

Support Vector Machine と RBF ネットワークによる追加学習

甲南大学 * 服部 篤史 (HATTORI Atsushi)

甲南大学 中山 弘隆 (NAKAYAMA Hiroataka)

要旨

Support Vector Machine(SVM) は、現在最も強力なパターン分類手法の1つとして注目されている。本論文では、SVM を用いてパターン分類を行い、追加学習の有効性を示し、RBF ネットワークとの比較、検討を行った。応用問題として、状況が次々に変化する株式売買問題に適用した。追加学習を行うことによって、環境が変化した時にも、柔軟な対応ができると考えられる。

1 Support Vector Machine の概要

SVM とは Vapnik らが 1992 年に提唱したもので、マージンを最大にする分離超平面を求めることにより最適な超平面を求める手法であり、次のような 2 次計画問題を解くことによって SVM によるパターン分類が実現できる。主問題として次のように定式化できる。

$$(SVM_p) \quad \begin{aligned} \text{目的} &: \frac{1}{2} w^T w \rightarrow \text{Min} \\ \text{制約} &: y_i (w^T \phi(x_i) + b) \geq 1, (i = 1, \dots, n) \end{aligned} \quad (1)$$

次に、(1) の問題の双対問題を考える。(1) の制約に対するラグランジュ乗数を α_i とすると、

$$(SVM_d) \quad \begin{aligned} \text{目的} &: \sum_{i=1}^p \alpha_i - \frac{1}{2} \sum_{i,j=1}^p \alpha_i \alpha_j y_i y_j \phi(x_i)^T \phi(x_j) \rightarrow \text{Max} \\ \text{制約} &: \alpha_i \geq 0, (i = 1, \dots, p) \\ & \sum_{i=1}^p \alpha_i y_i = 0 \end{aligned} \quad (2)$$

となる。ここで、双対問題での目的関数の中に高次元に変換したベクトルの内積 $\phi(x_i)^T \phi(x_j)$ が出てくるので、莫大な計算回数を要する。そこで、Mercer の定理により、 $K(x, x') = \phi(x)^T \phi(x')$ となるようなカーネル関数が存在することが証明されている。本論文ではカーネル関数として $K(x, x') = \exp(-\frac{\|x - x'\|^2}{r^2})$ のようなガウス関数を用いた。このようなカーネル関数を導入することにより、ベクトルの内積を計算することなく最適化を行うことができ、これをカーネルトリックと呼ぶ。ここで、ガウス関数の半径は $r = d_{max} / \sqrt{nm}$ によって決定され、 d_{max} はデータとデータとの最大距離、 n はデータ数、 m はデータの次元を表す。

2 追加学習

人間は、日々の生活で新しい体験をすると知識として吸収し、以前に比べて、より適切な行動をとることが出来るようになる。この追加学習という行為がより汎化能力を向上させているといえる。SVM にもこの概念を導入すると、より良い結果が得られると考えられる。以下にその方法として、受動的追加学習法と選択的追加学習法と呼ばれる 2 つの方法を説明する。

まず、受動的追加学習法とは、データの重要度を考えずにデータを追加する方法で、判別を行う際に毎回データを1つずつ追加していき、最終的には全てのデータを追加する。しかし、この方法では全てのデータを追加してしまうので、過学習をする恐れがある。

それに対して、選択的追加学習法とは、データの重要度を考慮して選択的にデータを追加する。つまり、追加データがどの程度重要であるかを判断し、重要だと判断したデータのみ追加し、重要でないと判断したデータは追加しないことにより判別局面を変化させる方法である。具体的な方法として、追加データを x_t とおくと、追加データ x_t に対する判別結果が不正解ならば、追加データ x_t を追加する。また、判別結果が正解、かつ、出力値の絶対値が閾値 θ 以下ならば、追加データ x_t を追加する。判別結果が正解ならば、無条件で追加しないという訳ではない。

3 単銘柄株式問題

SVM および RBF ネットワークを用いて変動する株価の予測を行った。今回は、ある企業の1985年1月から1994年11月までの119期間の月間株式データを用いた単銘柄株式問題について取り上げる。入力データの要素として7つの指標を用いて、“買い”か“買いでない”の2値分類を判別する。教師値として、“買い”を“+1”、“買いでない”を“-1”とした。初期学習データを1985年1月から1989年2月までの50期とし、テストデータを1989年3月から1994年11月までの69期間として実験を行った。

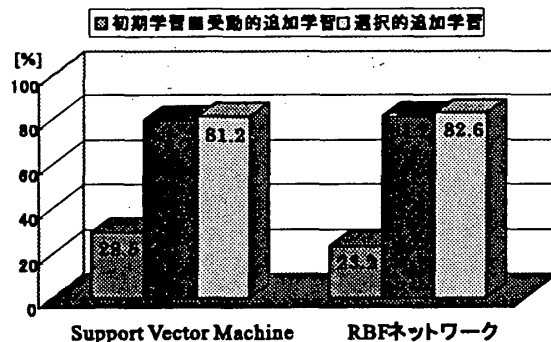


図1: 正解率の比較

図1はそれぞれの手法に対する正解率を表している。受動的追加学習法、さらに、選択的追加学習法を導入することにより、多少ではあるが良い結果を得ることができた。

4 おわりに

SVM と RBF ネットワークによって追加学習を行い、それぞれの結果を比較したが、両手法とも結果に大きな違いはなかった。むしろ、計算時間を考慮すると、RBF ネットワークの方が優れているとも言えた。今後の課題として、判別に邪魔をしているデータの影響力を積極的に弱める、能動忘却を取り入れるとさらに良い結果が得られると期待できる。

参考文献

- [1] Nello Cristianini, John Shawe-Taylor : An Introduction to Support Vector Machines and other Kernel-based learning method, Cambridge, 2000
- [2] S. Haykin, Neural Networks: A Comprehensive Foundation, Macmillan College Publishing Company, 1994.