Support Vector Machine におけるルールの利用

02502684　京都大学　*福永 拓郎　FUKUNAGA Takuro
01001374　京都大学　茨木 俊秀　IBARAKI Toshihide

1　はじめに

Support Vector Machine (以下 SVM) は学習理論，パターン認識などの分野で優れた性能を持つことで知られ，注目を集めている．本研究ではSVMで扱うデータに対して，あらかじめ得られたそのデータに関する知識を利用して前処理を施し，その学習過程に用いるデータ量を減らす方法について提案する．ただし，知識はルールの形式で表されているものとした．

2　Support Vector Machine

次のような問題を考える．

m 個のデータ $\boldsymbol{x_1}, \cdots, \boldsymbol{x_m} \in \boldsymbol{R}^n$ から成るデータ集合 $\boldsymbol{S}$ を考える．ただし，各データは n 次元の実ベクトルの形式で表されており，正クラスもしくは負クラスと呼ばれるクラスのうちのどちらかに属している．正クラスに属するデータを正データ，負クラスに属するデータを負データと呼び，それぞれのデータの集合を $\boldsymbol{P}, \boldsymbol{N}$ と表す．ラベル $y_i \in \{\pm 1\}$ はデータ $\boldsymbol{x_i}$ がどちらのクラスに属するかを示しており，$+1$ をとるならば正クラス，-1 をとるならば負クラスに属す．

SVMは以上のようなデータ集合 S から汎化能力の高い識別関数 $f : \boldsymbol{R}^n \to \boldsymbol{R}$ を求める．汎化能力とは，まだ観測されていない未知のデータ $\boldsymbol{x}$ を入力としたとき，$\boldsymbol{x}$ の正しいラベル y を出力として返す能力のことである．このとき，$\boldsymbol{S}$ を訓練データ集合と呼び，$\boldsymbol{S}$ から適切な識別関数を求めることを学習と呼ぶ．SVMの識別関数 f は次の式で表される．

$$f(\boldsymbol{x}) = sgn\left[\langle \boldsymbol{w}, \boldsymbol{x} \rangle - h\right] \tag{1}$$

ただし，$\boldsymbol{w} \in \boldsymbol{R}^n$ である．

SVMの学習は次の最適化問題に定式化される．

$$\begin{array}{ll} \min\limits_{\boldsymbol{w},h} & \|\boldsymbol{w}\|^2 \\ s.t. & y_i(\langle \boldsymbol{w}, \boldsymbol{x_i} \rangle - h) \geq 1 \quad (i = 1, \cdots, n) \end{array} \tag{2}$$

この際，最適解において i 番目の不等式制約の等号が成り立つとき，データ $\boldsymbol{x_i}$ のことをサポートベクトルと呼ぶ．これは訓練データのうち，識別関数の表す線形超平面との距離が最も近いデータのことである．

与えられた訓練データ集合によっては，線形超平面によって正例集合と負例集合を分離できる場合とできない場合がある．分離できる場合のことを線形分離可能，できない場合のことを線形分離不可能と呼ぶが，上記のSVMでは線形分離不可能な場合，正しくクラスを識別することができない．そのような場合，非線形な曲面でデータ空間を分離するように拡張された非線形SVMが用いられる．非線形SVMにおいては，学習に用いる最適化問題は次のようになる．

$$\begin{array}{ll} \max & \sum\limits_{i=1}^{n} \alpha_i - \frac{1}{2} \sum\limits_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x_i}, \boldsymbol{x_j}) \\ s.t. & \alpha_i \geq 0 \quad (i = 1, \cdots, n) \\ & \sum\limits_{i=1}^{n} \alpha_i y_i = 0 \end{array}$$

これは，最適化問題 (2) の双対問題において，ベクトルの内積 $\langle \boldsymbol{x_i}, \boldsymbol{x_j} \rangle$ をカーネル関数 $K(\boldsymbol{x_i}, \boldsymbol{x_j})$ で置き換えたものである．カーネル関数としては様々な関数が存在するが，一つの例としては次の関数が挙げられる．

$$K(\boldsymbol{x}, \boldsymbol{y}) \quad = \quad e^{-\rho \|\boldsymbol{x} - \boldsymbol{y}\|^2} \tag{3}$$

また，非線形SVMでは識別関数 (1) は次のようになる．

$$f(\boldsymbol{x}) = sgn[\sum_{i \in \boldsymbol{I}} \alpha_i y_i K(\boldsymbol{x_i}, \boldsymbol{x}) - h]$$

ただし，$\boldsymbol{I}$ はサポートベクトルの添字集合である．

3　訓練データ集合の削減

あるデータ集合 $\boldsymbol{S}$ に関する知識を抽出することによって，$\boldsymbol{S}$ の前処理を行う手法について提案する．この目的の一つは，訓練データ集合のサイズの削減である．知識を表現するには様々な形式が考えられるが，今回はルール形式によって表現された知識を用いることを考える．

ルールとは条件部 π と結論部 ω からなる知識のことであり，$r = (\pi, \omega)$ で表される．これは，「条件部 π の示す条件を満たすデータは結論部 ω の指すクラスに属する」ということを意味している．

条件部 π は，「(属性 i の値) $\geq c$」という条件式や，もしくは「(属性 i の値) $< c$」というような条件式の論理積で表される．つまり，データ空間においては座標軸

に垂直ないくつかの超平面で囲まれた領域を指している．また結論部 $\omega \in \{\pm 1\}$ は正もしくは負のクラスを指す．

前処理はルールの条件部の領域に存在する訓練データを削除し，その代わりにそのルールを訓練データ集合に組み込むことによって，訓練データ集合のサイズを減らす．ルールを組み込む方法については 3.1 と 3.2 において述べる．

3.1 代表点による方法

SVM では次のような性質が成り立つ．

性質 1 訓練データ集合 $\boldsymbol{S}$ に従って定まる識別関数は，サポートベクトルのみによって決定される．

サポートベクトルはクラスの境界付近に存在する．非線形 SVM の場合でも，カーネル関数として (3) を用いると，適当な条件のもとで境界付近に存在することがいえる．

そこで，ルールの条件部が示す領域を，その境界近くに置かれた代表点によって指し示すことにする．代表点の取り方としては次の方法を考えた．

データ選択法 ルールの条件部と結論部双方を満たすデータの集合から，条件部の示す領域のそれぞれの境界に対して最も近い $\boldsymbol{S}$の データ t 個を代表点として選ぶ．

ただし，t はパラメータである．

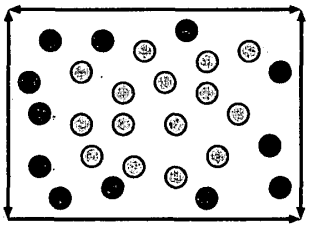

図 1: データ選択法

3.2 区間表現による方法

代表点による方法では，ルールの示す領域を正確に表したとはいいがたい．そこで，属性値の区間

$$(\text{属性 } i) = l : u$$

を用いたカーネル関数を定義する．ここで，$l, u \in \boldsymbol{R}$ であって，「属性 i は [l,u] の区間を示す」ことを意味している．このようなデータを SVM で扱うことは，カーネル関数 (3) における距離の定義を次のようにすることで可能になる．

$$\|\boldsymbol{x} - \boldsymbol{y}\|^2 = \sum_{i=1}^{n} d(x_{(i)}, y_{(i)})$$

ただし，

$$d(x_{(i)}, y_{(i)}) = \begin{cases} (x_{(i)} - y_{(i)})^2 & x_{(i)} \in \boldsymbol{R} \text{ のとき} \\ (y_{(i)} - u)^2 & y_{(i)} > u \text{ のとき.} \\ 0 & u \geq y_{(i)} \geq l \text{ のとき.} \\ (l - y_{(i)})^2 & l > y_{(i)} \text{ のとき.} \end{cases}$$

である．$x_{(i)}, y_{(i)}$ はそれぞれベクトル $\boldsymbol{x}, \boldsymbol{y}$ の i 番目の属性値である．$x_{(i)}$ が区間ではなく実数値の場合は，$l = u$ であるとみなせばよい．

削減方法	サイズ	認識率 (%)
削減前	674.6	94.87
データ選択法 (t=3)	113.6	93.4
区間表現	14	95.3

表 1: BCW に関する削減結果

4 実験結果

実験として，提案した訓練データ集合の削減手法を用いて実際に削減を試みた．表 1 は，ベンチマークとして広く用いられているデータ集合 BCW に対する削減結果である．ただし，ルールは決定木生成アルゴリズムである C4.5 を利用して生成した．認識率を下げることなく，サイズを削減することに成功している．

5 おわりに

本研究では SVM において，訓練データ集合に関する知識を利用し，そのデータ集合のサイズを削減する手法を提案した．このような知識の活用は削減だけでなく，性能の向上などの目的に用いることも考えられ重要である．その際にも，本研究で提案した手法を応用することが可能である．

参考文献

[1] B.Scholkopf, P.Simard, A. Smola, and V.Vapnik. "Prior knowledge in support vector kernels", In M.Jordan, M.Kearns, and S.Solla,editors, Advances in Neural Information Processing Systems 10, pages 640-646, Cambridge, MA, 1998.

[2] G. M. Fung, O. L. Mangasarian and J. W. Shavlik, "Knowledge-Based Nonlinear Classifiers",Data Mining Institute Technical Report 3-02, March 2003.