1999年度日本オペレーションズ・リサーチ学会
春季研究発表会

経済指標データの論理的解析とその回帰分析への適用について

02502134　京都大学　*橋口 浩隆　HASHIGUCHI Hirotaka
01704164　京都大学　柳浦 睦憲　YAGIURA Mutsunori
01001374　京都大学　茨木 俊秀　IBARAKI Toshihide

1　はじめに

様々な経済指標の過去のデータ系列を用いて, TOPIX(東証株価指数) の将来値を予測する問題を考える. 予測の基本的手法としては回帰分析を適用する. 回帰分析において各指標を説明変数として用いると, 目的変数と各説明変数間の (過去のデータ系列における) 線形関係式が導出される. 本研究では, 各指標間の論理的な関係を解析した上で, 変数の非線形結合を表す新たな説明変数を作成し, これらを回帰分析に用いることで予測精度の向上を目指す.

本研究では Tヶ月後の TOPIX の騰落の予測を目指す. 従って以下では与えられた過去のデータ系列中, 目的変数である Tヶ月後の TOPIX の騰落率が 0 以上となる系列 (正例) の集合を S^+ , 負となる系列 (負例) の集合を S^- とする.

2　データの論理的解析

2.1　有効な説明変数の抽出 —数値データの 2 値化の適用—

与えられた指標は数値データであるが, 論理的解析を行うため, まずはこれらのデータの 2 値化を考える. 数値データ対 (S^+, S^-) の属性 i において, カット点 α_i を導入し, データベクトル $\boldsymbol{u}^{(j)} \in S^+ \cup S^-$ における属性 i の値 $u_i^{(j)}$ に基づいて次のように $x_i^{(j)}$ を定義する;

$$x_i^{(j)} = \begin{cases} 1 & (u_i^{(j)} \geq \alpha_i) \\ 0 & (u_i^{(j)} < \alpha_i). \end{cases}$$

各属性 i のカット点 α_i の位置には, i に対し $S^+ \cup S^-$ 内の数値ベクトルがとる異なる値を

$$u_i^{(0)} < u_i^{(1)} < \cdots < u_i^{(K_i)}$$

と並べるとき, それぞれの隣接対 $u_i^{(s)}$ と $u_i^{(s+1)}$ の間の 1 点を考えることができる. $u_i^{(s)}$ と $u_i^{(s+1)}$ の間の 1 点には, 一般性を失うことなく

$$\alpha_i^{(s)} = (u_i^{(s)} + u_i^{(s+1)})/2, \quad s = 0, 1, \cdots, K_i - 1$$

とすることができる.

可能なカット点の集合から, 適当なカット点を選ぶと, 数値データ対 (S^+, S^-) を 2 値データ対 (T, F) に変換することができる. 全属性 $i = 1, \cdots, N$ において, 合わせて d 個のカットポイントを導入したとすると, (T, F) のベクトルの次元は d である. この (T, F) を部分定義論理関数 (partially defined Boolean function; pdBf) という. この時, T 内のベクトルに対しては 1 を, F 内のベクトルに対しては 0 を出力する (完全定義) 論理関数 (Boolean function) $f : \{0,1\}^d \to \{0,1\}$ を pdBf (T, F) の拡大と呼ぶ. このような拡大 f により, 任意の d 次元の 0-1 ベクトルを正例と負例のいずれかに区別することができる.

pdBf (T, F) が拡大 f を持つための必要十分条件は $T \cap F = \emptyset$ となることであり, 数値データ対 (S^+, S^-) が拡大を持つための必要十分条件は, 各属性 i において意味のある全てのカット点を用いて得られるマスター pdBf (T^*, F^*) が拡大を持つことである [1]. すなわち, 拡大の存在は 2 値データ (T, F) の完全な分離を意味し, またマスター pdBf (T^*, F^*) は数値データ対 (S^+, S^-) の持つ大小関係の情報を完全に有する.

ここで, マスター pdBf (T^*, F^*) が拡大を持つとする. このとき, (T^*, F^*) の全ての変数が (S^+, S^-) の拡大の存在に必要とは限らず, 適当に選ばれたカット点により得られる pdBf (T, F) が $T \cap F = \emptyset$ を満たせば (S^+, S^-) は完全に分離される. このように, 少ないカット点で (S^+, S^-) を分離できるならば, 変数とそのカット点の位置が (S^+, S^-) を分離する重要な情報を有していると考えられる.

そこで, (T, F) が拡大を持つための最小個数のカット点を求める問題が考えられるが, これは NP-困難であることが知られている [1]. しかし, 説明変数とし

て意味のある属性とカット点を知るには, 厳密な最小化はあまり意味を持たない. 本研究では, この問題を集合被覆問題に帰着した上で, 欲張り法を用いて好ましい属性とそのカット点を複数抽出する. このとき, 同じ属性に複数個のカット点が挿入されていたり, 統計的に相関の高い属性が繰り返し選ばれる可能性があるので, そのような場合, 類似の属性の近いカット点同士をより意味のあるカット点にまとめる, という調整の操作が必要となる. このようなカット点の取捨選択の方法の詳細については, 現在検討中の部分も含め, 当日報告する.

2.2 結合ルールの抽出

得られた pdBf (T, F) において, データの論理的解析であるパターン [1], あるいはデータマイニングの手法である結合ルール (association rule) [2] を適用し, 変数間の論理的な結合 (非線形結合) の中で結果に大きな影響を与えるものを抽出する.

ここで $R = \{X_1, X_2, \cdots, X_m\}$ を pdBf (T, F) における変数の集合とする. $T \cup F = \{\boldsymbol{t}_1, \cdots, \boldsymbol{t}_n\}$ ($n = |T| + |F|$) は変数集合 $R = \{X_1, X_2, \cdots, X_m\}$ に対する値を要素とする 0-1 ベクトルの集合である. ベクトルは正例 $(\in T)$ と負例 $(\in F)$ とに分けられる. W を変数の非線形結合とし (例えば $W = X_1\bar{X}_2$), データ $\boldsymbol{t}_i \in T \cup F$ をとる. $\boldsymbol{t}_i$ の各要素の 0, 1 の割り当てに対し, W が返す値を $t_i[W]$ とする. すると, 全てのデータ中 $t_i[W] = 1$ となるデータの割合 p_s , 及び $t_i[W] = 1$ となった場合に $t_i \in T$ である条件付き確率 p_c は, それぞれ

$$p_s = \frac{|\{i \,|\, t_i[W] = 1\}|}{n}$$

$$p_c = \frac{|\{i \,|\, t_i[W] = 1 \text{ and } t_i \in T\}|}{|\{i \,|\, t_i[W] = 1\}|}$$

となる. p_s, p_c がある一定値以上となるような W を求めることで, 正例と相関の高い非線形結合を求めることができる. なお, 負例の場合も同様に求めることができる. このような結合ルール及びパターンを求める方法は, 例えば [3, 4] などに紹介されているが, 基本的には W の候補を一つ一つ列挙していくことになるので, 変数集合, データ集合の大きさによってはかなり計算時間を必要とする.

3 回帰分析への適用

2.2 で得られた結合を元に, その論理的組合せの性質を有するような新たな説明変数を作成する. 例えば 2 つの属性 i, j において, それぞれのカットポイント α_i, α_j に対応する 2 値変数 X_i, X_j による結合ルール X_iX_j が得られた場合, 例えば次のような変数を新たに作成する.

1. $u_{ij} = \begin{cases} (u_i - \alpha_i) \cdot (u_j - \alpha_j) & (u_i \geq \alpha_i \text{ かつ } u_j \geq \alpha_j) \\ 0 & (\text{それ以外}) \end{cases}$

2. $u_{ij} = \begin{cases} (u_i - \alpha_i) + (u_j - \alpha_j) & (u_i \geq \alpha_i \text{ かつ } u_j \geq \alpha_j) \\ 0 & (\text{それ以外}) \end{cases}$

以上のようにして作成された変数を説明変数として, 重回帰式を求める. このようにすることによって, データの属性を直接説明変数として用い, それらの線形結合である回帰式を求める手法に比べ, データに対するより精度の高い解析が可能となることが期待できる.

4 数値実験

これまでの予備的な実験において, 目的変数との相関係数が, 生データの説明変数によるものよりも大きい説明変数をいくつか作成することができた. 詳細な実験結果は当日に報告する.

参考文献

[1] 茨木俊秀, データの論理的解析とブール関数, 離散構造とアルゴリズム 第 4 章, 近代科学社, 1998.

[2] R.Agrawal, T.Imielinski, and A.Swami, "Mining association rules between sets of items in large database," *Proceedings of the 1993 International Conference on Management of Data (SIGMOD 93)*, pp.207-216, May 1993.

[3] H.Mannila, H.Toivonen, and A.Inkeri Verkamo, "Efficient Algorithms for Discovering Association Rules," *AAAI Workshop on Knowledge Discovery in Database*, pp.181-192, July 1994.

[4] E.Mayoraz, "C++ Tools for Logical Analysis of Data," *RUCTOR(Rutgers University's Center for Operations Research) Technical Report*, July 1995.