2L-1

決定木を用いた質的データからのグラフ構造学習

川崎　隆史†　　鈴木　大輔†　　杉山　歩†　　Dam Hieu Chi†
†北陸先端科学技術大学院大学　知識科学研究科

1　はじめに

近年，グラフやネットワークから知識発見を行おうという試みが増加している．SNS からのコミュニティ抽出や，タンパク質の相互作用リンク予測などはその一例である．

しかしながら，グラフを用いた知識発見には，分析の目的に合わせてデータからグラフ構造を決定する必要があり，グラフ構造を如何にして決定するかという問題は非常に重要である．その中でも，近年増加している多次元の複雑なデータに対応するため，重要な関係性のみを取り出してグラフ構造を決定するという変数選択の問題を考慮した手法が求められている．

このような問題を解決する手法として N.Meinshausen -P.Bühlmann の手法 (MB 法)[1] が提案された．これは回帰分析を全変数に網羅的に行うことによってグラフ構造を学習する手法であり，変数選択には Lasso 正則化 [2] を用いている．これにより，多次元の複雑なデータから重要な関係性のみを抽出したグラフ構造を学習することに成功した．

しかし，MB 法は回帰分析を用いるために量的なデータを分析するには適切であるが，質的データの分析，とりわけ二値データに対する分析には不適切である．この問題に対し，ロジスティック回帰を用いて対応した既存研究が存在する [3]．しかし，目的変数に対する説明変数間の関係性を吟味出来ないなどの問題があり [4]，あらゆる場合において適切な手法であるとはいえないことがわかっている．

2　研究目的

本研究の目的は，質的データから説明変数間の関係性を考慮したグラフ構造を学習することである．そのために，決定木を用い，変数選択に変数減少法を採用する．グラフ構造学習のアルゴリズムに MB 法の発想を利用する．これにより，MB 法の優位点を継承した質的データを扱う手法の開発を目指す．

3　研究方法

3.1　決定木によるグラフ構造学習

MB 法によるグラフ構造の学習は，データの持つ N 個の変数 $(x_1, x_2, ..., x_N)$ について各変数 x_i を目的変数とし，その他を説明変数とした回帰分析を N 回行い，その結果を統合することにより行われる．本研究はこの方法を利用する．

決定木による構造学習の流れを下記に示す．なお，決定木のアルゴリズムは CART，変数重要度の算出に関しては，Gini-Importance を採用した．モデル推定には 10-fold-CrossValidation(CV) を用い，各分割データにおいて変数減少法を用いたモデル構築を行う．その結果から，最良なモデルの構築を行う．

Algorithm 1 決定木によるグラフ構造学習

1: Input: データ $X = (x_1, x_2, ..., x_N)$
2: Output: 変数重要度の行列
3: **for** $i = 1$ to N **do**
4:　Set $f(x_i) = (x_1, x_2, ..., x_{N-1})$
5:　$f(x_i)$：モデル推定
6:　$f(x_i)$：モデル構築
7:　$f(x_i)$：変数重要度の算出
8: **end for**

3.2　決定木とグラフ構造の対応

決定木モデルとグラフ構造の対応を図 1 に示す．変数 A,B,C が存在する時に，A が目的変数だった場合の決定木とグラフを表している．エッジの方向は B,C が A を説明することを意味し，実線は決定木の第一層，破線は第二層以降に用いられた変数で，それ以前の層による条件付きの関係性であることを表している．

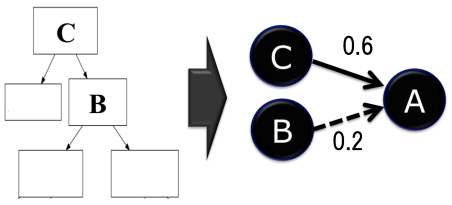

図 1: 決定木モデルとグラフ構造の対応

Graph Structure Learning from Categorical data by Decision Tree
†Takafumi KAWASAKI　†Daisuke SUZUKI　†Ayumu SUGIYAMA
†Dam Hieu Chi
†School of Knowledge Science, Japan Advanced Institute of Science and Technology

4 検証と評価

4.1 検証用データ

ニコニコ動画が提供する動画データ*(2012-6-03〜2012-11-01) のタグ情報を使用する．概要を以下に示す．

- 変数 (タグ) の数：24
- サンプル数：10456
- 変数要素：{0, 1}

上記のデータセットを構築するにあたり，使用するタグは出現率上位 0.001%のものに絞っている．さらに各動画データは絞られたタグのうち最低 4 つ以上を持つものに限定し，その上で総データ数の 5%以上は {1} を持つタグに絞った．

4.2 検証

本研究の手法と，変数減少法を採用しない決定木による学習結果とを比較した．これにより，変数減少法が有効であることを明らかにするとともに，両者のグラフ構造にどのような相違点があるのかを示す．

4.2.1 数値結果の評価

表 1 は両手法の予測精度平均と使用変数個数平均を表している．これは全変数 N 個分の決定木の予測精度と使用変数個数の相加平均である．これを見ると，本研究の手法である変数減少法を用いた決定木が，CV のみのものと同等の予測精度を保ちつつ，使用変数個数を減らせていることがわかる．これは，変数減少法を用いることで，重要な変数が選択されていることを示していると考えられる．

表 1: 両手法の数値結果

	CV のみの決定木	本研究の手法
予測精度平均	0.956	0.953
使用変数個数平均	22	16

4.2.2 グラフ構造の評価

図２，３は各々の手法によって学習されたグラフである．これを見ると，両者に構造上の違いはほとんどない．しかし係数に目を移すと，本研究の手法の方が全体的に値が大きくなっている．これは，本研究の手法が使用する変数を出来るだけ減らすことで，係数値の分散を防いでいることが要因と考えられ，より変数重要度を考慮した分析が可能であることを示している．

両手法に共通する点として，いくかの変数間に破線が見られる．これは，決定木による分析が説明変数間の条件付き関係を分析していることを示しており，グラフからの知識発見において，よりな知識を発見できる可能性があることを示唆する．

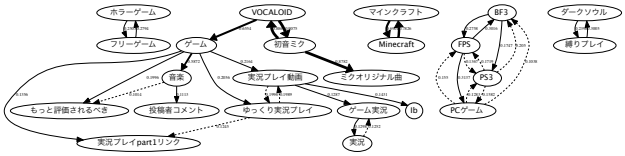

図 2: 本研究の手法から学習されたグラフ構造

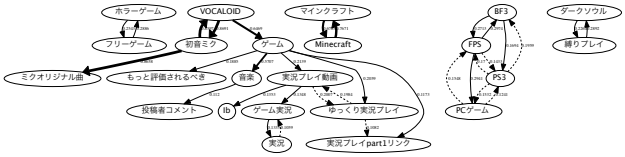

図 3: CV のみの決定木から学習されたグラフ構造

5 まとめ

本研究では，質的データからのグラフ構造学習手法として，変数減少法を採用した決定木を提案し，検証により変数減少法が有効であることを示した．今後は本手法によって得られたグラフが，他の手法によって得られたものと知識発見の面においてどのような特性があるのかを検証していきたい．

参考文献

[1] N.Meinshausen and P.Bühlmann,「High-dimensional graphs and variable selection with lasso」, Annals of Statistics, 34(3): 1436-1462, 2006

[2] Tibshirani, R., 「Regression shrinkage and selection via the lasso」, J. Roy. Statist. Soc. Ser. B, 58, pp.267-288, 1996

[3] Wainwright, Martin J and Lafferty, John D and Ravikumar, Pradeep K,「High-Dimensional Graphical Model Selection Using L1-Regularized Logistic Regression」, Advances in neural information processing systems,pp.1465-1472, 2006

[4] 奥喜正，本村猛能，前鶴政和，内桶誠二,「データマイニングにおける二値データ解析: 決定木とロジスティック回帰分析」, 物流問題研究, Vol44, pp.1-14, 2004

*http://www.nii.ac.jp/cscenter/idr/nico/nico.html