Support Vector Machine における変数選択基準

九州大学大学院数理学府 江田 智尊，九州大学マス・フォア・インダストリ研究所 西井 龍映

1 はじめに

Support Vector Machine (SVM) は高性能として知られる機械学習の判別手法である．本報告では SVM における高速な変数選択手法を提案する．まず判別性能の良さを表す目的関数を定義し，次に各変数がその目的関数に与える影響を評価して変数の重要度を測る．これにより変数選択を行う．そして判別精度の改善，および計算コストの削減を実現する．

2 変数選択基準の提案

$\{(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \{\pm 1\} \mid i = 1, \ldots, n\}$ を訓練データとする．説明変数 $\boldsymbol{x}$ を写像 $\boldsymbol{\phi}(\boldsymbol{x})$ により特徴空間 $\mathbb{R}^p$ に写し，特徴空間における線形判別を実現するための解 $(\boldsymbol{w}, b)$ を求める．次に判別関数 $f(\boldsymbol{x}_i) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i) + b$ の正負によってサンプル $\boldsymbol{x}_i$ をクラス ± 1 に分類する．そのため正しく判別されたサンプルは $y_i f(\boldsymbol{x}_i) > 0$ を満たすことになる．そこで最大化すべき目的関数を次で定義する．

$$\Delta(f) = \sum_{\boldsymbol{x}_i \in \mathrm{SV}} y_i f(\boldsymbol{x}_i) = \sum_{\boldsymbol{x}_i \in \mathrm{SV}} y_i \left(\sum_{\boldsymbol{x}_j \in \mathrm{SV}} a_j y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) + b \right)$$

ただし，$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{\phi}(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_j)$ はカーネル関数と呼ばれる特徴空間での内積，SV はサポートベクターの集合である．この目的関数 $\Delta(f)$ に個々の変数が与える影響を 2 通りの方法で評価する．

変数削除法： 変数 x_k $(k = 1, \ldots, d)$ を説明変数ベクトルから削除することで得られる次の式

$$\Delta(f)^{(-k)} = \sum_{\boldsymbol{x}_i \in \mathrm{SV}} y_i \left\{ \sum_{\boldsymbol{x}_j \in \mathrm{SV}} a_j y_j K\left(\boldsymbol{x}_i^{(-k)}, \boldsymbol{x}_j^{(-k)}\right) + b \right\}$$

によって変数 x_k が目的関数に与える影響を評価する方法である．

変数微分法： 変数 x_k の微小変動が目的関数に及ぼす影響を次で測る．

$$\nabla_k \Delta(f) = \sum_{\boldsymbol{x}_i \in \mathrm{SV}} y_i \sum_{\boldsymbol{x}_j \in \mathrm{SV}} a_j y_j \frac{\partial K(\boldsymbol{x}, \boldsymbol{x}_j)}{\partial x_k}\Big|_{\boldsymbol{x}=\boldsymbol{x}_i}$$

一旦訓練データで判別関数を求めると，両手法とも各説明変数の重要度を高速に評価することができる．

3 ベンチマークデータによる提案手法の評価

本手法を変数選択問題の 4 つの人工ベンチマークデータに適用した．これらのデータは各説明変数がクラスの情報を持っているかどうかが既知である．比較として同じ wrapper タイプの変数選択基準 [1][2] も併せて適用した．結果的に我々が提案した $\Delta(f)^{(-k)}$ を用いた基準のみが，クラスの情報を持つ説明変数を完全に選択することに成功した．詳細な結果と実データへの適用例に関しては当日報告する．

参考文献

[1] Byvatov, E. and Schneider, G. (2004). SVM-based feature selection for characterization of focused compound collections. *Journal of Chemical Information and Modeling*, 44(3):993-999.

[2] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:1357-1370.