視覚的データマイニングのための拡張自己組織化マップ*

4P−01
中村 雅之
芝浦工業大学 システム工学部 電子情報システム学科†

1 研究背景と目的

データマイニングの手法には多様なものが提案されているが，このうちデータを可視化することにより視覚的に行われるものを視覚的データマイニング(Visual Data Mining)と呼ぶ．例として，次のような株式銘柄の財務データの分析において，結果として得られる知識について考える．

$\{V_n\} = \{v_i | v_i = (\text{株価収益率 純資産倍率 } \cdots)^T\}$

分析結果として望まれる知識には，例えば次のようなものがある．

1. 入力空間 $\{V_n\}$ のおおよその密度分布
2. 興味ある特徴を有する銘柄 v_c
3. 銘柄間の関係 $f(v_i, v_j)$ とその構造

このような知識は，v_i を一般に特徴ベクトルと捉えたうえで v_i 間の特徴距離を適切に可視化することにより得られる．この特徴距離の測度は，データに応じてEuclid測度などが選ばれる．ここで，あるデータに対し有益な距離測度は1つとは限らない．財務データにおいても，例えば何等かの重み付きEuclid距離を設定することで，データの異なる側面を表現できると考える場面は多々あるだろう．さらにデータによっては，複数の距離測度にわたる特徴・関係・構造などが意味ある結果を持つ．

ところが，こうした複数の距離測度の存在を前提とした分析に対し，汎用的に適用可能な可視化技術は存在しない．可視化の結果はいずれも距離測度ごとに別々の可視表現になるため，距離測度を越えた関係を視認することが極めて困難だからである．

本研究では，可視化技術のうちKohonenの自己組織化マップ(SOM)[1]を取り上げ，この射影空間と幾何学的表現を拡張することにより，前述の問題の解決を試みる．これを以下では拡張SOMと呼ぶ．

2 SOMにおける課題の定義

基本的なSOMは，高次元ベクトル集合からなるデータの2次元平面への非線型射影であり，入力空間の位相的な情報が保存される．ここで，入力空間とは射影時に用いる特定の距離測度によって決まり，複数の距離測度における射影は複数の射影結果（マップ）になる．このため，従来のSOMを用いて「距離測度を越えた関係を視認する」ためには，複数のマップに分散した情報を交互に読み取り，分析担当者が自ら連続的な関係を考える必要がある．しかしこれでは，データ項目間の関係が連続的に順序づけされるというSOMのメリットが生かされておらず，ある程度複雑な関係となると発見はほぼ不可能であった．

以上より解決すべき問題は，マップの生成において次の要件を満たすことであると考えられる．

I. 異なる入力空間の同一データ項目が，視覚的に容易に対応づけられる

II. 異なる入力空間の相異なるデータ項目において，位相的関係が視覚的に明確に識別できる

III. 条件1及び2を満たした上で学習過程が収束する

3 拡張SOMの提案と考察

3.1 拡張SOMの幾何学的表現

従来のSOMは，入力空間における位相関係をニューロンマップ(射影空間)上のEuclid距離[1]で対応させて射影していた．これに対し提案する拡張SOMは，Euclid距離に加えて同時にもう一つの別の距離を設定し，この距離に対して別の入力空間から射影することを可能にする．このような距離の性質としては，Euclid距離から独立しており，Euclid距離と同時に視認可能で，かつ直感性を失わないものが望ましい．しかし，2次元平面でEuclid距離の束縛を逃れることは容易ではなく，条件をすべて考慮すると独立性を妥協せざるをえない．

このため，今回はSOMの比較的緩やかな拡張となる距離を実験的に採用した．この距離は，ニューロンマップ上において，あるノードAからノードBへ最短で辿る際にノード間を移動した回数で定義され，これを「経路距離」と呼ぶ．この経路距離を用いた拡張SOMの幾何学的表現の例をFig.1に示す．このようにノードの形を様々に変化させることで，

*The Extended Self-Organizing Maps for Visual Data Mining

†Masayuki NAKAMURA; Department of Electronic Information Systems, Faculty of Systems Engineering, Shibaura Institute of Technology

[1] 幾何学的距離

Euclid 距離では遠いが経路距離では近いというように，異なった距離としての射影を可能とする．例えば図中の黒いマスの場合，そのＰ１近傍 (経路距離１の範囲) とＰ２近傍は斜線の範囲になる．このように，視覚的な直感でも Euclid 距離と経路距離の違いがみてとれる．ただし，独立性は損なわれており，他に何等かの改善を必要とすることは明らかである．

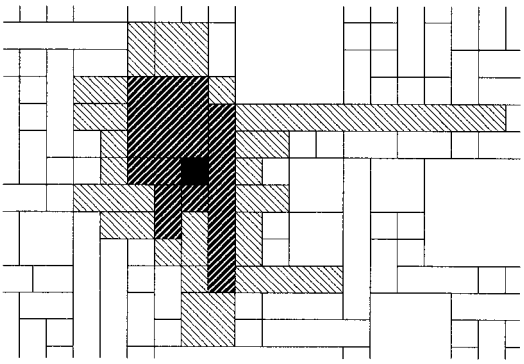

Fig. 1: 拡張 SOM の幾何学的表現の例

3.2 拡張 SOM のアルゴリズム

Fig.1 のような可視表現を生成する方法として，次の２つのアプローチがある．

A. それぞれの距離測度で従来の SOM を生成した後これを１つに合成する

B. SOM の訓練において２つの距離測度を同時に用いる

本提案の拡張 SOM を，可視表現を合成するだけの問題として考えるとＡが自然であるが，順序づけの精度に問題があることと，本提案の正当性と妥当性の検証には従来の SOM アルゴリズムとの比較が必要であることから，Ｂを採用する．

この具体的なアルゴリズムの議論は，大きく次の３点に分かれる．

3.2.1 近傍関数

経路距離による近傍関数 hp_{ci} の最も単純な形は次式で表される．

$$hp_{ci} = \alpha \cdot f(t, \|G_c - G_i\|_{path})$$

ここで，α は学習率係数，f は近傍カーネルを表す関数，G_i はノード i の幾何学的表現における位置情報，$\|G_c - G_i\|_{path}$ はノード c(最整合ノード) とノード i 間の経路距離である．

3.2.2 学習過程

拡張 SOM の学習過程は次式で表される．

$$m_i(t+1) = m_i(t)+[c_e \cdot he_{ci}(t)+c_p \cdot hp_{\tilde{c}i}(t)][x(t)-m_i(t)]$$

ここで，$x = [\xi_1 \xi_2 ... \xi_n]^T$ を入力ベクトル，$m_i = [\mu_{i1} \mu_{i2} ... \mu_{in}]^T$ をノード i の参照ベクトル，c と $\tilde{c}$ はそれぞれ Euclid 距離と経路距離による最整合ノードの添え字，he_{ci} と $hp_{\tilde{c}i}$ はそれぞれの近傍関数，c_e と c_p はそれぞれの影響を調整するパラメータである．

3.2.3 G_i の初期値と更新則

現状では G_i の設定に関し，理論的な洞察を得られていない．G_i の設定には，Euclid 距離との適切な対応と，学習の安定な収束性が求められる．このために検討すべきことは次の２点である．

適切な初期化

Euclid 距離からの独立性を持たせるために，ランダムでは都合が悪く，入力サンプルを用いた適切な初期化が望まれる．

更新の必要性とその方法

G_i の更新とは，時間 t によってノード間の距離が変化するということであり，これは従来の SOM には存在しない問題であるため，性能評価の観点から安易に導入できない．また，一つのノードの修正が周囲に及ぼす影響が局所的なものに留まらないために単純な逐次的更新が不可能であり，更新則の設定は困難である．しかし，Euclid 距離との柔軟な対応づけのために何等かの枠組みを必要とする可能性がある．

4 まとめ

本発表では，複数の距離測度の存在を前提とした分析に対して有効な SOM の拡張方法を提案した．現在，提案した拡張 SOM をフリーのプログラム・パッケージである SOM_PAK に対して実装しており，今後はこれを用いて本提案の評価と考察を行う予定である．

参考文献

[1] T.Kohonen: 自己組織化マップ, シュプリンガー・フェアラーク東京 (1996)