족한다는 성질을 이용하여 구할 수 있다.

3. PoMEN

본 장에서는 Partially observed Maximum Entropy Discrimination Markov Network (PoMEN)에 대해서 간략히 설명한다. PoMEN은 Maximum Entropy Discrimination Markov Network (MaxEnDNet)의 확장으로, 숨겨진 변수(hidden variable)를 모델에 추가하였다 [7].

MaxEnDNet은 Maximum Entropy Discrimination (MED)의 확장으로, Max-Margin Markov Network (M^3N) 혹은 structural SVM과 같은 max-margin 학습과 Bayesian-style의 학습을 결합하여 사용하며, 다음과 같이 정의된다 [8].

$$\min_{p(\mathbf{w}),\xi} KL(p(\mathbf{w}) \| p_0(\mathbf{w})) + U(\boldsymbol{\xi}), \quad (5)$$
$$s.t. \int p(\mathbf{w}) \left[\mathbf{w}^\top \Delta \mathbf{f}(\mathbf{x}_i, \mathbf{y}) - \mathrm{L}(\mathbf{y}_i, \mathbf{y}) \right] d\mathbf{w} \ge -\xi_i,\ \forall i, \forall \mathbf{y}$$

위 식에서 $\Delta f(x_i,y) = f(x_i,y_i) - f(x_i,y)$ 이며, KL은 KL-divergency이며, U는 convex 함수이다. 수식 (5)를 풀기위해 라그랑지 변수 α_i를 도입하면 다음과 같은 식을 얻을 수 있다.

$$p(\mathbf{w}) = \frac{1}{Z(\alpha)} p_0(\mathbf{w}) \exp\left\{ \sum_{i,\mathbf{y}} \alpha_i(\mathbf{y}) \left[\mathbf{w}^\top \Delta \mathbf{f}(\mathbf{x}_i, \mathbf{y}) - \mathrm{L}(\mathbf{y}_i, \mathbf{y}) \right] \right\}$$

위 식에서 $Z(\alpha)$는 정규화 부분(normalization factor)이며, 라그랑지 변수 α_i값은 수식 (5)의 dual problem으로부터 구할 수 있으며, dual problem은 다음과 같다.

$$\max_{\alpha} -\log Z(\alpha) - U^*(\alpha) \quad (6)$$
$$s.t.\ \alpha_i(\mathbf{y}) \ge 0,\ \forall i,\ \forall \mathbf{y}$$

위 식에서 U*는 U의 conjugate이다. $U(\xi) = C\sum_i \xi_i$이고 $p_0(\mathbf{w})$가 가우시안 분포 $N(\mathbf{w} \mid 0, I)$를 따른다면, $p(\mathbf{w})$도 가우시안 분포를 따르며 평균값은 $\sum_{i,\mathbf{y}} \alpha_i(\mathbf{y}) \Delta \mathbf{f}(\mathbf{y}_i, \mathbf{y})$이 되고, $\alpha_i(\mathbf{y})$값은 수식 (6)으로부터 구할 수 있으며, 수식 (6)은 M^3N 혹은 structural SVM의 dual problem과 같아진다. 즉, 이 경우 MaxEnDNet과 M^3N 혹은 structural SVM의 결과가 같아지게 된다. 따라서 M^3N 혹은 structural SVM은 MaxEnDNet의 특수한 경우라고 말할 수 있다.

PoMEN은 MaxEnDNet 모델에 숨겨진 변수(hidden variable)를 추가한 것으로 다음과 같이 정의된다.

$$\min_{p(\mathbf{w},\{\mathbf{z}\}),\xi} KL(p(\mathbf{w},\{\mathbf{z}\}) \| p_0(\mathbf{w},\{\mathbf{z}\})) + U(\boldsymbol{\xi}), \quad (7)$$
$$s.t. \sum_{\mathbf{z}} \int p(\mathbf{w},\mathbf{z}) \left[\mathbf{w}^\top \Delta \mathbf{f}(\mathbf{x}_i, \mathbf{y}, \mathbf{z}) - \mathrm{L}(\mathbf{y}_i, \mathbf{y}) \right] d\mathbf{w} \ge -\xi_i,$$
$$\forall i, \forall \mathbf{y}$$

위 식에서 $\Delta f(x_i,y,z) = f(x_i,y_i,z) - f(x_i,y,z)$이며, $\mathbf{z}$는 숨겨진 변수(hidden variable)로 $\{\mathbf{z}\} = (z_1, \ldots, z_n)$이다. 계산량의 문제로, 일반적으로 $p_0(\mathbf{w},\{\mathbf{z}\})$와 $p(\mathbf{w},\{\mathbf{z}\})$를 다음과 같이 가정한다.

$$p_0(\mathbf{w},\{\mathbf{z}\}) = p_0(\mathbf{w}) \prod_{i=1}^{n} p_0(\mathbf{z}_i),$$
$$p(\mathbf{w},\{\mathbf{z}\}) = p(\mathbf{w}) \prod_{i=1}^{n} p(\mathbf{z}_i)$$

학습을 위해서 EM-style의 알고리즘을 사용하며, 1단계에서는 $p(\mathbf{z})$를 고정한 후 $p(\mathbf{w})$를 구하고, 2단계에서는 $p(\mathbf{w})$를 고정하고 $p(\mathbf{z})$를 구한다. 1, 2 단계를 반복하여 학습을 수행한다.

4. PoMEN 기반의 Latent One-class SVM

본 장에서는 기존의 One-class SVM의 성능을 높이기 위해서 숨겨진 변수(hidden variable) 혹은 토픽을 도입한 Latent One-class SVM를 제안한다. 이를 위해서 PoMEN으로부터 max-margin 학습과 Bayesian-style의 학습을 이용하여 Latent One-class SVM을 아래와 같이 도출한다. 먼저 결정(decision) 함수 h(x)를 다음과 같이 정의한다.

$$h(\mathbf{x}) = \mathrm{sign}\left(\sum_{z} \int p(\mathbf{w}, z) \left[\mathbf{w}^\top \mathbf{f}(\mathbf{x}, z) - \rho \right] d\mathbf{w} \right) \quad (8)$$

위 식에서 z는 $\mathbf{x}$에 대응하는 숨겨진 변수(hidden variable) 혹은 토픽이다. Latent One-class SVM을 max-margin 학습과 Bayesian-style의 학습을 이용하여 다음과 같이 정의한다.

$$\min_{p(\mathbf{w},\{z\}),\xi} KL(p(\mathbf{w},\{z\}) \| p_0(\mathbf{w},\{z\})) + U(\boldsymbol{\xi}), \quad (9)$$
$$s.t. \sum_{z} \int p(\mathbf{w}, z) \left[\mathbf{w}^\top \mathbf{f}(\mathbf{x}_i, z) - \rho \right] d\mathbf{w} \ge -\xi_i,\ \forall i$$

PoMEN과 같이 계산량의 문제로 $p_0(\mathbf{w},\{z\})$와 $p(\mathbf{w},\{z\})$를 다음과 같이 가정한다.

$$p_0(\mathbf{w},\{z\}) = p_0(\mathbf{w}) \prod_{i=1}^{n} p_0(z_i),$$

$$p(\mathbf{w},\{z\}) = p(\mathbf{w})\prod_{i=1}^{n} p(z_i)$$

학습을 위해서 PoMEN과 같이 EM-style의 알고리즘을 사용하며, 1단계에서는 p(z)를 고정한 후 p(w)를 구하고, 2단계에서는 p(w)를 고정하고 p(z)를 구하며, 1, 2 단계를 반복하여 학습을 수행한다. 1, 2 단계는 다음과 같이 수행된다.

- 1 단계: p(z)를 고정한 후, 다음과 같이 정의된 문제를 이용하여 p(w)를 구한다.

$$\min_{p(\mathbf{w}),\xi} KL(p(\mathbf{w})\|p_0(\mathbf{w})) + C\sum_i \xi_i, \quad (10)$$
$$s.t. \int p(\mathbf{w}) E_{p(z)}\left[\mathbf{w}^\top \mathbf{f}(\mathbf{x_i},z) - \rho\right] d\mathbf{w} \geq -\xi_i, \forall i$$

위 식을 풀기위해 라그랑지 변수 α_i를 도입하면 다음과 같은 식을 얻을 수 있다.

$$p(\mathbf{w}) = \frac{1}{Z(\alpha)} p_0(\mathbf{w}) \exp\left\{\sum_i \alpha_i E_{p(z)}\left[\mathbf{w}^\top \mathbf{f}(\mathbf{x_i},z) - \rho\right]\right\}$$

위 식에서 $p_0(\mathbf{w})$가 가우시안 분포 $N(\mathbf{w} \mid 0,I)$를 따른다고 가정하면 $p(\mathbf{w})$도 가우시안 분포 $N(\mathbf{w} \mid \mu_\mathbf{w},I)$를 따름을 증명할 수 있다. 평균값은 $\mu_\mathbf{w} = \sum_i \alpha_i E_{p(z)}\left[\mathbf{f}(\mathbf{x_i},z)\right]$로 구할 수 있다. 라그랑지 변수 α_i값은 수식 (10)의 dual problem으로부터 구할 수 있으며, dual problem은 다음과 같다.

$$\max_\alpha \sum_i \alpha_i - \frac{1}{2}\left\| \sum_i \alpha_i E_{p(z)}\left[\mathbf{f}(\mathbf{x_i},z)\right]\right\|^2 \quad (11)$$
$$s.t.\ 0 \leq \alpha_i \leq C, \forall i$$

위 식은 M^3N 혹은 structural SVM의 dual form과 거의 동일하다. 따라서 기존의 M^3N 혹은 structural SVM의 학습 알고리즘을 그대로 사용할 수 있다. 또는 다음과 같은 primal problem으로 변환하여 기존의 학습 알고리즘을 사용하면 된다.

$$\min_{\mathbf{w},\xi} \frac{1}{2}\mathbf{w}^\top \mathbf{w} + C\sum_{i=1}^{n} \xi_i,$$
$$s.t.\ \mathbf{w}^\top E_{p(z)}\left[\mathbf{f}(\mathbf{x_i},z)\right] \geq \rho - \xi_i;\ \xi_i \geq 0, \forall i$$

- 2 단계: p(w)를 고정한 후, $p_0(\{z\}) = \prod_{i=1}^{n} p_0(z_i)$와 $p(\{z\}) = \prod_{i=1}^{n} p(z_i)$라고 가정한 것을 이용하여 각각의 학습데이터 $\mathbf{x_i}$에 대한 p(z) 분포를 다음과 같이 정의된 문제를 이용하여 구한다.

$$\min_{p(z),\xi_i} KL(p(z)\|p_0(z)) + C\xi_i, \quad (12)$$
$$s.t. \sum_z p(z) \int p(\mathbf{w})\left[\mathbf{w}^\top \mathbf{f}(\mathbf{x_i},z) - \rho\right] d\mathbf{w} \geq -\xi_i$$

위 식의 조건식은 $\sum_z p(z)\left[\mu_\mathbf{w}^\top \mathbf{f}(\mathbf{x_i},z) - \rho\right] \geq -\xi_i$로 표현할 수 있다. 위 식을 풀기위해 라그랑지 변수 β_i를 이용하면 다음과 같은 식을 얻을 수 있다.

$$p(z) = \frac{1}{Z(\beta_i)} p_0(z) \exp\left\{\beta_i\left[\mu_\mathbf{w}^\top \mathbf{f}(\mathbf{x_i},z) - \rho\right]\right\}$$

위 식에서 $p_0(z)$는 uniform 분포를 사용하며, 라그랑지 변수 β_i는 다음과 같은 dual problem으로 구할 수 있다.

$$\frac{\partial \log Z(\beta_i)}{\partial \beta_i} = \mu_\mathbf{w}^\top E_{p(z)}\left[\mathbf{f}(\mathbf{x_i},z)\right] - \rho$$

5. 실험

실험을 위해서 stochastic gradient descent (SGD)를 이용해서 Latent One-class SVM을 C++로 구현하였다. One-class SVM과의 비교 실험을 위해서 LIBSVM을 이용하였다 [9]. 모든 실험은 linear kernel을 사용했으며, 16GB 메모리의 Intel Core i7-2600 CPU PC를 사용하였다.

One-class SVM(LIBSVM)과 본 논문에서 제안하는 Latent One-class SVM을 실제 데이터인 US postal service OCR data set을 사용하였다 [10]. 이 데이터 셋은 총 9298개의 16*16=256 사이즈 숫자 이미지로 구성되어 있으며 마지막 2007개의 이미지는 테스트에 사용하였다. Latent One-class SVM의 성능을 측정하기 학습셋의 숫자 2와 7만을 positive example로 사용하였고, 테스트 셋에서는 2와 7을 positive example으로, 나머지 숫자들을 negative example로 사용하였다. Latent One-class SVM의 hidden variable의 종류, 즉 토픽의 개수는 2로 지정하여 실험하였다.

그림 1은 위 데이터 셋에서의 숫자 2와 7의 탐지에 관한 Receiver Operating Characteristic (ROC) 그래프이다. x축은 숫자 2와 7의 탐지의 False positive ratio를 의미하고 y축은 True positive ratio를 의미한다. False positive ratio의 값이 작을수록 정확율(precision)이 높은 것이고, True positive 값이 클수록 재현율(recall)이 높은 것이다. 그림 1로부터 Latent One-class SVM이 대부분의 구간에서 One-class SVM (LIBSVM) 보다 성능이 높음을 알 수 있다. 특히 False positive ratio 값이 작은 경우 (즉, 정확율이 높은 경우), True positive ratio값이 5% 이상 높은 것을 알 수 있다 (즉, 재현율이 5% 이상 높음). 따라서 Latent One-class SVM은 높은 정확율을 요구하는 경우에 더욱 효과적임을 알 수 있다.

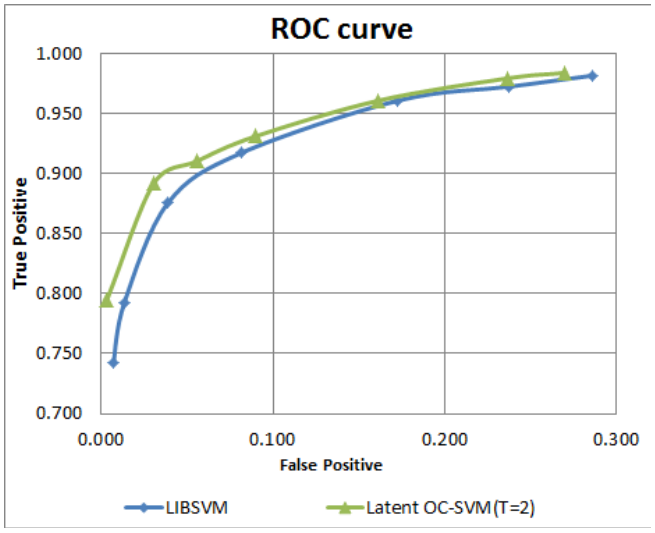

그림 1. US postal service OCR 데이터 셋에서의 ROC curves (숫자 2와 7의 탐지)

6. 결론

One-class SVM은 데이터가 존재하는 영역을 추출하고, 이 영역을 서포트 벡터로 표현하며 표현된 영역 밖의 데이터들은 아웃라이어(outlier)로 간주된다. 본 논문에서는 데이터 포인트마다 숨겨진 변수(hidden variable) 혹은 토픽이 있다고 가정하고, 이를 반영하기 위해 PoMEN에 기반한 Latent One-class SVM을 제안하였다. 실험결과 Latent One-class SVM이 대부분의 구간에서 One-class SVM 보다 성능이 높았으며, 특히 높은 정확율을 요구하는 경우에 더욱 효과적임을 알 수 있었다.

향후 연구 방향은 Latent One-class SVM의 검증을 위해 좀더 다양한 응용에 적용하고, 더욱 다양한 방법의 hidden variable 적용 방법을 연구하는 것이다.

참고문헌

[1] Scholkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C., Estimating the support of a high-dimensional distribution. Neural Computational 13 (7), 1443-1471, 2001.

[2] Tax, D.M.J., Duin, R.P.W., Support vector domain description. Machine Learning 54, 45-66, 2004.

[3] Chen, Y., Zhou, X., Huang, T.S., One-class SVM for learning in image retrieval. In Proc. IEEE International Conf. on Image Processing, 2001.

[4] Deng, H., Xu, R., Model selection for anomaly detection in wireless ad hoc networks. In Proc. IEEE Symp. On Comput. Intell. Data Mining (CIDM 2007), pp. 540-546, 2007.

[5] Manevitz, L.M., Yousef, M., One-class SVMs for document classification. Journal of Machine Learning Research 2, 139-154, 2001.

[6] Zhang, Y., Li, X., One-class classification of text streams with concept drift. 2008 IEEE International Conference on Data Mining Workshops, 2008.

[7] Zhu, J., Xing, E.P., Zhang, B., Partially observed maximum entropy discrimination markov networks. In NIPS, 2008.

[8] Zhu, J., Xing, E.P., Zhang, B., Maximum entropy discrimination markov networks. Technical Report CMU-ML-08-104, Machine Learning Department, Carnegie Mellon University, 2008.

[9] Chang, C.C., Lin, C.J., LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1-27:27, 2011.

[10] Hull, J.J., A database for handwritten text recognition research. IEEE Transaction on Pattern Analysis and Machine intelligence, 16 (5), 550-554, 1994.