F-022

異常検出サポートベクトルマシン
Anomaly Detection Support Vector Machine

藤巻遼平 †
Ryohei Fujimaki

1 はじめに

一般に，分類問題ではテストデータが学習されたいずれかのクラスに属する事を暗黙に仮定している．しかし，実問題ではこの前提条件が成立しない状況にしばしば直面する．例えば，故障分類問題では学習期間には発生しなかった，あるいは学習に足るデータ数が得られなかった故障がテスト期間では発生しうるし，文章分類問題ではテスト期間に新たなトピックが発生する事は容易に想像できる．さらに，テストデータはクラスラベルがわからないため，前処理でそのような未学習クラスのデータを取り除く事はそれほど簡単な問題ではない．

この状況では，分類が不明瞭なテストデータは未学習のクラス (異常クラス) に属する可能性が高いため「異常」として検出して，人間に判断を仰ぐことで誤分類リスクを避けることが望ましい．これは，誤分類リスクの低減に加えて，データが真に異常クラスに属するデータであれば，新たなクラス情報を獲得可能という利点もある．

従来，上述の問題に対しては，1) 異常検出アルゴリズムによってテストデータから異常データを除外する，2) 通常の分類アルゴリズムによって学習済みのクラスへ分類する，という 2 段階の戦略がとられ [2]，1 段目の異常検出問題と 2 段目の分類問題は別々に研究されてきた．この 2 つの問題をつなぐ経験則として，異常データを擬似的に一様分布から生成し，通常の訓練データとその異常データを通常の多クラス分類問題として解く手法が提案されている [4, 9]．Steinwart らは，この経験則を一般化し，任意の分布から異常データを生成する事によって，真の分布への一致性を保って異常検出問題を分類問題へ変換可能である事を示している [7]．しかし，実問題では異常クラスのデータに対する事前知識を得ることは難しく，異常クラスの分布選択に対する明快な答えは今のところ与えられていない．

本稿で提案する異常検出サポートベクトルマシン (Anomaly Detection Support Vector Machine; ADSVM) は，各クラスに対して 2 つの分離境界を設定し，学習クラスのデータに対して分類と異常検出に関する二つの評価基準を同時に最適化することによって，上記の問題を解決する．実証実験では，SVM [6, 10] と One-class SVM (1SVM) [5] を組み合わせた素朴な方法と比較して分類精度が向上することを確認した．提案手法は，前述の素朴な手法に対する精度的優位性に加えて，Steinwart ら [7] とは異なる観点から分類問題と異常検出問題を統一的に扱うための指針を与えるという意味でも重要な意味を持つ．

2 異常検出サポートベクトルマシン

本稿では，学習クラスが $\mathcal{C} = \{C_1, C_2\}$ の 2 クラス問題を扱い，異常クラスは $\mathcal{A}$ と表記する．

まず，学習時には $\{\mathcal{C}\}$ に属するデータのみが得られているとし，学習データおよびクラスラベルを $\mathcal{X} = \{x_i^j, y_i^j\}$ $(j = 1, 2; i = 1, \ldots N_j)$ とする．ここで，x_i^j は C_j に属する i 番目のデータを意味し，$y_i^j = C_j$ である．この時，テストデータ x に対するラベル y を予測する事がタスクとなる．

以下では，関数 ϕ で決まる特徴空間が再生核ヒルベルト空間である事を仮定する．これは，特徴空間に写像された $\phi(x)$ に関し，その内積がカーネル関数 $k(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle$ によって計算される事を意味している．

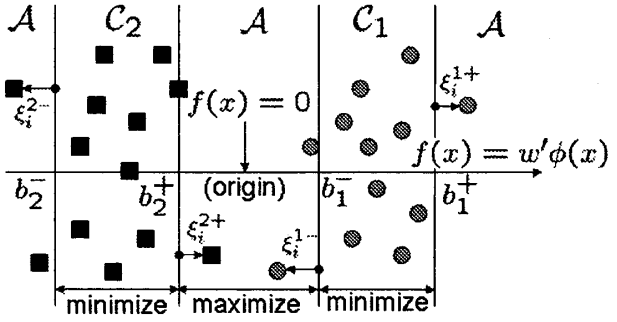

図 1: 異常検出サポートベクトルマシンの概念図.

2.1 概念説明

ADSVM は，通常の SVM と同様に分類関数を $f(x) = w'\phi(x)$ として，重みベクトル w を学習する．ただし，$'$ はベクトルまたは行列の転置を表す．この時，ADSVM では分類問題と異常検出問題を同時に解くために，通常の SVM で利用される正則化リスク [10] に加え，以下の 3 つの概念に従って最適化の基準を設計する (図 1)．

A) 平行超平面 ADSVM では，各学習クラスに対して一組の平行超平面を割当，それらを同時に最適化する (つまり 2 クラス問題では 4 つの超平面)．各超平面組は，各クラスの分類関数 $f(x)$ の上限と下限に対応し，図 1 に示されるように $C_1, C_2, \mathcal{A}$ の領域を規定する．

B) 学習クラス領域の最小化 ADSVM では，各学習クラスに割り当てられた領域，すなわち各超平面組の間の領域を最小化する．これは，support vector data description [8] などの異常検出器が採用する「各クラスを包含するクラス領域最小化」という基準と同様のものであり，異常検出力を高める基準として機能する．

C) $f(x) = 0$ の近傍領域最大化 ADSVM では，$f(x) = 0$ の近傍領域を $\mathcal{A}$ に割当て，その領域を最大化する．これによって，C_1 と C_2 を $f(x) = 0$ の対岸に配置した場合，この基準は SVM のマージン最大化と同じ働きをする (図 1) ため，分類誤差の低減効果が期待される．これと同時に，異常クラスのデータは $f(x) = 0$ 付近に分布しやすいという観察から，異常検出力の向上も期待される．

ここで，異常クラスのデータが $f(x) = 0$ 付近に分布しやすいという点に関して，直感的な説明を与える．いま，$\mathrm{span}(\mathcal{X})$ および $\overline{\mathrm{span}(\mathcal{X})}$ を，$\{\phi(x_i^j)\}$ によって張られる空間とその補空間とし，各空間の基底を ψ_τ および $\overline{\psi_{\tau'}}$ とする．まず，異常クラスのデータは，学習クラスのデータとは異なる分布から生成されるため，学習クラスのデータよりも $\overline{\mathrm{span}(\mathcal{X})}$ に分布しやすいと仮定することは不自然ではないだろう．次に，$f(x)$ がある重み a_i^j に対して $f(x) = \sum_{i,j} a_i^j \phi(x_i^j)$ と表現され，かつ $\phi(x_i^j)$

†NEC 共通基盤ソフトウェア研究所

が ψ_τ $(\tau = 1, \ldots)$ の線形結合として表現される事に注意すると，$f(x)$ はある重み b_τ に対して，$f(x) = \sum_\tau b_\tau \psi_\tau$ と表現される．この時，補空間の定義より，$f(\overline{\psi}) = \sum_\tau b_\tau \langle \psi_\tau, \overline{\psi} \rangle = 0$ であるため，異常クラスのデータは $f(x) = 0$ 付近に分布しやすくなる．

2.2 最適化問題の導出

前節で説明した概念を元に，ADSVM の最適化問題は以下のように定式化される．

$$\min \left\{ \frac{1}{2} \boldsymbol{w}'\boldsymbol{w} + \frac{1}{N} \sum_{j=1}^{2} \sum_{i=1}^{N_j} (\xi_i^{j+} + \xi_i^{j-}) \right. \tag{1}$$

$$\left. + \nu_1 \sum_{j=1}^{2} (b_j^+ - b_j^-) - \nu_0 (b_1^- - b_2^+) \right\}$$

subject to

$$\boldsymbol{w}'\phi(x_i^j) - b_j^+ \leq \xi_i^{j+} \quad \boldsymbol{w}'\phi(x_i^j) - b_j^- \geq -\xi_i^{j-} \tag{2}$$

$$b_j^+ - b_j^- \geq 0 \quad \xi_i^{j\pm} \geq 0 \quad b_1^- \geq 0 \quad b_2^+ \leq 0 \tag{3}$$

ここで，$\xi_i^{j\pm}$ は x_i^j に対するスラック変数で $b_j^\pm$ に対する hinge loss を表す．ν_1 および ν_0 は，モデルの複雑性と訓練誤差を制御するパラメータで，N は学習データの総数を表す $(N = \sum_j N_j)$．

初項および第 2 項は，SVM と同様の正則化リスクを表す．SVM (soft-margin loss) とは異なり，hinge loss を利用した理由は紙面の都合上説明を省く．第 3 項は $b_j^\pm$ $(j = 1, 2)$ に挟まれた領域の最小化であり，前節の **B)** に対応する．第 4 項は b_1^- および b_2^+ に挟まれた領域の最大化を表現し，前節の **C)** に対応する．各クラスの位置に関しては，(3) の条件によって制約が課されている．

上記最小化問題または，その双対問題を解くことで，テストデータに対するラベルは，

$$y = \begin{cases} \mathcal{C}_j & \text{if} \quad b_j^- \leq f(x) \leq b_j^+ \\ \mathcal{A} & \text{otherwise} \end{cases} \tag{4}$$

と推定される．

本稿では，紙面の都合上 (1) から (3) の双対問題，Karush-Kuhn-Tucker 条件，バイアス項 $b_j^\pm$ の導出方法などは割愛する．

3 実験と考察

UCI データ [1] 中の，Iris，Wine データ (各々3 クラスデータ) を利用し，1 クラスを異常クラスとして学習には利用せず，残り 2 クラスを学習に利用した．また，評価は 1SVM を適用し，異常が検出された場合には $\mathcal{A}$ に，異常が検出されなかった場合には SVM を適用し $\mathcal{C}$ のいずれかに分類する，2 段階アルゴリズム (Two-step) と精度を 10-fold 交差検定によって比較した．なお，1SVM および SVM の実装には，LIBSVM [3] を利用した．

分類および異常検出の精度は，$F_C = 2P_C R_C/(P_C + R_C)$ および $F_A = 2P_A R_A/(P_A + R_A)$ を利用した．ただし，$P_C = \sum_{i=1}^{M} TP_i/(\sum_{i=1}^{M}(TP_i + MC_i) + FN)$，$R_C = \sum_{i=1}^{M} TP_i/(\sum_{i=1}^{M}(TP_i + MC_i + FP_i))$，$P_A = TP/(TP + FN)$，$R_A = TP/(TP + \sum_{i=1}^{M} FP_i)$ であり，F_A は異常検出に対する通常の F 値を表し，F_C は分類精度評価のために F 値と同様の定義をした指標となっている．ただし，TN_i, MC_i, FP_i, FN, FP は図 2 に示されるように定義される．

		Predicted Label		
		$\mathcal{C}_i$	$\mathcal{C} \setminus \mathcal{C}_i$	$\mathcal{A}$
True Label	$\mathcal{C}_i$	TN_i	MC_i	FP_i
	$\mathcal{A}$	FN		TP

図 2: 推定されたラベルに対する真のラベルの関係 $(i = 1, 2)$．行列の中はデータ数．

表 1: UCI データに対する実験結果

	F_C / F_A	
データ	ADSVM	Two-step
Iris	0.83 / 0.48	0.53/0.48
Wine	0.57 / 0.46	0.40/0.47

表 1 に各データ，各手法に対する F_C および F_A をまとめる．ADSVM と Two-step では，異常検出精度 (F_A) に関しては優位水準 5% の t 検定で優位な差は見られなかった．一方で，分類精度 (F_C) では ADSVM が Two-step を大きく凌駕しており，分類および異常検出を同時に扱う問題に対する ADSVM の有効性が確認された．

4 結論

本稿では，分類問題においてテストデータが未学習のクラスに属するデータを含む問題を解決するために，異常検出サポートベクトルマシン (ADSVM) を提案した．ADSVM は，分類問題と異常検出問題を統一的に扱うための新しい枠組みを与えている．UCI データによる実証実験を通じて，通常の SVM と 1SVM を組み合わせた素朴な方法に対し精度の面で優位性を持つ事を確認した．本稿では，2 クラスの場合のみを扱ったため，学習クラスが 3 クラス以上の多クラス問題への拡張が今後の課題として挙げられる．

参考文献

[1] A. Asuncion and D. Newman. UCI machine learning repository. http://www.ics.uci.edu/~mlearn/MLRepository.html, 2007. University of California, Irvine, School of Information and Computer Sciences.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006.

[3] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[4] W. Fan, M. Miller, S. Stolfo, W. Lee, and P. Chan. Using artificial anomalies to detect unknown and known network intrusions. *Knowledge and Information Systems*, 6:507–527, 2004.

[5] B. Scholkopf, J. Platt, J. Shawe-Taylor, A. J. Smola, and R. Williamson. *Estimating the support of a high-dimensional distribution*. Technical Report MSR-TR-99-87, Microsoft Research, 1999.

[6] B. Scholkopf, A. Smola, R. Williamson, and P. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.

[7] I. Steinwart, D. Hush, and C. Scovel. A classification framework for anomaly detection. *Journal of Machine Learning Research*, 6:211–232, 2005.

[8] D. Tax and R. Duin. Support vector data description. *Machine Learning*, 54:45–66, 2004.

[9] D. M. J. Tax and R. P. Duin. Uniform object generation for optimizing one-class classifiers. *Journal of Machine Learning Research*, 2:155–173, 2001.

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