日本オペレーションズ・リサーチ学会
2005年春季研究発表会

Domain Description based on Reliability Learning

02203360 筑波大学 鈴木伸彦 SUZUKI Nobuhiko
01105930 筑波大学 *香田正人 KODA Masato

1 はじめに

本研究では，信頼性測度を用いた信頼性学習を定義し，訓練集合から高密度領域を推定するための新しいアルゴリズムを提案する．

その中でも，信頼性測度の小さい訓練パターンを用いたアルゴリズムは，信頼性測度の選択によって，代表的な高密度領域推定問題（Domain Description 問題，1クラス分類問題）のアルゴリズムである One-Class SVM（OCSVM）[1]，Support Vector Domain Description（SVDD）[2] と一致する．これによって，両アルゴリズムの内容をより論理的に理解することができる．さらに，それらが本質的に等価であることが示される．

また，全ての訓練パターンを用いたアルゴリズムでは，信頼性測度の種類によっては解析的に最適解を求めることができる．これを，重み付けされた訓練パターンを用いて繰り返し実行することによって，より精度の高い推定量を求めることが可能になる．数値実験の結果，この提案するアルゴリズムが誤り率の低い推定量を実現することが確認された．

2 高密度領域推定問題

β 高密度領域 $C(\beta)$ とは，密度関数 p を用いて以下のように表される集合のことである．

$$C(\beta) = \{\mathbf{x}|p(\mathbf{x}) \geq \alpha\}, \quad such\ that \quad P(C(\beta)) = \beta \quad (1)$$

また，密度関数 p のサポート S は確率 1 となる最小閉集合として定義される．

高密度領域推定問題では，クラスラベルの無い訓練パターンから，密度関数の推定を行わずに高密度領域（またはサポート）を推定する．この高密度領域推定問題は，外れ値検出や特異値検出などに用いられている．

3 Domain Description based on Reliability Learning

集合 $C(\beta)$ が β 高密度領域であるならば，高密度領域の定義より明らかに，集合 $C(\beta)$ は入力空間上で有界である．これを，高密度領域の必要条件と呼び，これに基づいて信頼性測度を以下のように定義する．

定義 [信頼性測度] あるパラメータ Θ が与えられたときに関数 $g(\Theta, \mathbf{x})$ が以下の条件を満たすとき，$g(\Theta, \mathbf{x})$ を信頼性測度 (Reliability Measure) と呼ぶ．

(i) $g(\Theta, \mathbf{x})$ は Θ で微分可能である．
(ii) 集合 $\{\mathbf{x}|g(\mathbf{x}) \geq \alpha\}$ は高密度領域の必要条件を満たす．

ある関数 $g(\Theta, \mathbf{x})$ に対して，集合 $\{\mathbf{x}|g(\mathbf{x}) \geq \alpha\}$ が高密度領域の必要条件を満たすならば，$g(\Theta, \mathbf{x})$ は上に有界である．つまり，信頼性測度は上に有界な関数になる．

信頼性測度の値が大きい訓練パターンは高密度領域内のパターンであることの信頼性が高いと考えられる．この信頼性測度を用いて学習を行うことを信頼性学習 (reliability learning) と呼び，様々な問題に適用することができると考えられる．

ここでは，信頼性学習に基づいて β 高密度領域 $C(\beta)$ を推定することを考える．

Domain Description based on Lower Reliability Learning（DDLRL）

信頼性測度が小さい訓練パターンを用いた DDRL を DDLRL と呼ぶ．DDLRL では，信頼性測度の下位 $\nu \times 100\%$ の訓練パターンの条件付き期待値 $\phi_{\text{lower}\nu}$ を最大にするパラメータ Θ を求める．$\phi_{\text{lower}\nu}$ は以下のように与えられる．

$$\phi_{\text{lower}\nu}(\Theta) = E\{g(\Theta, \mathbf{x})|g(\Theta, \mathbf{x}) \leq \alpha\} \quad (2)$$

$\phi_{\text{lower}\nu}$ を最大にするパラメータ Θ は以下の最適化問題を解くことで得られる．

$$\max_{\Theta, \alpha} \quad \alpha - \frac{1}{\nu n} \sum_{i=1}^{n} [-g(\Theta, \mathbf{x}_i) + \alpha]^+ \quad (3)$$

この最適化問題 (3) は，以下の最適化問題と等価である．

$$\begin{aligned} \max_{\Theta, \alpha, \mathbf{z}} \quad & \alpha - \frac{1}{\nu n} \sum_{i=1}^{n} z_i \\ \text{s.t.} \quad & z_i \geq 0, \quad i \in [n], \\ & z_i \geq -g(\Theta, \mathbf{x}_i) + \alpha, \quad i \in [n] \end{aligned} \quad (4)$$

つまり，DDLRL は，最適化問題 (4) を解けばよい．これによって得られる推定量は，高密度領域内の訓練パターンであることに対する信頼性が低い訓練パターンから得られる．そのため，これらの訓練パターンの信頼性を最大にする推定量であると考えられる．

信頼性測度を，ガウシアンカーネルに対応した非線形写像 Φ を用いた $g(\mathbf{w}, \mathbf{x}) = \langle \mathbf{w}, \Phi(\mathbf{x}) \rangle$ としたときの

DDLRLは，OCSVMと一致する．そのため，OCSVMでは$\langle \mathbf{w}, \Phi(\mathbf{x}) \rangle$を信頼性測度とし，信頼性測度の下位$\nu \times 100\%$の訓練パターンを用いて，その期待値を最大にする$\mathbf{w}$を求めていることがわかる．これより，OCSVMで多項式カーネルを用いた場合に，意味の無い推定量が得られるのは，多項式カーネルに対応した特徴空間上での$\langle \mathbf{w}, \Phi(\mathbf{x}) \rangle$が信頼性測度ではないためであると考えられる．逆に，特徴空間上で$\langle \mathbf{w}, \Phi(\mathbf{x}) \rangle$が信頼性測度である場合には，対応したカーネルを用いたOCSVMによって，高密度領域を推定することが可能になる．

また，信頼性測度を$g(\mathbf{a}, \mathbf{x}) = -\|\mathbf{a} - \mathbf{x}\|^2$としたときのDDLRLは，SVDDと一致する．そのため，SVDDでは$-\|\mathbf{a} - \mathbf{x}\|^2$を信頼性測度とし，信頼性測度の下位$\nu \times 100\%$の訓練パターンを用いて，その期待値を最大にする$\mathbf{a}$を求めていることがわかる．

これより，OCSVMとSVDDは信頼性測度が異なるが，本質的には等価であることがわかる．

naive-DDRL

全ての訓練パターンを用いたDDRLをnaive-DDRLと呼ぶ．naive-DDRLでは，全ての訓練パターンの信頼性測度の期待値を最大にするパラメータΘを求める．その際，しきい値αは最適化問題には組み込まず，Θの最適解が得られた後に求めることにする．これはDDLRLのパラメータνの値を$\nu = 1$とした特殊な場合である．求めるパラメータはΘのみであるので，非常に単純なDDRLの1つであり，信頼性測度によっては最適解が解析的に求まる．そのため，DDLRLに比べて，計算コストが非常に小さいと考えられる．

4 Extension of naive-DDRL

ここでは，naive-DDRLを拡張したアルゴリズムを提案する．

重み付けされた訓練パターンを用いて，naive-DDRLを繰り返し実行することを考える．訓練パターンの重みは，naive-DDRLによる信頼性測度の高いパターンに重みをおいて更新される．最終的に得られるβ高密度領域は信頼性測度の上位$\beta \times 100\%$の訓練パターンを用いた推定量に近付くと期待される．

これはnaive-DDRLを拡張したものであるので，Extension of naive-DDRL（EDDRL）と呼ぶ．

EDDRLのアルゴリズムは図1のようになる．

最終的に得られる高密度領域の推定量は以下のようになる．

$$C_n(\beta) = \{\mathbf{x} | g(\Theta_M, \mathbf{x}) \geq \alpha\} \tag{5}$$

ここで，重み関数は各訓練パターンの重みを決定するための関数で，任意の単調増加関数を用いる．重み関数の選択によって，各訓練パターンの重みが変化する．重み関数としては，ガウシアン関数（$f(g) = \exp(-(g-1)^2/s')$）や，ロジスティック関数（$f(g) = 1/(1+\exp(-bg))$）などが考えられる．ガウシアン関数は推定する高密度領域の割合が比較的小さいときに有効であり，ロジスティック関数は推定する高密度領域の割合が大きいときに有効であると考えられる．重み関数やそのパラメータを変えることによって，高密度領域内に含まれる訓練パターンの割合βを調整することができる．重み関数によって違いはあるものの，繰り返し回数が大きくなるとβ，目的関数値ともに収束する．

naive-DDRLの最適解は，信頼性測度によっては解析的に求めることが可能であるから，EDDRLのように繰り返し実行しても大きな計算コストはかからないと考えられる．

step 1: 各訓練パターンの重みd_iを，$d_i = 1/n$によって初期化する．

step 2: 重み付けしたデータを用いてnaive-DDRLを行う．

step 3: 重み関数$f_w(\mathbf{x})$を用いて，以下のようにデータの重みを更新する．

$$d_i = f_w(g(\Theta, \mathbf{x}_i)) / \sum_{i=1}^{n} f_w(g(\Theta, \mathbf{x}_i))$$

step 4: **step 2,3**を$M(\geq 2)$回繰り返す．

step 5: step 4終了時のΘとd_iをそれぞれΘ_M，d_{Mi}とし，

$$\alpha = \arg \min_{g(\Theta, \mathbf{x}_i)} \{d_{Mi} | d_{Mi} \geq 1/n\}$$

とする．

図1: EDDRLのアルゴリズム

5 数値実験

本研究では，以下の信頼性測度を用いた場合について数値実験を行った．

$$g(\Theta, \mathbf{x}) = g(\mathbf{w}, \mathbf{x}) = \langle \mathbf{w}, \Phi(\mathbf{x}) \rangle \tag{6}$$

ここで，Φはガウシアンカーネルに対応した非線形写像である．数値実験結果については当日報告する．

参考文献

1. B. Schölkopf and J.C. Platt and J. Shawe-Taylor and A.J. Smola and R.C. Williamson. Estimating the Support of a High Dimensional Distribution. *Neural Computation* , Vol. 13, No. 7, pp. 1443-1471, 2001.
2. D.M.J. Tax and R.P.W. Duin. Support Vector Domain Description. *Pattern Recognition Letters*, Vol. 20, pp. 1191-1199, 1999.