修士論文　要旨

信用格付に対するサポートベクターマシンの応用

076226　星野 禎
指導教員　藤井　眞理子　教授

2009 年 2 月 3 日

1　はじめに

SVM（サポートベクターマシン，Support Vector Machine）[1] は判別分析手法の 1 つであり，様々な分野に応用され，近年でも盛んに研究がなされている．SVM は学習データに外れ値があると判別能力が低くなることが知られている [3] が，Lin and Wang による Fuzzy-SVM[3] は，データの取扱いに柔軟さを持たせ，外れ値の影響を小さくすることにより，判別能力の向上が可能であることを示した．また，Li and Lin [2] は，順序構造があるクラス分類問題を通常の 2 クラス分類に帰着させて解く手法（順序付け SVM）を提案している．

本研究では，これらの SVM 手法の応用として格付の分析を考え，実験を通してそれぞれの手法の比較を行った．また，Fuzzy-SVM と順序付け SVM を組み合わせた方法を新たに提案し，格付を判別する事例を通して，その特徴をみた．

2　サポートベクターマシン

d 次元実数データ $\boldsymbol{x}$ が与えられた時，適当な非線形変換 $\Phi(\cdot) : \mathbb{R}^d \to \mathbb{R}^l$ を用い，判別関数 $g(\boldsymbol{x}) = \boldsymbol{w}^\top \Phi(\boldsymbol{x}) + b$ の正負により 2 クラスへの分類をすることを考える．2 クラスの判別を行うソフトマージン付き非線形 SVM は，学習データ $(y_i, \boldsymbol{x}_i), y_i \in \{-1, 1\}$ $(i = 1, 2, \cdots, N)$ が与えられた時に次の最適化問題を解くことにより判別関数を得る手法である:

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{n} \xi_i, \tag{1}$$

$$\text{s.t.} \quad y_i(\boldsymbol{w}^\top \Phi(\boldsymbol{x}_i) + b) - 1 + \xi_i \geq 0, \quad \xi_i \geq 0 \quad (i = 1, 2, \cdots, n). \tag{2}$$

式 (1) の 1 項目の最小化はマージンの最大化に対応する．マージンは線形分離可能な場合判別面と面に一番近い点との距離に対応する量であり，マージンが大きいほど判別能力が良くなるとされる．2 項目は誤判別に対応するコストであり，その定数 C は実験的に決定される．実験において非線形変換はガウスカーネルを用いた．

Fuzzy-SVM[3] は，SVM の誤判別コストにデータに応じた重要度を加味することにより高い判別能力が得られると考える．具体的にはデータ i の重要度を $m_i \in (0, 1]$ とし，

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{n} m_i\xi_i \tag{3}$$

を最小化する．この定式化により，外れ値などの重要度が小さくなるように定めることで，その影響が小さくなると期待できる．

2 クラス判別 SVM を用いて K クラスへの判別を行う方法として広く使われている方法が one-against-one（OAO）法である．これは全てのクラスの組み合わせに対応して $K(K-1)/2$ 個の SVM を作成し，判別をする際にはそれらの「投票」によって結果を得る方法である．

Li and Lin[2] はクラス間に順序構造がある場合に適した順序付け SVM を考案した．この手法は，判別関数 g とその閾値 θ_k $(k = 1, 2, \cdots, K-1)$ を同時に学習するものである．従って $g(\boldsymbol{x})$（つまり $\boldsymbol{w}$ と b）および $\boldsymbol{\theta}$ を動かす．この手法の学習は，3 クラス判別の場合次のように説明される: データ $(y_i, \boldsymbol{x}_i)$ に対応して，新たにデータ $(y_i^1, (g(\boldsymbol{x}_i), 0)), (y_i^2, (g(\boldsymbol{x}_i), \eta))$ $(y_i^k = 2 \cdot \mathbf{1}_{\{k<y_i\}} - 1, \eta$ は適当な定数，$\mathbf{1}_{\{\cdot\}}$ は指示関数) を作り，これらのデータを 2 クラス判別 SVM を用いて解いたときのマージンを最大化する．ソフトマージンも考えた時，これは次の最適化問題を解くことに帰着される:

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{\theta}} \frac{1}{2}||\boldsymbol{w}||^2 + \frac{1}{2\eta^2}||\boldsymbol{\theta}||^2 + C\sum_{i=1}^{N}\sum_{k=1}^{K-1} \xi_i^k, \tag{4}$$

$$\text{s.t.} \quad y_i^k(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - \theta_k \geq 1 - \xi_i^k, \quad \xi_i^k \geq 0 \quad (i = 1, 2, \cdots, N, \text{ and } k = 1, 2, \cdots, K-1).$$

この場合，解は $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_{K-1}$ となる [2]．実験では [2] に従って $\eta = 1$ とした．

本研究では Lin and Wang[3] の考えを組み入れた順序付け Fuzzy-SVM を提案した．この手法は，順序構造に合うように Fuzzy-SVM の重要度を定めることで，外れ値の影響が小さくなることを期待するものである．式 (4) の最小化項を，

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{\theta}} \frac{1}{2}||\boldsymbol{w}||^2 + \frac{1}{2\eta^2}||\boldsymbol{\theta}||^2 + C\sum_{i=1}^{N}\sum_{k=1}^{K-1} m_i^k\xi_i^k \tag{5}$$

とし，重要度 m_i^k は通常の順序付け SVM の判別関数の値を用いる．従って，まず通常の順序付け SVM を作成し，その後に順序付け Fuzzy-SVM を作成する．重要度 m_i^k の定め方は，通常の順序付け SVM を作成した際のランク k より大きいか否かという判別関数，$g^k(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b - \theta_k$ を用い，$(0, 1]$ にスケーリングする単調非減少な関数 $s(\cdot)$ を使って，正のデータに関しては $m_i^k = s(g^k(\boldsymbol{x}_i))$，負のデータに関しては $m_i^k = 1 - s(g^k(\boldsymbol{x}_i))$ とする．概念図は図 2.1 のようになる．

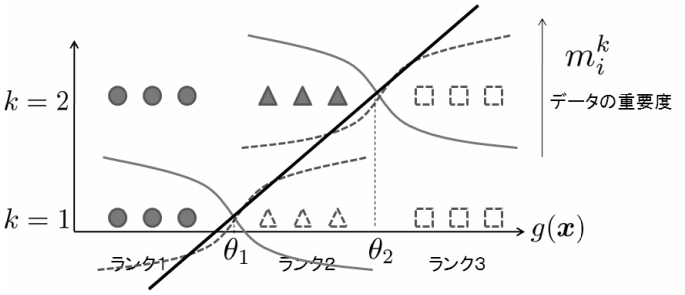

図:2.1 順序付け SVM の重要度

実験ではロジスティック関数 $L(x) = \frac{\exp(x)}{1+\exp(x)}$ を用いてスケーリングを行った．

3　SVM による格付の判別

格付は財務データを反映していると考え，財務変数による格付の判別分析を考えた．格付データは R&I 社による発行体格付を使用した．財務変数は他の信用リスク評価モデルなどを参考に代表的なものを 2 つまたは 6 つ用いることにし，それぞれ Box-Cox 変換を行った後に $[-1, 1]$ に収まるように線形変換を行った．

実験は以下を 100 回行い，平均と分散を見た．

1. データを学習データとテストデータに個数比が 3:1 になるようランダムに分割する．

2. できた学習データで Cross-Validation+Grid-Search を行い,「最適な」 (C, γ) を得る．

3. 得られた (C, γ) を用い，学習データで学習し，テストデータを判別し，指標を計算する．

指標は正答率と絶対コストを用いた．これらはデータ i の判別結果を $\hat{y}_i$ とした時，正答率は $\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\{y_i \neq \hat{y}_i\}}$，絶対コストは $\frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|$ により計算される．

4　格付判別分析結果と考察

OAO 法と順序付け SVM について比較を行った．98 年度のデータを，2 変数を用い，6 段階の判別を行なった場合，判別面は図 5.1, 5.2 のようになった．6 変数を用いた場合の実験結果が表 5.1 である．なお，括弧内の数字はそれぞれの分散である．順序付け SVM の方が滑らかな判別面を描き，また判別能力が高いことがわかる．

順序付け SVM と順序付け Fuzzy-SVM の比較を行った．輸送機器業種は自動車産業とその部品会社などであり，その企業の人気や企業戦略，取引先の業績も格付に大きく影響するなどの特徴を持つため，財務変数のみでは判別のしにくい業種である．2 変数を用いて 6 段階の判別の場合の判別面は図 5.3, 5.4 のようになり，結果をまとめたものが表 5.2 である．順序付け Fuzzy-SVM が解釈のしやすい判別面を得ており，結果も良いことが見て取れる．

5　まとめと今後の課題

格付判別問題に対しては，順序構造を考慮した順序付け SVM が高い判別能力を持つことがわかった．また，提案手法である順序付け Fuzzy-SVM も有効である場合があった．このように状況に応じた手法を選択することが重要である．今後の課題としては Fuzzy 化手法の理論的考察や，格付問題を考えた時の過去データの取り扱い方をどうするか，などがある．

参考文献

[1] N. Cristianini and J. Shawe-Taylor. サポートベクターマシン入門. 大北剛訳, 共立出版, 東京, 2005.

[2] L. Li and H. Lin. Ordinal Regression by Extended Binary Classification. *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, Vol. 19, p. 865, 2007.

[3] C. Lin and S. Wang. Fuzzy support vector machines. *IEEE Trans Neural Netw*, Vol. 13, No. 2, pp. 464–71, 2002.

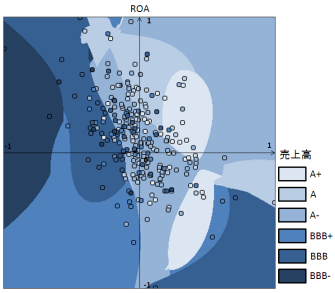

図 5.1: OAO 法

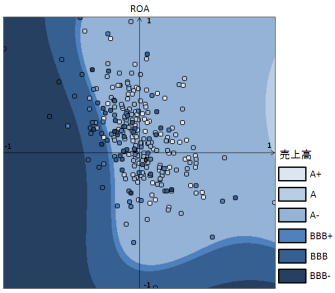

図 5.2: 順序付け SVM

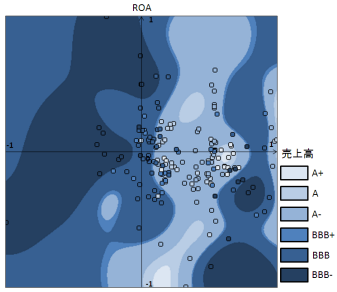

図 5.3: Non-Fuzzy

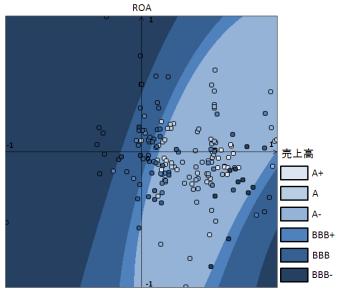

図 5.4: Fuzzy

表 5.1: OAO 法と順序付け SVM の比較

	OAO 法	順序付け
正答率	36.8%(0.2%)	39.8%(0.3%)
絶対コスト	0.932(0.009)	0.812(0.010)

表 5.2: Non-Fuzzy と Fuzzy の比較

	Non-Fuzzy	Fuzzy
正答率	38.5%(0.4%)	44.4%(0.5%)
絶対コスト	0.839(0.004)	0.786(0.041)