2A3-03

グラフィカルモデリングを用いた空間特徴抽出

Spatial feature extractions by graphical modeling

今原 修一郎[*1]
IMAHARA Shuichiro

佐藤 誠[*2]
SATO Makoto

[*1] (株) 東芝 研究開発センター
TOSHIBA CORPORATION.

[*2] (株) 東芝 研究開発センター
TOSHIBA CORPORATION

In this paper, we apply a graphical modeling method to spatial data and analyze where turns into housing area. It should select variables that spatial data have many attributes. We use the graphical modeling method to select variables that finds pairs which have relationship. We should convert from spatial data in consideration of neighbor relations to table data needed by the graphical modeling method. Statistical test are subject to be significant in such case of large amount of data as spatial data, so we use non-centered chi-square test in stead of chi-square test.

1. はじめに

昨今のストレージ容量の増大と計算機能力の向上により大量データの蓄積及び処理が可能となり、空間データを扱うことが容易になってきている。また政府の政策により国土空間データ基盤(基礎的な電子地図)が Web 上でダウンロードできるようになるなど、空間データは身近になりつつある。

こうした空間データにおいても表データにおける学習問題と同様、標高、国勢調査、地質といった空間情報と各地域につけられた情報(専門家による教師情報)からモデルを作成し、未知の空間情報が与えられた際に、該当する地域の属性を推定したい場合がある。

空間データは表データと比べて隣接する地域の情報も利用することが多いため、学習に使用する変数が多くなりがちである。これは変数の重要性に関する先見知識が無い場合(専門家の暗黙的な知識をルール化するような場合)にはこの問題はさらに顕著である。

これらの問題を踏まえ、本論文ではグラフィカルモデリングを使用して変数間の関係を求め、空間情報を使用したモデルの変数選択を行った。以下、グラフィカルモデリングの概要と空間データへの適用、実験手順とその結果について述べる。

2. グラフィカルモデリング

グラフィカルモデリング[Edwards 2000] [宮川 2003]とは変数間の直接的な関係を独立グラフとして表現するための手法である。変数間の直接的な関係とは対象とする2変数以外の全ての影響を取り除いた2変数間の関係を表しており、連続変数の場合は偏相関に、離散変数の場合には分割表における変数間の交互作用に相当する。

離散変数の場合には、分割表におけるセルの同時確率の対数を各変数間の交互作用の和で表現したグラフィカル対数線形モデルで変数間の関係をモデル化している。全ての交互作用が存在するモデルをフルモデルと呼び、いくつかの交互作用がないモデルを縮小モデルと呼ぶ。

最も単純なグラフィカルモデリングのアルゴリズムを以下に示す。

(1) フルモデルをモデル1とし、その逸脱度を計算する
(2) モデル1から1つだけ関係を削除したモデルをモデル2とし、その逸脱度を計算する
(3) 逸脱度の差をカイ二乗検定する
(4) 全ての関係について検定し、全ての P 値が棄却域を越えなければモデル1を出力する
(5) 一番大きいP値を持つ関係を切断してモデル1とし、(2)へ戻って繰り返す

ここで、逸脱度とはフルモデルの尤度と縮小モデルの尤度の差の 2 倍であり、以下の式で表される。

$$dev = 2\sum_i n_i \log \frac{n_i}{m_i}$$

ここで n_i はセル度数を、m_i は縮小モデルの元での期待セル度数を表している。期待セル度数を求めるには IPS アルゴリズムによって反復計算で求めるか、あるいは分解可能モデルの場合のみを考えてグラフをクリークとボーダーに分解して計算する方法がある。後者は切断できない関係が存在するかわりに単純で高速な計算方法である。

このようにしてフルモデルから順に1つずつ関係を削除し、最終的に得られたモデルは変数間の直接的な関係を表している。

3. 空間データへの適用

空間データに対してグラフィカルモデリングを適用する際には、空間データを表データにする際の空間的隣接情報の表現方法、変数間の関係ではなく変数のカテゴリ値間の関係を得る方法、大量データ量への対応方法について考える必要がある。

以下、これら3つの問題について考える。

3.1 空間データの表データ化

空間データには大きく分けてポリゴン、ライン、ポイントに代表されるベクトル型データと、配列状にデータが格納されているラスタ型データの2つがある。前者は公示地価などのポイントデータ、行政界ごとに格納された国勢調査データなどがあり、後者は標高や土地利用、国勢調査メッシュデータなどがある。今回は隣接関係演算が単純で表データの形式に変換しやすいラスタ型のデータを対象とする。

グラフィカルモデリングは表データを対象としているため、空間データから表データに変換する必要がある。空間データの特徴である隣接関係による相関を考慮するために、メッシュ全体の左上から右下に向かってZ字に一次元化したものに加え、中心のメッシュと中心メッシュに隣接する 8 つのメッシュから代表

*1 shuuichiro.imahara@toshiba.co.jp, +81-44-549-2140
*2 makoto12.sato@toshiba.co.jp, +81-44-549-2140

値（例えば最頻値）を計算したものを同様に一次元化して表データの属性として追加する。教師データ以外は全ての空間データに関してこのように表データ化する。

3.2 カテゴリ値空間データの 2 値表データ化

カテゴリ値を持つ空間データ、例えば土地利用データを表データ化する際、カテゴリ値のまま表データ化するのではなく、回帰分析におけるダミー変数のように各カテゴリ値の有無(2 値)に分解することにより、土地利用データそのものではなく各カテゴリ値に対してグラフィカルモデリングで関係を調べることができる。

グラフィカルモデリングは変数間の関係を調べるためのものであるが、このように 2 値に分解することで値そのものの関係を調べることができ、より詳細な調査が可能となる。

このように 2 値に分解すると分解されたカテゴリ値ごとの変数は明らかに大きな相関を持つ。これらの変数間の関係は常に成立するため、グラフィカルモデリングによる関係の評価を省略して高速化を行うことができる。

3.3 空間データのデータ量の問題

差異があるかどうかの統計的検定は、差異が 0 か否かを検定するため、データ量が多くなると検定が有意になりやすいという特性がある。社会調査などの分野ではデータ取得に多大なコストが発生するため、必要十分なデータ量のみを取得して検定に使用するので問題は起きないが、空間データのように大量のデータが存在する場合には対策が必要となる。[保田 2004]

対策は2つあり、1つ目は大量のデータから適切な量になるようサンプリングを行うことである。2つ目は非心カイ二乗分布を導入し、差異があるかどうかを検定するのではなく、非心度で表される値の範囲を越えるかどうかを検定するようにすることである。これにより検定はデータ量に依存しなくなる。

このうち前者はせっかくのデータを捨てていることに等しくなるため、本論文では後者の方法を使用する。ただし、非心度の決め方に主観が入ってしまうため、アルゴリズムの実行前に許容度を信頼率の 3%と設定した。

4. 実験

標高データと土地利用データから、宅地化する際に関係のある変数を調査する。つまり、宅地以外から宅地に変化した地域の関係を調査したい。

使用した空間データは、(1)標高 50m メッシュ、(2)1984 年細密数値地図、(3)1989 年細密数値地図で、溝の口駅周辺の地域のみを対象とした。(1)は連続値であるのでデータのレンジを 3 等分して高中低と離散化した。またメッシュの大きさをあわせるため、(2)と(3)は 5×5 の 25 マスの中から最も頻度の高い値をそのメッシュの値とする前処理を行った。

教師データは(2)から抽出した住宅地が偽で(3)から抽出した住宅地が真のメッシュのみを真、それ以外は偽としたものとする。入力データは(1)から抽出した標高の離散値、(2)から抽出した商業地、農業地、森林、工業地の真偽値である。入力データは中央メッシュのみの場合と隣接メッシュを考慮した場合の 2 通りがあるため、与えた変数の合計は 15 個である。またデータサイズは横 160 個、縦 120 個をZ字に一次元化した 19200 個である。

上記のデータを使用し、これら 15 個の変数に対してグラフィカルモデリングを実行した。データサイズが大きく、前述した統計的検定の問題が起こりやすくなるため、非心カイ二乗分布を使用した統計的検定を行っている。このとき、使用した非心度は 0.0009×19200＝17.28 である。これらのデータ対してグラフィカルモデリングを行った結果を図1に示す。

5. 考察

図 1 を見ると、宅地化した地域は標高に関係なく満遍なく宅地化されていることが分かる。また、隣接メッシュがなんであるかは問わず、元が農業地である地域は宅地に変化しやすいということが分かる。これらの結果は 84 年から 89 年にかけての宅地建設の傾向を表していると思われる。

地図を見たところ、森林は標高が高めの地域に多数存在していた。隣接メッシュに標高との関連が出ているのは森林が固まって存在しているためで、おそらく広い範囲で成り立つ方が優先された結果だと推測される。このような現象を避けるために隣接メッシュ同士は接続できないという条件を設けたほうが結果がわかりやすいかもしれない。

結果では標高が低の地域に関しては隣接との間に関係はないとされてしまっているが、地図を見ると低い地域は広い範囲で存在しており、結果と合わない。グラフィカルモデリングで関係を切断する様子を観察すると、この問題は切断初期の段階で標高に関する関係の P 値がほぼ等しく無関係を表す 1.0 の値を示していることに起因すると思われる。これによりプログラム的に先に来る関係が優先的に切断されてしまった結果、標高が低い場合における関係がほとんど切断されてしまったものと考えられる。

この問題を解決するためには、切断アルゴリズムを改良して同一の値を持つ場合には両方の場合を試してみるなどの工夫が必要となる。

6. まとめ

変数が多くなりがちな空間データにおいて、変数の重要性に関する経験的な知識があまりない場合に変数選択を考慮した分析手法としてグラフィカルモデリングを使用した空間分析を提案した。

空間データを適用するにあたり、空間特有の問題である隣接関係やデータ量の問題を対処し、さらにカテゴリ値を 2 値に分解してからグラフィカルモデリングすることでカテゴリ値の関係が得られるようになった。

また、P 値が 1 になるという問題が分かったため、今後は切断アルゴリズムの改良に努めたい。

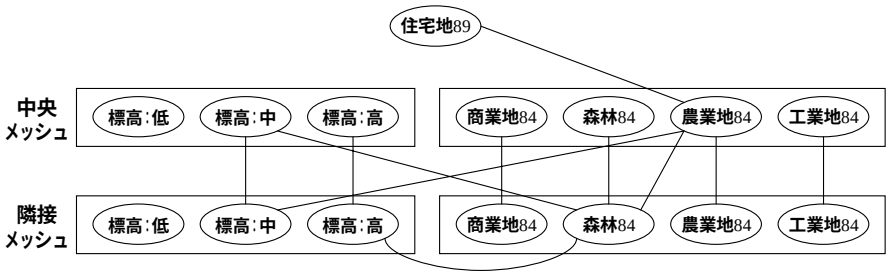

図1：変数間の関係

参考文献

[Edwards 2000] David Edwards: Introduction to Graphical Modelling, Springer, 2000

[宮川 2003] 宮川雅巳: グラフィカルモデリング, 朝倉書店, 2003

[保田 2004] 保田時男: 大規模サンプルに対する一般化χ2適合度検定―JGSS データへの適用例―, 日本版 General Social Surveys 研究論文集[3] JGSS で見た日本人の意識と行動, 大阪商業大学比較地域研究所·東京大学社会科学研究所編, 2004

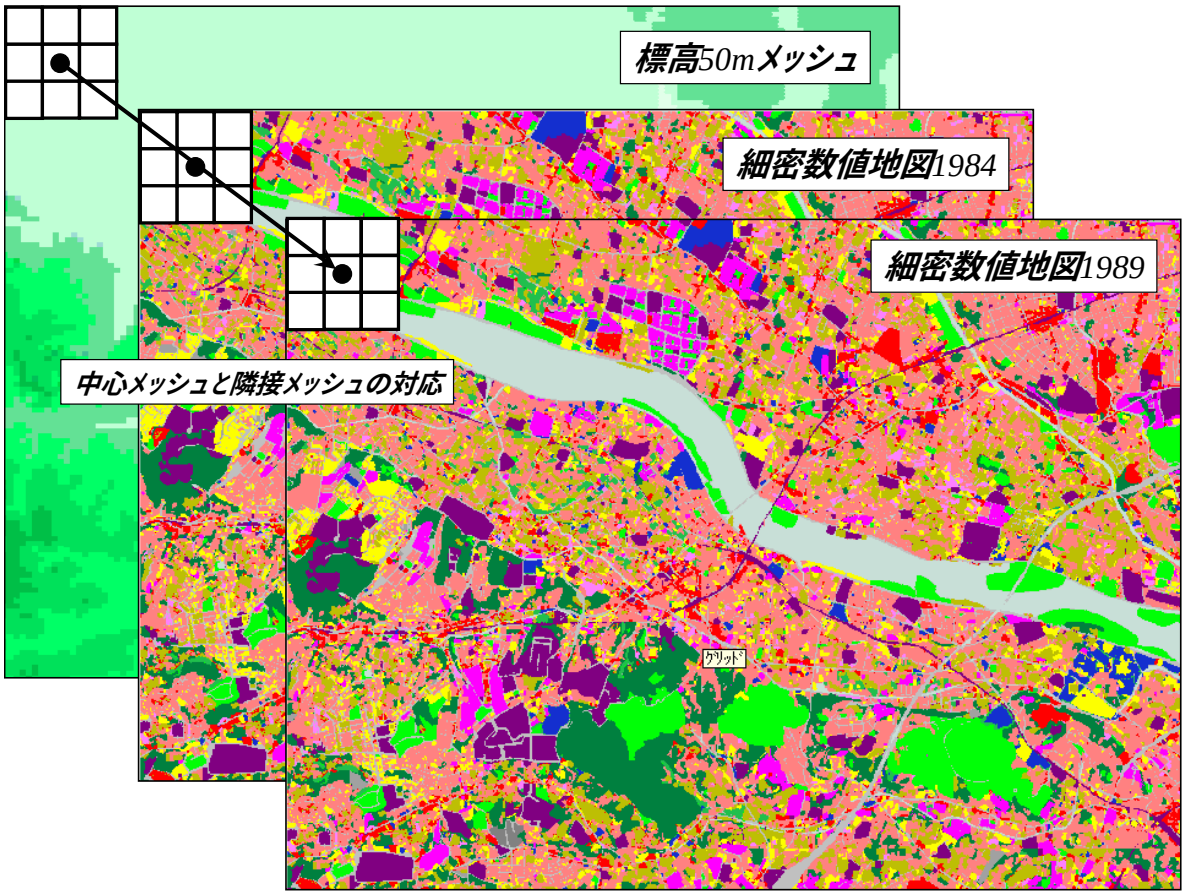

図２：空間データにおける中心メッシュと隣接メッシュの対応