マルチクラスサポートベクターマシンを用いた信用格付け
Credit rating using multi-class support vector machine

経営システム工学専攻　山形　拓

1　はじめに

今日, 私たちの身の回りにおいて様々なものが格付けの対象となっており, 格付けを参考にして物事を見る事も多いと思われる. 格付けには様々な種類があるが, 本研究で取り上げたのは信用格付けと呼ばれる格付けである. 企業の信用格付けは, 社債投資における投資指標として資本市場において重要な役割を担っている. 各格付け機関はそれぞれの格付けメカニズムを持って, 格付けを決定しているが, その格付けメカニズム自体は世間に公表されていない. このように, 公表されていない格付けの決定メカニズムの中身に対して, そのメカニズムによって出された格付けと同等の精度, あるいはそれに近い精度を判別する手法を見いだすことができれば有益であると考え, 今回の研究を行った.

信用格付けに関する研究として, これまでに多くの研究が行われている. 例えば格付けデータを用いる手法がある. これは, 将来のデフォルト率及び格付推移確率などを格付けデータによって推定する方法である. その他にも財務データを用いる方法やニューラル・コンピューティング, マクロファクターを用いる手法などが研究されてきている. 本研究では, これら格付に関する研究において用いられてきた方法の中から, サポートベクターマシン （SVM）と総称される方法に注目し, 研究を行った.

本研究の目的としては, 既存の複数クラス SVM で出した信用格付けの精度と順序付け複数クラス SVM において出した信用格付けの精度, さらに, 改良方法を用いた順序付け複数クラス SVM の精度を比較し, 順序付け複数クラス SVM を用いる事で, より良い精度を出す事ができているかの評価を行うことである.

2　サポートベクターマシン（SVM）

本研究では, サポートベクターマシン（SVM）と総称される方法に注目し, 研究を行った.SVM とは入力ベクトルを非線形に写像した高次元の特徴空間上で,1 つの線形識別関数を構成する方法である.SVM でのクラス分類の目的は, 高次元特徴空間において,2 値問題について, 正しく分類するような超平面を, 計算量的に効率良く学習するような方法を提供する事である. 高次元特徴空間において写像された分離超平面は, 元の入力空間では局面になって, 最終的に非線形な識別関数を構成する. また,SVM は数多くある現在のパターン認識手法の中でも, 最も性能が優秀な学習モデルの一つとして知られている.SVM のモデルとしては, 線形 SVM 手法やソフトマージンの手法などが知られている. しかし,SVM は 2 クラス識別問題において基本的には定式化されているので, 格付け問題など, 多クラスの問題を扱うような識別器を構成するには, 多数の 2 クラス SVM を組み合せる事などが必要となる.

2.1　複数クラス SVM

複数クラス分類とは, 基本的に各データが予め用意されたクラスのどれかに分類をされることを言う. そのような中で複数クラス SVM において,2 値分類問題に限ると,SVM のようなマージン制御に従った高い汎化性能を持つような分類器を用いる事が出来る. 前述したように, 基本的に多クラスの識別問題を 2 値分類器で扱うためには,2 クラスの判別モデルを組み合せる事になる. 代表的な手法として, 任意のクラスとその他全てのクラスで構成された識別関数を利用し, 2 値クラス SVM を複数クラス問題に対応させる手法である One-Against-All 手法（OVA）や任意のクラスのペアどうしでの学習データを用いて識別関数を構成し, それを多クラス分類問題に利用している One-Against-One 手法（OVO）などがある.

2.2　順序付け複数クラス SVM

本研究では複数クラス SVM を用いてより精度の高い信用格付けを決定していく事を目的としている. そこで, 複数クラス SVM に対して順序を考慮する事で, 格付けの精度を上げる事を考えた. 理由として, ここまでの手法では, クラスがカテゴリーごとの複数クラス分類問題のために設計されており, 順序といったものはなかった. それに対して信用格付けにおいては, 各格付けは AAA,AA,$\cdots$,C といったように順番がつけられている. そのような理由から, 順序を考慮に入れた複数クラス SVM 技術の適切な修正は, 信用格付けのための識別器の性能を向上させるかもしれないと考えたためである. そこで,1 つの複数クラス SVM の手法として, 順序付き複数クラス SVM を考え, 研究を行った.

本研究で参考にした手法として,Kyoung-jae Kim,Hyunchul Ahn が発表した順序付け複数クラス SVM の手法がある. この手法におけるアプローチでは, 分類方法（One-against-The-Next,One-against-Followers）, 組み合せ方法（forward method,backrward method）の 2 つの方法について考える必要がある. またデータの分類方法と組み合せ方法によって,4 つのタイプの順序付き複数クラス SVM の組み合せパターンが考えられる. 以下の図において実際にどのように各クラスに順位がついていくのか,2 つの例について説明している.（図 1,2）

図において, 赤色の数字が決定するクラスを表している. また, 赤い線が分類器を指している.

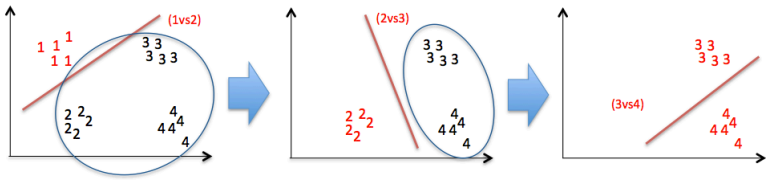

図 1: One-Against-The-Next+forward method

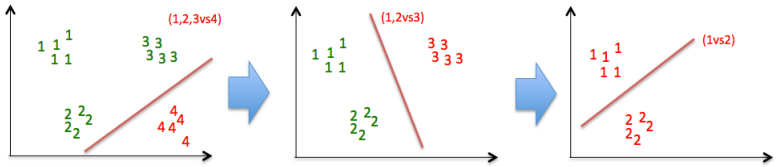

図 2: One-Against-Followers+backward method

図 1 にあるように, 分類方法として One-against-The-Next を用いると,OVO に似たアプローチを取っている事が分かる. 任意のクラスのペアどうしでの学習データを用いて識別関数を構成しているが, 順序が決定したクラスを次の計算時にデータから除いている. また, 組み合せ方法として forward method を用いると, 最高クラス（格付けで言う AAA）から順序を決定していく事が分かる. 同じように図 2 では分類方法として One-against-Followers を用いているが,OVA に似たアプローチを取っており, 任意のクラスとその他全てのクラスで構成された識別関数を用いている事が分かる. 加えて, 決定したクラスを次の計算時にデータから除いている. また, 組み合せ方法として backward method を用いると, 最低クラス（格付けで言う C）から順序を決定している.

2.3　順序付け複数クラス SVM の改良

上記で紹介した順序付け複数クラス SVM を改良する事で, 従来の手法よりも, より精度の高い信用格付けを行う事を試みた. 改良方法として, 本研究では以下の 4 つのアプローチを順序付け複数クラス SVM に適用し, 計算結果を出している.

1. カーネルを用いて, 非線形分類を考慮する方法. 本研究で使用したカーネルは, 多項式カーネル, ガウスカーネル, シグモイドカーネルの三つのカーネルである.

2. 順序の付け方を変化させる方法. 順序付け複数クラス SVM の手法について, 順序のつけ方は最高クラス (AAA) から決定していく場合か, 最低クラス (C) から決定していく場合かの 2 通りだった. そこで, まず始めにその中間のクラス同士での判別を行う事で, 上位クラスのグループと下位クラスのグループを判別し, その後, その 2 グループ（上位クラス, 下位クラス）の中で各クラスを決定していく事で精度を挙げる事ができないか検証してみた. 実際の各クラスの決定方法は以下の図 3 のようになっている.

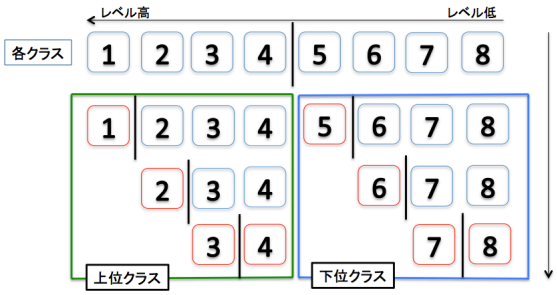

図 3: 各分類方法,forward method+

赤色の線で囲ってあるクラスが決定クラスとなっている. また, この手法については, 本研究における名前として, 分類方法, 組み合せ方法共に+をつけて表記する. したがって, 図 3 はそれぞれの分類方法において forward method+を選択した場合を表している.

3. 特徴選択を行う方法. 一般的にパターン認識で用いるデータベースは, 事例とそれを表す特徴量で表されており, 高次元ベクトルで構成されることが多い. 識別に有効な情報を得るためには, 特徴数を多く必要とするが, 逆に識別に関して意味を持たない不要な特徴により, 最良の識別を行うことができないこともある. そこで, 特徴選択を行うことで最良の特徴数を選択し, 識別精度を向上させる事を試みた. 本研究では SBS 法を用いた.

4.SMOTE アルゴリズムを用いてサンプリングを行う方法.SMOTE アルゴリズムとはサンプリング法の一つである. サンプリング法とは, 判別モデルを構築する前に, フィットデータに含まれるケースを増加もしくは減少させることで, データセットの偏りを補正する前処理を行うことである. 本研究で用いる順序付け複数クラス SVM の手法において, 正例と負例の学習データ間に偏りが生じてしまう事がある. これは, 不均衡データの問題であると考える事ができ, 各分類器の分類精度低下を招いている可能性がある. その問題に対して, 不均衡データの状態を解消する事で精度を上げる事ができないかと考え,SMOTE の適用を行った.

3 使用データ

本研究で使用したデータは以下の通りである. 信用格付データとして 2011 年 3 月末時点と 2012 年 3 月末時点の 2 時点における, 東証 1 部上場銘柄（金融除く）で格付情報投資センター（R&I）の格付が付与されている企業の財務データ及び社債格付データを用いた. それぞれ 341 社（2011 年）,338 社（2012 年）という社数となっていた. 本研究では, 2011 年（3 月）データで変数選択やパラメータの推定を行い, 2012 年（3 月）データをテストデータとして精度の検証を行った.

社債の格付の内訳について上記の 2 時点における R&I の格付けデータでは各社の格付けクラスが 14 に分かれていた. しかし, 数に偏りが大きいため, 実用性の観点から 6 つ のクラスに統合して実験を行った. また, 格付が付与されている企業の財務データに関しては, 企業の格付を説明する指標として, 5 カテゴ リー（収益性, 安全性, 効率性等, 規模,CF） 38 指標を使用して検証を行った.

4 結果

格付けデータ, 指標データに対して, 順序付け複数クラスサポートベクターマシンを用いた数値実験を行った. 本研究では, 既存の複数クラス SVM に手法として One-against-One の手法を採用している. また, 実験結果を出すに当たって交差検定を用いている. 交差検定とは, 学習に使うために集めたデータをいくつかに分割する方法である. なお, 今回の実験では 10 分割交差検定を用いている.

ここでは結果として, 事前データ, 事後データ別に通常の各複数クラス SVM とガウスカーネルを用いた各複数クラス SVM の中で, もっとも高い精度を出した順序付け複数クラス SVM における組み合せパターン（分類方法と組み合せ方法）の結果を以下表 1 に表した.「事前」と表記される上段のデータが 2011 年のデータを表しており,「事後」と表記されている下段データが 2012 年のデータを指している. 表中で表記されている正答率は完全に予測クラスと実際のクラスが同じだった場合の精度を表している.

表 1: 各複数クラス SVM における最高精度

事前 (2011 年)	ovo	omsvm (F-f)	順序変化 (F-f+)	特徴選択 (F-f)	smote (F-f)	ovo (ガウス)	omsvm (N-f)	順序変化 (N-f+)	特徴選択 (N-f)	smote (N-f)
正答率	62.6%	75.3%	74.4%	80.6%	76.5%	79.7%	85.0%	85.3%	85.3%	92.6%
順位	5	3	4	1	2	5	4	2	2	1
事後 (2012 年)	ovo	omsvm (F-f)	順序変化 (N-f+)	特徴選択 (F-f)	smote (F-f)	ovo (ガウス)	omsvm (N-f)	順序変化 (N-f+)	特徴選択 (N-f)	smote (F-f)
正答率	60.1%	62.7%	64.8%	64.5%	63.3%	66.6%	71.6%	73.4%	73.7%	73.4%
順位	5	4	1	2	3	5	4	2	1	2

表 1 の結果から, 組み合せ方法として forward method を選択する事で backward method を選択する場合よりも良い正答率になっている事が分かる. これは他のカーネルの場合も同様の傾向が見られた. この事から, 組み合せ方法としては forward method を選択する事が有効だと考えられる.（表 1 では通常の場合とガウスカーネルを用いた場合の例を取り上げたが, 他カーネルにおいても同様の実験を行っている.）

全体的な正答率を見ていくと, 各順序付け複数クラス SVM の方法は,OVO の方法よりも正答率が高くなっている事が分かった. さらに順序付け複数クラス SVM に対して改良方法を適用した方法の方が高い正答率につながっている事が分かる. また, カーネル別に見るとガウスカーネルを使用した場合, 他の全ての方法に比べよい結果につながっている事が分かった. 全ての方法の中で最も事後における正答率が高かったのは, 特徴選択を行った場合の順序付け複数クラス SVM の方法（ガウスカーネル）で 73.7%であった. これらのことから, 今回用いた順序付け複数クラス SVM の各方法は格付けの予測精度をいう面で複数クラス SVM の方法よりも好ましい方法だと考えられる。

主な参考文献

1. T.G. Ditterich,G. Bakiri. Solving multiclass learning problems via error-correcting output codes
2. Kyoung-jae Kim,Hyunchul Ahn,Combining Pairwise SVM Classifiers for Bond Rating
3. T. MARILL, D. M. GREEN,On the Effectiveness of Receptors in Recognition Systems
4. Nitesh V. Chawla , Kevin W. Bowyer,Lawrence O. Hall,W. Philip Kegelmeyer,SMOTE: Synthetic Minority Over-sampling Technique
5. Jun-ya Gotoh , Akiko Takeda ,Rei Yamamoto,Interaction between Financial Risk Measures and Machine Learning Methods