修 士 論 文

決定木を用いた質的データからのグラフ構造学習

北陸先端科学技術大学院大学
知識科学研究科

川崎　隆史

2014 年 3 月

修 士 論 文

決定木を用いた質的データからのグラフ構造学習

指導教官 Dam Hieu Chi 准教授

北陸先端科学技術大学院大学
知識科学研究科

1250014 川崎　隆史

審査委員: Dam Hieu Chi 准教授 (主査)
Ho To Bao 教授
藤波 努 教授
Huynh Nam Van 准教授

提出年月: 2014 年 2 月

目 次

第 1 章　序論 **1**

1.1 研究背景 1

1.1.1 グラフを利用した研究 1

1.1.2 関係性の種類 2

1.1.3 グラフとは 2

1.1.4 既存研究 3

1.2 研究目的 5

1.3 知識科学的意義 5

1.4 本論文の構成 5

第 2 章　関連研究 **6**

2.1 N.Meinshausen -P.Bühlmann の手法 6

2.1.1 MB 法による学習 6

2.1.2 Lasso 正則化 7

2.1.3 MB 法のまとめ 8

第 3 章　本手法の概要 **9**

3.1 各手法の概要 9

3.1.1 決定木 9

3.1.2 変数減少法 11

3.1.3 交叉検証法 11

3.1.4 Gini-Importance 12

3.2 構造学習の行程 13

3.2.1 構造学習行程の全容 13

3.2.2 モデル推定 14

3.2.3 モデル構築 17

3.3 決定木構造とグラフ構造 18

第 4 章 検証実験と考察 **19**

4.1 検証実験の目的 19
4.2 実装環境 19
4.3 使用データ 20
4.3.1 データ成形 20
4.4 実験結果と考察 22
4.4.1 変数減少法の有効性 22
4.4.2 学習されたグラフの妥当性 25
4.4.3 説明変数間の関係性 26
4.4.4 予測的分析と相関分析の比較 27

第 5 章 結論と今後の課題 **29**

5.1 まとめ 29
5.2 今後の課題 29
5.2.1 変数選択手法について 29
5.2.2 変数重要度の評価について 30
5.2.3 多重共線性の問題について 30

参考文献 **32**

謝辞 **33**

付録 **34**

本研究に関する発表論文 **36**

図 目 次

1.1 関係性の種類 2
1.2 両手法によるグラフ構造学習 3

2.1 MB 法によるグラフ構造学習 6
2.2 Lasso 正則化による変数選択 8

3.1 決定木の例 10
3.2 k-fold-CrossValidtion 12
3.3 Gini-Importance による変数重要度の表現 13
3.4 決定木によるグラフ構造学習のフローチャート 14
3.5 CV と変数減少法によるモデル推定 15
3.6 モデル推定の目的 15
3.7 変数減少法のフローチャート 16
3.8 最終的に構築に使用される変数 17
3.9 決定木構造とグラフ構造 18

4.1 成形後のデータ 21
4.2 学習過程 23
4.3 検証用データから学習されたグラフ 25
4.4 本手法によるグラフ 27
4.5 クラメールの連関係数によるグラフ 27
4.6 目的変数による変数重要度の変化 28

5.1 動画情報 JSON 34
5.2 タグ情報 JSON 35

表 目 次

1.1 データタイプと対応する手法 3

4.1 使用タグ一覧 21
4.2 両手法の予測精度平均 23
4.3 両手法の使用変数個数平均 24

第1章　序論

1.1　研究背景

1.1.1　グラフを利用した研究

現在，Web 上には多種多様な情報が存在し，日々増加している．それに加え、近年急速に成長したSNS(ソーシャルネットワーキングサービス) の流行などもあり，それらを利用することによる社会ネットワーク分析が盛んに行われている [14][12]．これは，グラフを用いて社会システムを理解しようという試みであり，様々な成果が報告されている．

グラフを用いてシステムを理解しようという研究は，社会科学だけでなく，バイオインフォマティクス等の分野にも広がっている．遺伝子やタンパク質などの要素間相互作用をグラフで表現して分析する研究などがその一例である [9].

また，データマイニングの一分野として，より巨大なネットワーク構造を扱うリンクマイニングと呼ばれる分野も現れている [15]．これは，ネットワークの動的な変化や，新たに生成されるであろうリンクを予測するなど，より高度なグラフ分析を扱っている．

このようなグラフの利用は，グラフをデータ（入力）として利用することで，システムの理解という結果を得る事を目指すものであるが，グラフは法則・関係性を表現（出力）するのに利用することもできる．単純なものでは，SNS の友人間の関係性を可視化，複雑なものでは，センサーの時系列データを元にデータの類似性からセンサー間の関係性を定義してグラフ化するものなどがある [2]．また，ベイジアンネットも表現の一例と言える．ベイジアンネットは，因果関係という高度な関係性をグラフで表現しており，グラフが多様な関係性を表現可能なことを示している．

このように，グラフは様々な分野においてデータあるいは表現として利用されている．しかしグラフを利用するには，グラフの構造を決定しなくてはならず，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．

1.1.2　関係性の種類

グラフ構造は，グラフから何を発見するのかという目的に合わせて決定される．SNS では友人間関係，Web ページではリンクの有無などを利用してグラフ構造を決定するのが一般的である．こういった，見た目上の形式的関係を利用したグラフ構造の決定は比較的容易である．

しかし，関係性は必ずしも形式的なものによって決定されるとは限らず，潜在的な類似性をもって決定することも可能である．形式的な関係と潜在的な関係によるグラフ構造決定の違いを図 1.1 に示す．

潜在的な関係性とは，人間を例に説明すると，年齢・性別・出身・年収・etc などの様々なデータの類似度によって決定されるものである．形式的な関係と比較して，潜在的な関係を明らかにするのは難しく，相関分析や予測的分析による構造学習を必要とする．

図 1.1: 関係性の種類

1.1.3　グラフとは

既存研究について触れる前に，数学的なグラフの定義と，本研究におけるグラフの定義について述べる．

数学的なグラフの定義

数学的なグラフの定義を「グラフ・ネットワーク・組合せ論」[10] を参考に述べる．
１つのグラフを定義するのに次の４つの材料（構成要素）が必要である．

1. 集合 V：V の要素を点（ノード）といい，V を点集合という．
2. 集合 A：A の要素を枝（エッジ）といい，A を枝集合という．
3. 写像 ∂^+（$A \to V$）：各枝 $a \in A$ に対して $\partial^+ a$ で指定される点を枝 a の始点という．

4. 写像 ∂^-（$A \to V$）：各枝 $a \in A$ に対して $\partial^- a$ で指定される点を枝 a の終点という．

これらを合わせて，グラフ $G = (V, A, \partial^+, \partial^-)$ のように書く．ここで，混乱のおそれのない場合，しばしばグラフ $G = (V, A)$ のように略記する．

グラフにおいて，枝の向きに関心がないときは，そのグラフを無向グラフといい，通常のグラフを無向グラフと対比していう場合には特にこれを有向グラフという．

本研究におけるグラフの定義

本研究では，グラフのノードをデータの持つ変数，エッジを変数間の関係性を表すものとして定義する．

1.1.4 既存研究

潜在的な関係を明らかにする方法は，相関分析と予測的分析の二つの方法が存在する．両者の違いを図 1.2，量的・質的データに対応する手法を表 1.1 にそれぞれ示す．

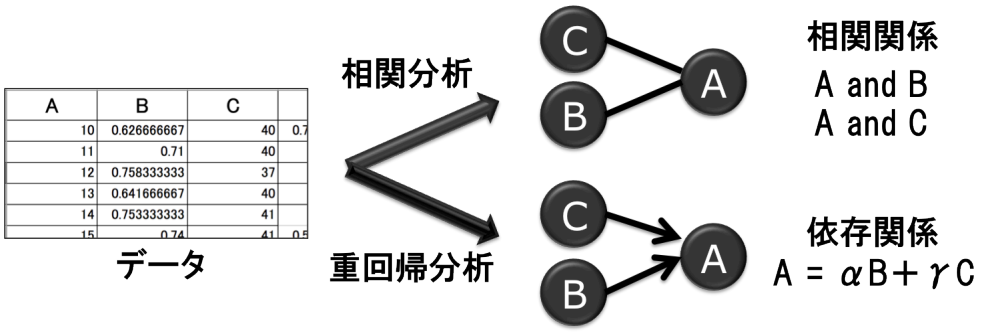

図 1.2: 両手法によるグラフ構造学習

表 1.1: データタイプと対応する手法

	量的データ	質的データ
相関分析	ピアソンの積率相関係数	クラメールの連関係数
予測的分析	N.Meinshausen -P.Bühlmann の手法（重回帰分析）	M.Wainwright らの手法 (ロジスティック回帰分析)

相関分析と予測的分析の大きな違いは，記述を目的とするものと，予測を目的とするものという違いもあるが，単変量分析か多変量分析かという違いも大きい．図 1.2 を見るとわかるように，相関分析によって得られるグラフが無向グラフなのに対し，重回帰分析によるものは有向グラフとなっている．これは，多変量による分析の場合，必ずしも両者の関係性を示す数値が一致するわけではないことを意味している．

また，両者から得られるグラフは，そこから得られる知識の質においても違いがある．単変量分析である相関分析では，得られる知識は，変数ペアにおける関係性という相関関係に留まる．それに対し多変量分析では，目的変数に対する説明変数群全てを考慮した関係性を分析できる．例えば重回帰分析では，目的変数に対する説明変数の影響を線形結合によって表した式を得る事が出来る．知識の質としては，依存関係の方が高度であるといえる．

表を見ると，予測的分析において，量的データでは N.Meinshausen -P.Bühlmann の手法（以降 MB 法）[5]，質的データでは M.Wainwright らの手法 [7] が存在する．

MB 法は，重回帰分析の変数選択に Lasso 正則化 [6] を取り入れてグラフ構造学習を行うことで，重要な関係性のみを反映したグラフを学習することに成功した．これは,「オッカムの剃刀」の原則に従ったシンプルな法則によってのみ構成されたグラフを学習できていることと同義であり，グラフからの知識発見において，より洗練された知識の発見を可能とするものである．M.Wainwright らの手法は，MB 法の発想をロジスティック回帰分析に取り入れたものである．MB 法は，本研究において最も重要な関連研究であるので，第 2 章で内容について詳しく述べる．

M.Wainwright らの手法は，ロジスティック回帰分析を用いてグラフ構造学習を行っている．しかし，ロジスティック回帰分析には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 [8]，あらゆる場合において適切であるとはいえないことがわかっている．また，回帰分析では説明変数の中で質的データを持つ変数をダミー変数に変換して分析を行うのが一般的に行われているが，これは目的変数が量的で説明変数の中に質的変数が存在する場合に用いるものであり [11]，目的変数・説明変数ともに質的変数のみで構成される場合に用いることを想定されたものではない．

そこで本研究では，質的変数のみで構成されるデータに適し，かつ説明変数間の関係性を考慮出来る予測的分析手法の開発を目指す．

1.2 研究目的

本研究の目的は，質的変数のみで構成される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，分析手法には質的データに対応し，かつ説明変数間の関係を階層的に表すことのできる決定木を用いる．また，洗練されたグラフを学習するための変数選択には，変数減少法 [1] を採用し，グラフ構造学習のアルゴリズムには MB 法の発想を利用する．

1.3 知識科学的意義

グラフ構造学習とは，データが持つ潜在的な法則・関係性を得るという知識発見であり，それをグラフで表現するという知識表現の一つでもある．本研究が提案する手法により，質的データから説明変数間の関係性を考慮したグラフを得る事が出来たならば，データからの知識発見・表現という分野の幅が広がることとなる．また，得られたグラフをデータとして用いリンクマイニングを行うことで，さらに高度な知識発見に繋がることが期待でき，データからの知識発見という意味で本研究には大きな意義があるといえる．

1.4 本論文の構成

本論文の構成を述べる．本章では，研究背景と目的について述べた．第 2 章では，本研究を考えるにあたり最も参考とした関連研究である MB 法について詳しく述べる．第 3 章では，本研究の提案する手法に使用される手法・基準などについて簡単に述べた後，本手法による構造学習の概要を述べる．第 4 章では，本手法の妥当性・有効性等に関する結果と考察を述べる．第 5 章は，本研究のまとめと今後の課題を述べる．

第2章　関連研究

本章では，本研究の関連研究について述べる．

2.1　N.Meinshausen -P.Bühlmannの手法

2.1.1　MB法による学習

グラフ構造を学習するというのは，隣接行列を学習するということと同義である．N.Meinshausen -P.Bühlmann の手法（MB 法）は重回帰分析を用いて隣接行列を求めることで構造学習を行う．MB 法による学習の概念図を図 2.1 に示す．

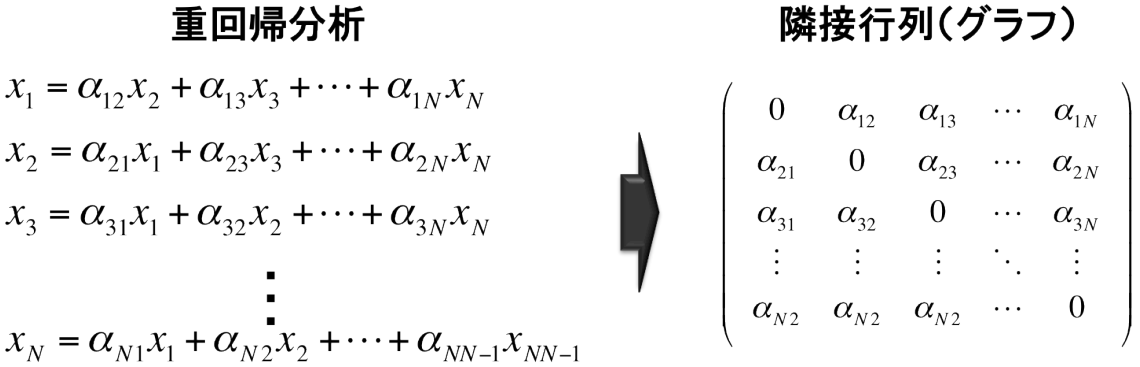

図 2.1: MB 法によるグラフ構造学習

MB 法の発想というのは，データの持つ N 個の変数に対し，各変数を目的変数とした重回帰分析を網羅的に行う事により，隣接行列を求めるものである．本研究においても，この発想を利用する．

2.1.2 Lasso 正則化

MB 法には分析の発想以外に，もう一つ重要な概念が用いられている．それが Lasso 正則化である．これは目的変数にとって重要な説明変数のみを残すという変数選択を行うものであり，回帰分析においては，推定式に正則化項を追加することで行われる．Lasso 正則化の利点は，正則化項を推定式に埋め込むことで，変数選択とモデル推定を同時に行う事ができる点にある．ここでは，回帰分析において一般的に使用される最小二乗法によるモデル推定を例に説明する．

最小二乗法は次式を最小化することで得られる．

$$\frac{1}{m}\sum_{i=1}^{m}(y_i^{predict} - y_i^{obs})^2 \tag{2.1}$$

ここで，m はデータ数，$y_i^{predict}$ は学習されたモデルによる予測値，y_i^{obs} は実測値を表す．また，$y_i^{predict}$ は次式をのように表される．

$$y_i^{predict} = \sum_{j=1}^{n} \alpha^j x_i^j + \alpha^0 \tag{2.2}$$

ここで，n は全説明変数の数，x_i^j は j 番目の説明変数の値，α^j は x_i^j に対応する回帰係数，α^0 は切片である．

Lasso 正則化を取り入れた回帰分析を Lasso 回帰という．これは，最小二乗法に正則化項を加えた次式によって表される．

$$\frac{1}{m}\sum_{i=1}^{m}(y_i^{predict} - y_i^{obs})^2 + \gamma \sum_{j=1}^{n} |\alpha^j| \tag{2.3}$$

正則化項の γ を操作することにより，正則化の影響力を操作することができる．γ の操作により，重要な変数が選択されている様子を図 2.2 に示す．図 2.2 における縦軸は各説明変数の係数 α^j の値，横軸は γ の値を示している．これを見ると γ の値を大きくして，正則化項の影響力が高められるにつれて，説明変数群の中で係数値が 0 となるものが増加していくのがわかる．

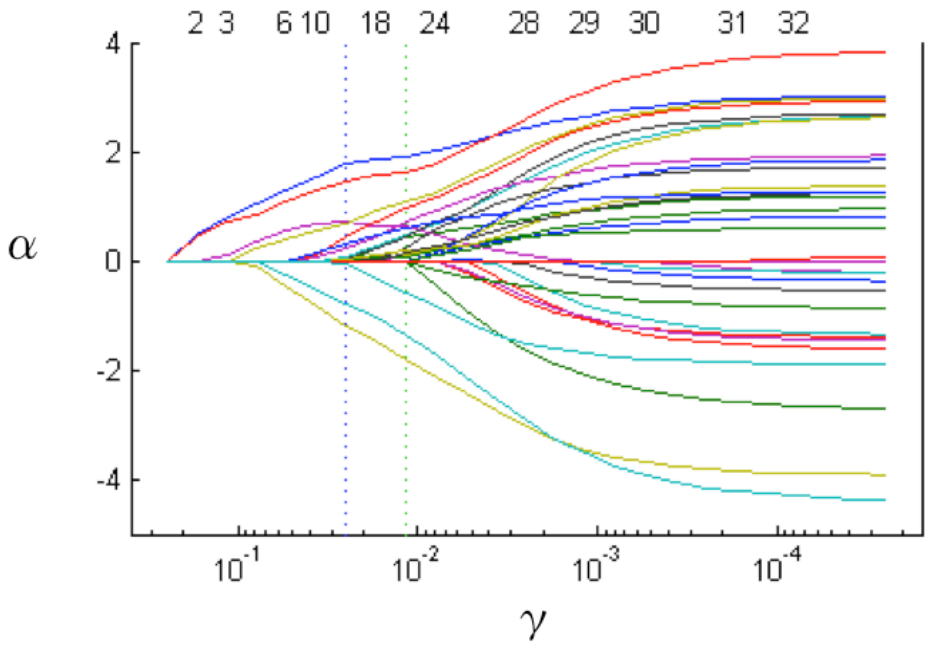

図 2.2: Lasso 正則化による変数選択

2.1.3 MB 法のまとめ

MB 法は，重回帰分析を網羅的に行うという発想と，Lasso 正則化による変数選択によって構成される．これにより，学習されるグラフは重要な関係性のみを抽出されたものとなり，洗練されたグラフ表現を得ることが可能となった．本研究は MB 法の論理を取り入れることで，MB 法の利点を取り入れた手法の開発を行う．

第3章　本手法の概要

本章では，本手法に使用される決定木・変数減少法等の各手法について概要を述べる．その上で，本手法のグラフ構造学習行程を述べる．

3.1　各手法の概要

3.1.1　決定木

概要

決定木の説明は,「データマイニングの基礎」[13] を参考に行う．

決定木は，意思決定や事象の分類を多段階で行う場合に，その多段の分岐過程を階層化して樹形図で表現したグラフ表現である．目的変数が質的なデータである場合を分類木,量的なものである場合を回帰木という．本研究は質的データを扱うので，分類木を用いることとなる．よって以後の説明は分類木に焦点を当てたものとする．

図 3.1 に決定木の例を示す．これはプログラミング言語 R[1]の mvpart パッケージ[2]を使用して作成したものであり，データには iris データ[3]を使用した．決定木において，各分割が行われる部分を親ノードといい，分割後の部分を子ノードという．特に木の終端に存在する子ノードのことを葉ノードという．決定木は，子ノードの全ての事例が同一クラスに属すまで，子ノードを親ノードとして再帰的に分割を行うトップダウン式の分割統治法に相当する．

[1]http://www.r-project.org/
[2]http://cran.r-project.org/web/packages/mvpart/index.html
[3]http://stat.ethz.ch/R-manual/R-devel/library/datasets/html/iris.html

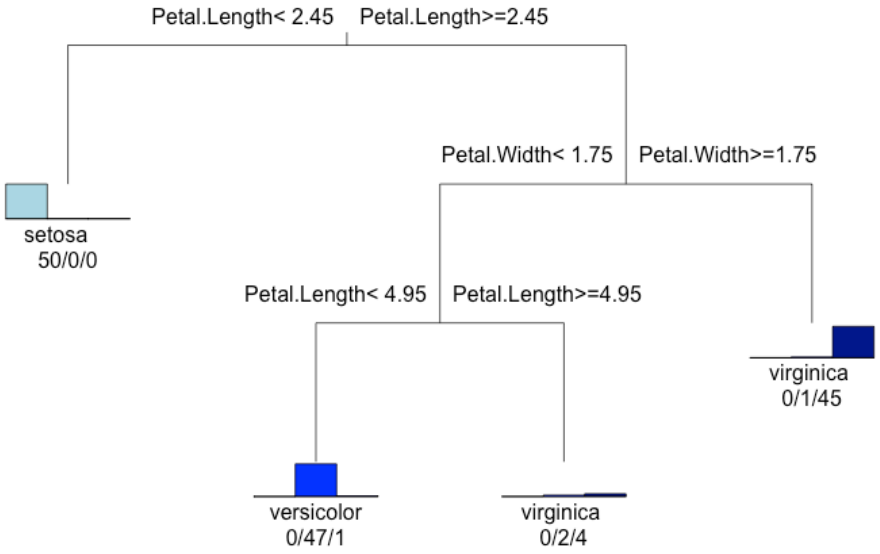

図 3.1: 決定木の例

アルゴリズム

決定木のアルゴリズムとして， 多分木を学習する C4.5 と二分木を学習する CART が有名であるが，本手法では CART を用いた決定木を採用する．両者の違いは，分割において C4.5 は多分割を行い，CART は二分割を行うという点である．しかし，基本的な学習アルゴリズムは同じである．下記に決定木の学習プロセスを示す．

1. 親ノードに置く説明変数を決定し，分割を行う．
2. データ集合を各分岐に応じて部分集合に分割して子ノードを作成し，親ノードとする．
3. １，２のプロセスを再帰的に行い，決定木の学習を行う．

4. 子ノードのすべての事例が同一クラスに属したら学習を終える．

このプロセスの中で問題となるのは，親ノードに置く説明変数を所与の説明変数の中からどう選択するのかということである．最良の決定木を学習するという問題はNP困難として知られており，全ての可能性を探索するわけにはいかない．よって良い分割変数を選択するための基準が必要となる．

選択基準

一般的に知られる選択基準として，ジニ係数とエントロピー（情報利得）が存在する．本研究ではエントロピーを採用しているので，それに焦点を当てて述べる．

エントロピーとは，情報の不確実性を表すものであり，平均情報量 $-\sum p_i \log_2 p_i$ で示される．良い決定木とは，各葉ノードに所属する事象が単一のクラスに属するものであり，エントロピーが0となることである．つまり，各分割変数において，その子ノードのエントロピーが最大限低くなるものを選択すれば，最終的に良い決定木が学習されることとなる．

3.1.2 変数減少法

変数減少法 [1] は，全ての説明変数でモデルを学習した後に，モデルにとって最も必要性の薄いと考えられる変数から一つずつ取り除いていくという操作を，何らかの基準に達するまで逐次的に行うものである．基準としては，F検定から得られる値や赤池情報量基準 (AIC) などが用いられる．

3.1.3 交叉検証法

交叉検証法（CrossValidation）[4] は，モデルの精度を評価する際によく使用される方法の一つである．特に，学習に使用するデータをk個の部分集合に分割し，各集合を一度はテストデータとして用いて評価する方法をk-fold-CrossValidationという．本研究はこれを採用している．よって以後はk-fold-CrossValidationに焦点を当てて説明する．

k-fold-CrossValidationの概念図を図3.2に示す．k-fold-CrossValidationの目的は，各

テストデータに対し精度評価を行い，その平均をとることで平均精度を得ることである．平均精度は，使用する学習データから得られるモデルの推定精度を意味している．

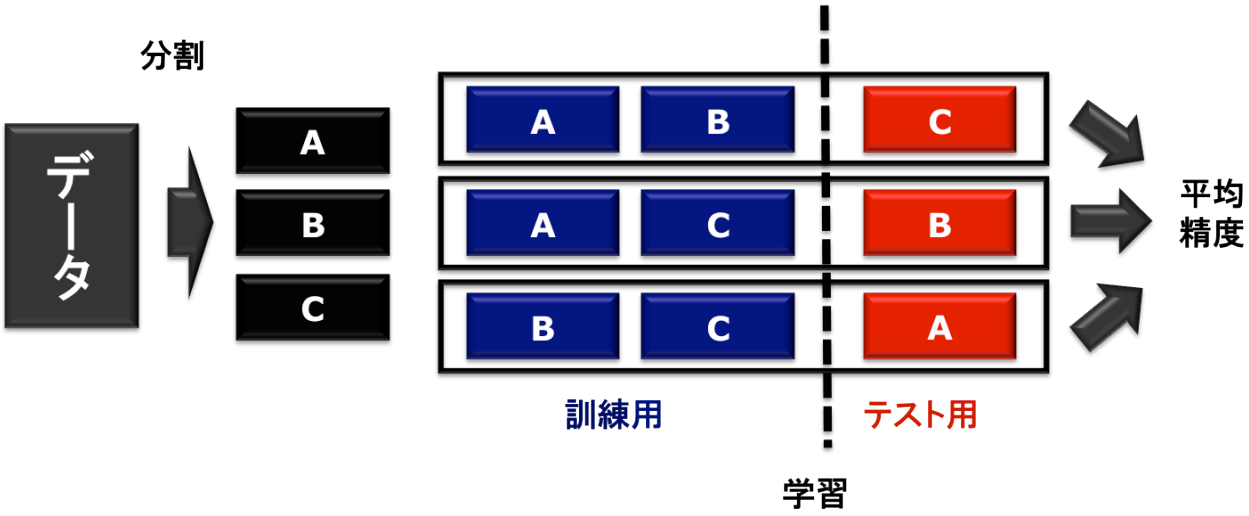

図 3.2: k-fold-CrossValidtion

3.1.4 Gini-Importance

Gini-importance[3] は，分割変数を評価する基準の一つである．その評価方法は，分割前の親ノードにおけるデータ集合と，分割後の子ノードにおけるデータ集合のジニ不純度を比較し，分割変数がどれだけジニ不純度を減少させたかによって行われる．

本研究では，学習用データにおける元のジニ不純度から，決定木の学習によって減少したジニ不純度の合計を 1 とすることで，各分割変数が減少させたジニ不純度の値を割合として表している．これは，各分割変数の目的変数に対する重要度を表している．概念図を図 3.3 に示す．

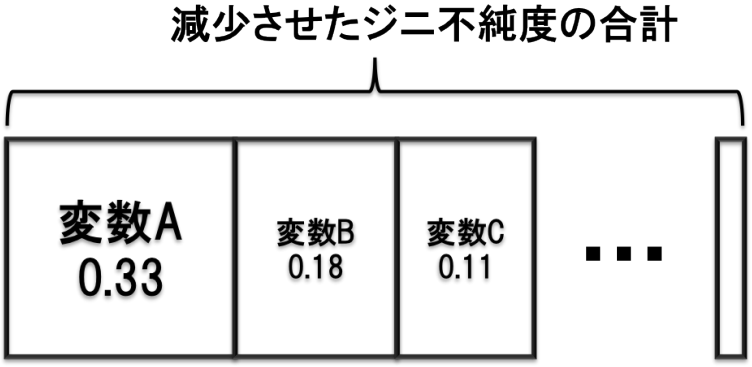

図 3.3: Gini-Importance による変数重要度の表現

3.2 構造学習の行程

3.2.1 構造学習行程の全容

本手法によるグラフ構造学習行程のフローチャートを図 3.4 に示す．これは，MB 法の発想である重回帰分析を全変数に網羅的に行うという方法を利用したもので，決定木による網羅的な分析を表している．

フローチャートの各処理について述べる．N 個の変数 $(x_1, x_2, \cdots, x_N)$ を持つデータが入力されると，ループ処理に入る．このループは，データの持つ全変数の各変数を目的変数とした場合の決定木分析を網羅的に行うものであり，N 回ループ処理が行われる．モデル推定・構築では，x_i 番目の変数が目的変数で，その他が説明変数である場合の決定木学習が行われる．決定木が構築されると，その木における各変数の重要度が算出できる．これにより，隣接行列における i 番目の行が求められることになる．変数重要度は Gini-importance により定義している．

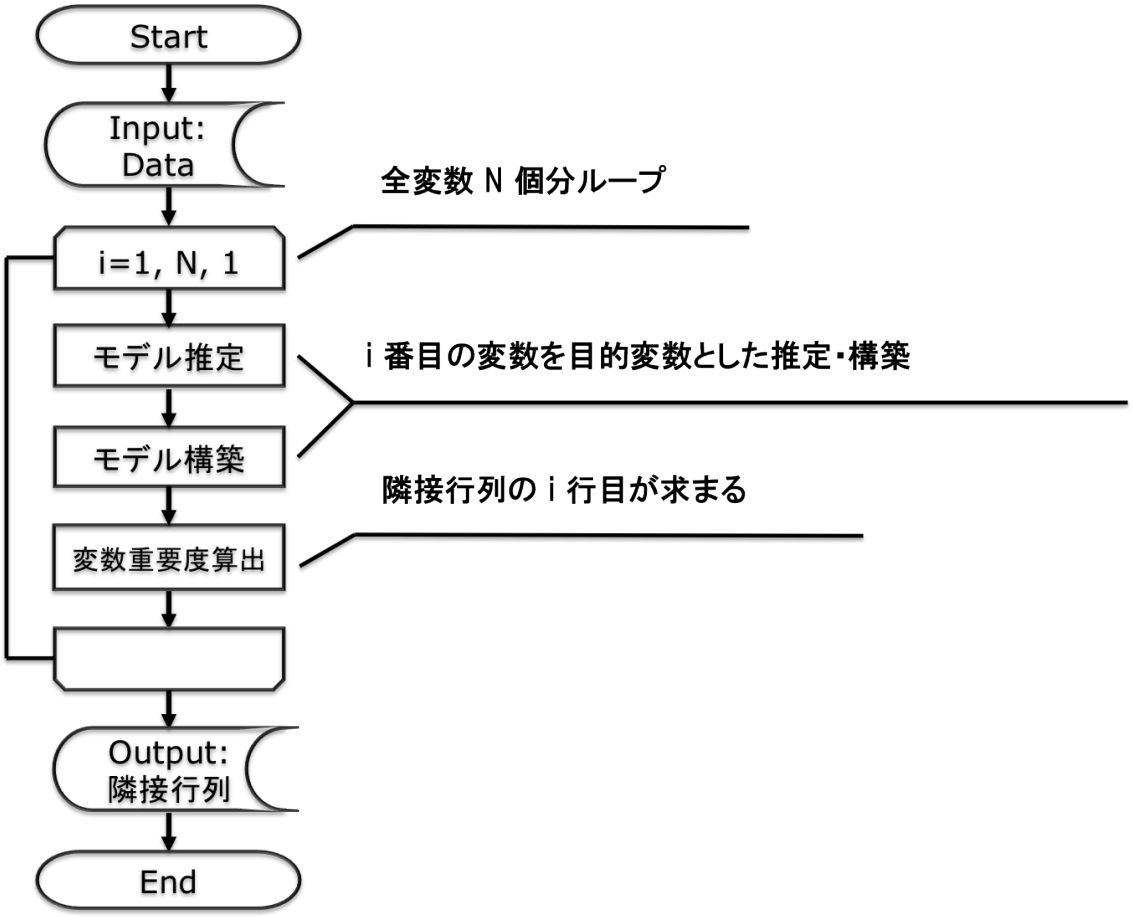

図 3.4: 決定木によるグラフ構造学習のフローチャート

3.2.2 モデル推定

推定には，10-fold-CrossValidation（以降 CV）と変数減少法を用いる．CV は最善のモデルを推定するためであり，変数減少法は重要な変数のみを残す変数選択に用いる．

変数 x_1 が目的変数であった場合のモデル推定の概念図を図 3.5 に示す．これは，CV によって分割された各分割パターンデータにおいて変数減少法を用いて決定木を学習した場合の変数選択の様子を表している．図 3.5 のように，各分割パターンデータからは，それぞれ異なった変数選択の結果が得られる．CV と変数減少法による推定の目的は，図 3.6 に示すように，各分割パターンデータから得られた変数選択の結果を統合したものを得ることである．統合結果は，モデル構築において用いられる．

変数減少法による変数選択の過程を図 3.7 に示す．このフローチャートは，精度が悪化

しない限りは，変数を削減し続けることを示している．入力データは，CV によって分割された各パターンデータである．パターンデータからモデル構築・精度の算出を行うと，初回以外は前回の精度と比較を行う．ここで精度が悪化した場合には，削除した変数を元に戻して処理を終了する．変数重要度は Gini-Importance により定義している．

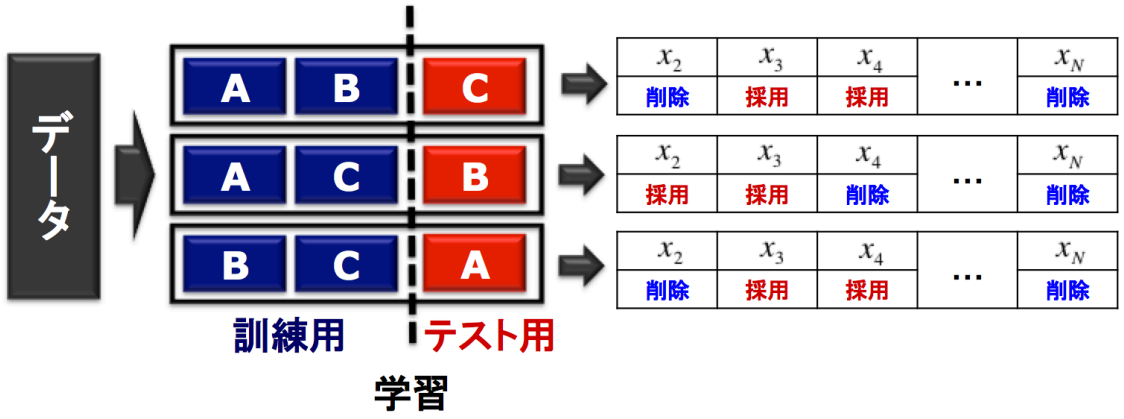

図 3.5: CV と変数減少法によるモデル推定

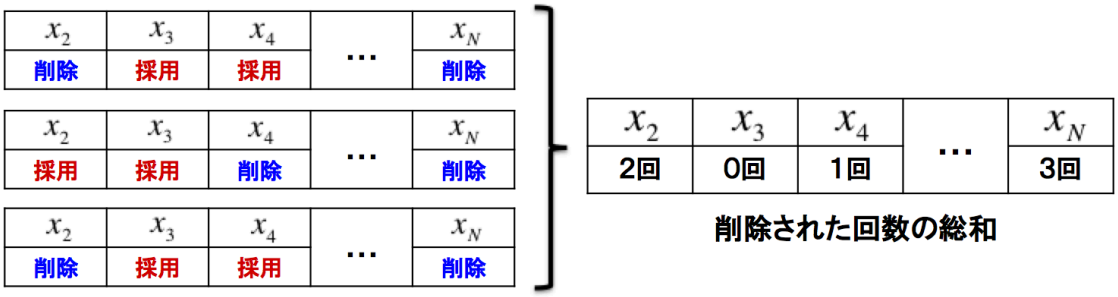

図 3.6: モデル推定の目的

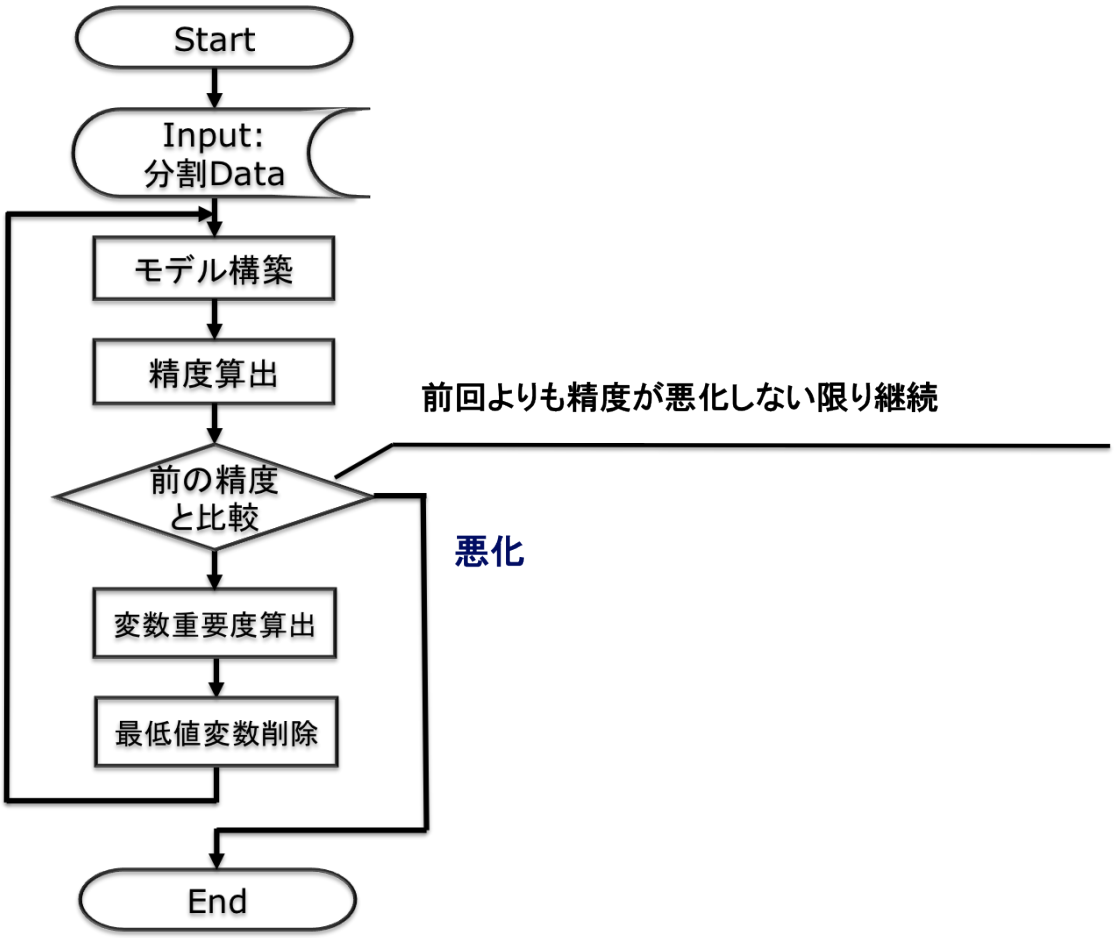

図 3.7: 変数減少法のフローチャート

3.2.3　モデル構築

モデル構築は，先の説明にあった図 3.6 にある統合結果を使用する．統合結果は，削除された回数の総和であるが，本手法は 10-fold-CrossValidation を利用しているので，値は 0〜10 の間をとることになる．これに関して閾値を設定することで，閾値以上削除されたものを学習において説明変数から除外するという方法でモデル構築を行う．

例えば図 3.6 の統合結果に対し，２回以上削除されたものを除外するという条件でモデル構築を行った場合を図 3.8 に示す．図 3.8 のように，閾値を超え削除され多変数はモデルに使用されることはない．モデルは採用された変数でのみ構築される．

閾値の決定方法は，実際に閾値を 1〜9 まで試し，最も精度が良かったものを選択する．0 と 10 については，行う意味がないので試行しない．同率の精度となる場合には，より使用変数が少ない閾値を選択する．

最終的なモデル構築が終了すると，最後に変数重要度の計算を行う．これにより，隣接行列のある一行が求められたことになる．

x_2	x_3	x_4	・・・	x_N
2回	0回	1回		3回
削除	採用	採用		削除

図 3.8: 最終的に構築に使用される変数

3.3　決定木構造とグラフ構造

決定木構造とグラフ構造の対比を図 3.9 に示す．これは変数 A，B，C が存在し，A が目的変数であった場合の決定木とグラフの例を表している．

グラフは有向グラフとなっているが，これは説明変数が目的変数を「説明」するという意味で，C・B から A に向かって矢印が伸びることとなっている．矢印の色の違いは，赤が決定木の頂点ノードに選ばれた変数との関係性を表し，青は頂点以外で使用された変数との関係性を意味している．頂点以降に使用される変数は，決定木アルゴリズムの性質から，以前の変数による分割の影響を受けた条件付きの関係性となる．グラフの重みは，Gini-Importance によって決定されたものである．

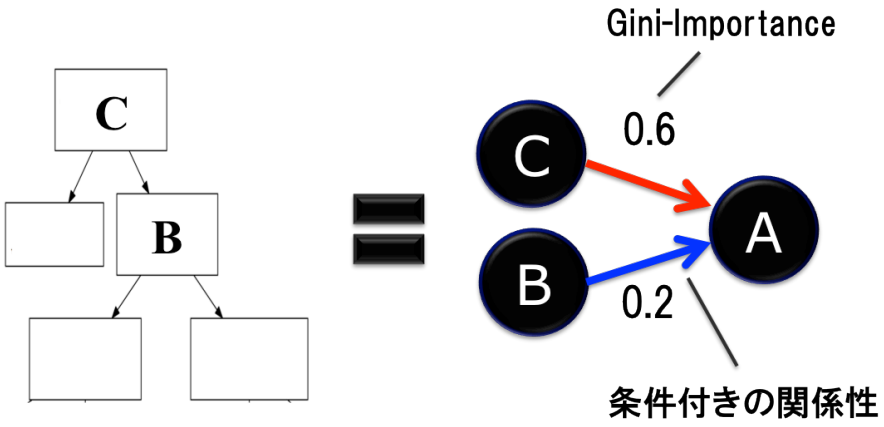

図 3.9: 決定木構造とグラフ構造

第4章　検証実験と考察

本章では，本手法の有効性等を検証実験から明らかにする．

4.1　検証実験の目的

下記の項目について検証を行うことで，本手法の有効性等を明らかにする．

1. 変数減少法による変数選択が有効であるか．
2. 本手法により学習されるグラフに妥当性があるか．
3. 本手法により学習されるグラフは説明変数間の関係を反映できているか．
4. 予測的分析である決定木と相関分析によるグラフの相違点，また決定木による優位点は何か．

1 については，変数減少法を用いない決定木と本手法との結果を比較することにより行う．2 は，学習されたグラフを現実の事実と照らし合わせる事で検証する．3 は 2 で学習されたグラフを参考に評価する．4 は質的データに対する相関分析手法のクラメールの連関係数を用いたグラフと比較することで検証する．

4.2　実装環境

決定木の構築においては python[4]及び機械学習パッケージ scikit-learn[5]中の DecisionTree クラスを利用している．クラメールの連関係数においては統計処理ソフト R を用いた．なお，グラフの作成は python のパッケージ pygraphviz[6]及びグラフ描画ツール Graphviz[7]を利用した．

[4]http://www.python.org/
[5]http://scikit-learn.org/stable/
[6]http://pygraphviz.github.io/
[7]http://www.graphviz.org/

4.3 使用データ

データは，ニコニコ動画が提供する動画データ[8](2012-6-03〜2012-11-01) のタグ情報である．検証用データにおける変数とはタグのことであり，各タグは”0”または”1”の値のみをとる．これは，動画データにおいて，タグが ”存在しない ”または”存在する”ということを意味している．元データはタグ総数 592014 個，サンプルデータ総数 869437 個であるが，あまりにも疎であったために，データスパースネスの問題が発生し，意味のある分析ができなかった．よってデータを分析可能なよう成形してある．元データの詳細については，付録を参照して頂きたい．

4.3.1 データ成形

データスパースネスの問題を解決するため，使用するタグは出現率上位 0.001%のものに絞った．これにより，60 個のタグが抽出できた．次に，データを 60 個のタグのうち，最低 4 つ以上を持つサンプルデータのみに絞った．しかし，まだデータの 99%が”0”というタグが存在していたので，最低 5 %以上は”1”を持つタグのみに絞ると，最終的に 24 個のタグが残ることとなった．使用タグの一覧が表 4.1 である．成形後のデータ概要は以下の通りである．また，データの内容を図 4.1 に示す．

- 変数の数（タグ）：24
- サンプル数：10456

[8]http://www.nii.ac.jp/cscenter/idr/nico/nico.html

表 4.1: 使用タグ一覧

1	ホラーゲーム	13	PS3
2	Ib	14	実況プレイ part1 リンク
3	マインクラフト	15	実況
4	BF3	16	投稿者コメント
5	ダークソウル	17	初音ミク
6	ミクオリジナル曲	18	もっと評価されるべき
7	PC ゲーム	19	実況プレイ
8	FPS	20	ゆっくり実況プレイ
9	縛りプレイ	21	VOCALOID
10	フリーゲーム	22	音楽
11	Minecraft	23	実況プレイ動画
12	ゲーム実況	24	ゲーム

	変数1	変数2	変数3		変数24
	ホラーゲーム	Ib	マインクラフト		ゲーム
動画1	0	0	0		1
動画2	0	0	0		0
動画3	0	0	1		1
動画4	0	0	0		1
動画5	0	0	0		1
動画6	0	0	0		1
動画7	0	0	0		1
動画8	1	0	0		1
動画9	0	0	0		1
動画10	0	0	0		0
動画11	0	0	0		0
動画10456	0	0	0		1

図 4.1: 成形後のデータ

4.4　実験結果と考察

目的と内容の部分で挙げた四つの事柄について実験結果を示し，考察を行う．

4.4.1　変数減少法の有効性

変数減少法の有効性を検証するため，以下の二つの手法を比較した．比較は予測精度と使用変数個数によって行う．

- CV のみを用いた決定木
- CV と変数減少法を用いた決定木（本手法）

検証実験における両者の学習過程について述べる．学習過程は図 4.2 の通りである．順に説明すると，まずは元のデータを学習用と評価用に分割する．ここで評価用とされたデータは学習にまったく関わらない最終評価用のデータとなる．学習用のデータは前述した 10-fold-CrossValidation(CV) によるモデル学習に用いられる．学習後，CV のみの決定木では，CV による分割データの中で最も予測精度の高かったモデルを最終モデルとして選択する．本手法では，モデル構築の部分で述べたように，統合結果から削除された回数が閾値未満のものを用いて最終モデルを構築する．最後に，評価用のデータを用いて最終モデルを評価する．

今回，初期の学習用・評価用データの分割においても 10-fold-CrossValidation を用いている．つまり，二重に CV を行ってモデル評価を行うこととなる．ここで，学習用データと評価用データを分割する部分を外側 CV と呼び，最終モデルを学習するための CV を内側 CV を呼ぶこととする．

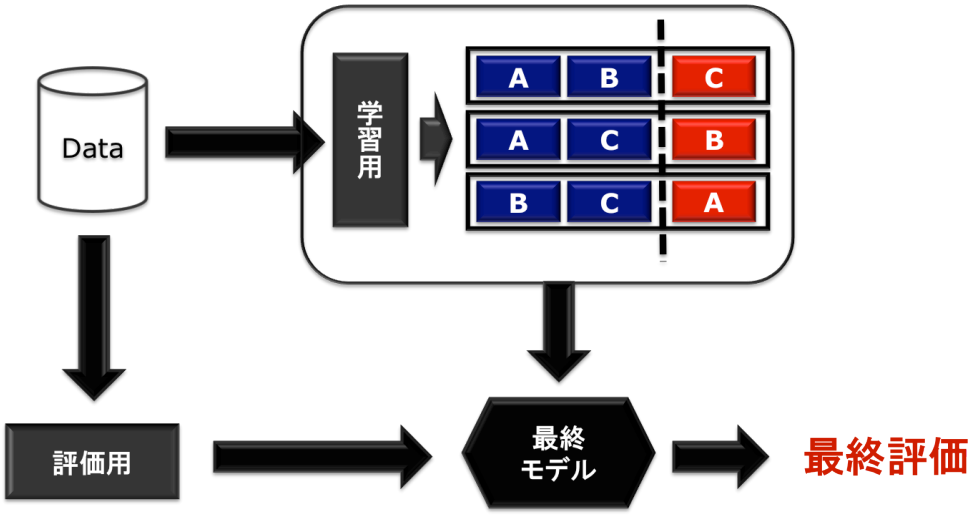

図 4.2: 学習過程

検証用データは 24 個の変数を持つので，外側 CV での最終モデル評価は 24 個のモデル評価の結果を得る．表 4.2 は外側 CV の各分割データにおける両手法の 24 個のモデルの予測精度平均を示している．同様に，表 4.3 は両手法の使用変数個数平均を示している．予測精度平均については，小数点下 3 桁未満を四捨五入し，使用変数個数平均は整数未満を四捨五入している．

表 4.2: 両手法の予測精度平均

	CV のみの決定木	本手法
1	0.949	0.952
2	0.954	0.955
3	0.954	0.956
4	0.961	0.963
5	0.956	0.958
6	0.955	0.956
7	0.948	0.951
8	0.952	0.954
9	0.951	0.952
10	0.943	0.944
平均	0.952	0.954

表 4.3: 両手法の使用変数個数平均

	CV のみの決定木	本手法
1	22	17
2	22	17
3	22	18
4	22	17
5	22	17
6	22	18
7	22	18
8	22	18
9	22	18
10	22	17
平均	22	18

考察

表 4.2，表 4.3 を見ると，本手法が精度を落とさずに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，重要な変数が正しく選択されていることを示しており，その有効性が示されたものと考えられる．

使用変数個数を減らすことは，変数重要度の評価にも影響を与える．Gini-Importance による評価は，減少させたジニ不純度の合計の中で，各変数がどの程度貢献しているかを割合で示すものである．よって，使用される変数が多いと，重要度が各変数に分散してしまい，重要な変数が何であるかが曖昧になってしまう．変数減少法を用いて使用する変数を減らすことは，この問題を避けることに繋がり，より変数重要度を考慮した分析を可能とする．これは価値あるグラフを学習する上で有効と考えられる．

4.4.2　学習されたグラフの妥当性

学習されたグラフを図 4.3 に示す．このグラフは先述の変数減少法の有効性を確かめる実験において，予測精度が最良であった 4 番目の分割データにおいて学習されたものである．また，このグラフは 0.15 以上の変数重要度を持つ関係性のみを表すように表現の閾値を設定している．

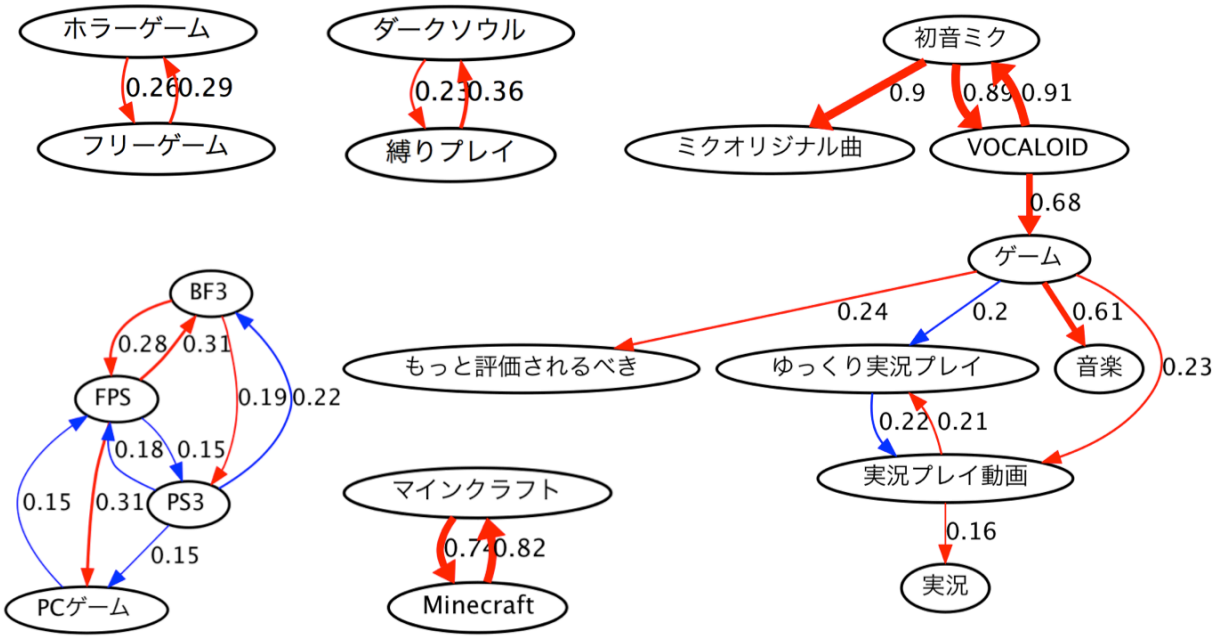

図 4.3: 検証用データから学習されたグラフ

考察

図 4.3 から，グラフがいくつかのクラスタに分かれているのがわかる．まず，右半分を占める最大クラスタの内容から考察する．これを見ると，クラスタの上側は「初音ミク」を中心とした VOCALOID 関連のタグが強い関係性で結びついていることがわかる．次に下側に目を移すと，「実況」という言葉が並んだタグが結びついていることがわかる．「初音ミク」，また「実況」という概念を中心にしたそれぞれのクラスタは，似た概念同士の結びつきにより出来ており，現実の認識と一致する．これは，グラフの妥当性を示すものと考えられる．

次にグラフの左半分について考察する．左半分は，全てゲームに関するタグである．こ

こで興味深いのは，全てゲームに関するタグでありながら，それぞれ異なるジャンルで纏まっているということである．例えば「ダークソウル」は，PlayStation3 のゲームでありながら「PS3」タグとは結びつかず,「縛りプレイ」というタグと結びついている．これは,「ダークソウル」が PS3 のゲームであることは周知であり，載せる必要の無い情報と判断されているか，あるいは「縛りプレイ」というゲームプレイジャンルと「ダークソウル」に切り離せない強い関係性があることを意味している.「フリーゲーム」というジャンルと「ホラーゲーム」が結びついていること，また「FPS」というジャンルと「PS3」が結びついていることにも同様のことがいえる．

このこと以外にも,「ゲーム」という最も中心的なタグが，どのジャンルのクラスタとも結びついていないのは興味深い．これは,「ゲーム」タグの意味が抽象的で，あらゆるジャンルのゲームにのせられるために，どれか一つと強い関係性を持ち得ないか，あるいはどのジャンルも「ゲーム」であることが周知であるためにタグとして登録されないといった可能性が考えられる．

これらのことから，本手法が学習するグラフは，個々のタグが持つ特徴を上手く反映できており，妥当なグラフを学習できているものと考えられる．

4.4.3 説明変数間の関係性

図 4.3 のグラフでは，いくつかの変数間関係に青のエッジが存在している．これは，決定木における頂点以外，つまり第二層以降で使用された説明変数と目的変数との関係である．これは説明変数間に存在する階層的関係をグラフに反映したものであり，この情報を考慮したグラフ分析を行うことで，ロジスティック回帰分析から学習されたグラフでは発見できなかった知識を発見できる可能性が示唆される．

4.4.4　予測的分析と相関分析の比較

予測的分析と相関分析のそれぞれから学習されるグラフを比較し，相違点と本手法の優位性を明らかにする．相関分析で用いる手法は，質的データに対する手法であるクラメールの連関係数を用いた．両者から学習されたグラフが，それぞれ図 4.4 と図 4.5 である．比較のため，グラフの特定部分を抽出し，両者とも構造が完全グラフになるようにした．図 4.4 は図 4.3 のグラフと同様の学習結果を用いている．

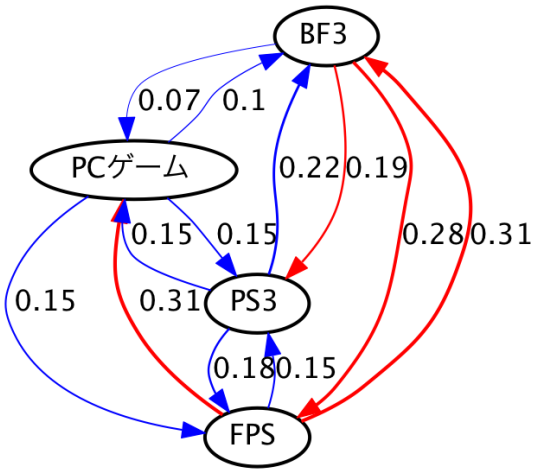

図 4.4: 本手法によるグラフ

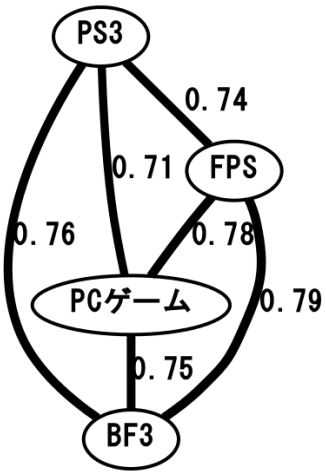

図 4.5: クラメールの連関係数によるグラフ

考察

両者の違いを生み出す要因は，多変量分析か単変量分析であるかという違いである．決定木は多変量分析であり，クラメールの連関係数は単変量分析である．

単変量分析では，使用される変数は２つであり，純粋にその２つの関係性だけを分析して，それ以外を考慮しない．よって学習される行列は対称行列となり，無向グラフとなる．

多変量分析では，目的変数とそれに対する説明変数間の条件付け関係を含めた複雑な関係性を考慮する．また，Gini-Importance による評価は，各説明変数がどれだけ目的変数の予測に貢献しているかの割合であり，序列と割合は目的変数と説明変数の関係、そして説明変数の組合せによって変化する（図 4.6）．よって，行列は非対称なものとなり，目的変数と説明変数という依存関係から，グラフは有向グラフとなる．

これらの違いは，獲得できる知識の違いに繋がる．単変量分析では，変数ペアの関係性

という知識しか得られない．それに対し，多変量分析である決定木からは，複数の説明変数と目的変数を絡めた if-then ルールを得る事が出来る．よって知識発見という意味では，本手法の方が，より高度な知識を発見できる可能性がある．

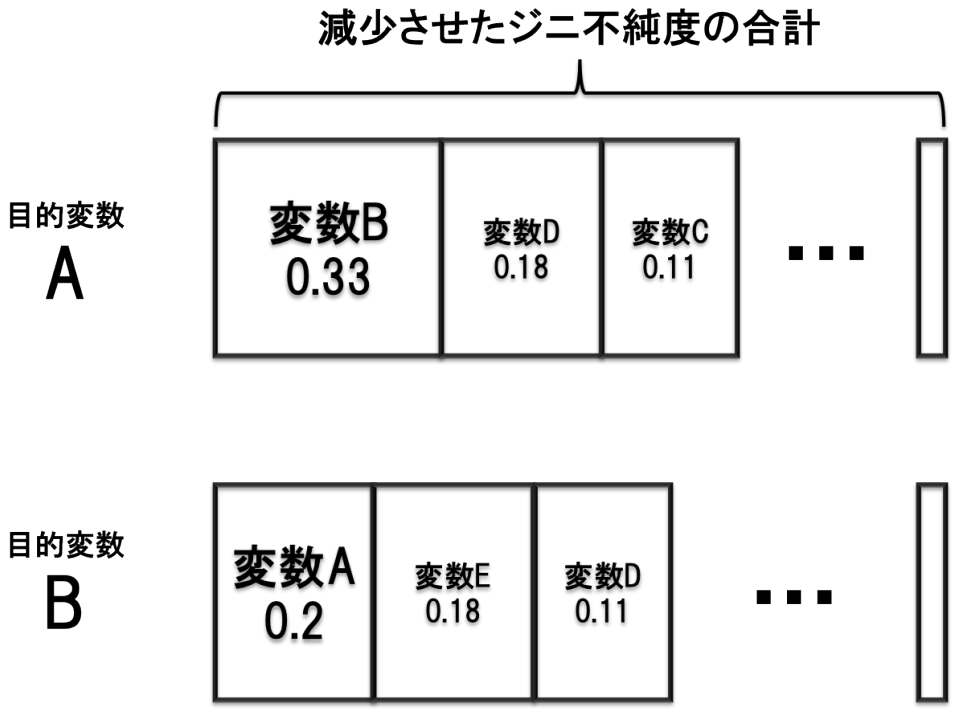

図 4.6: 目的変数による変数重要度の変化

第5章　結論と今後の課題

5.1　まとめ

本稿では，質的データからグラフ構造を学習する手法を提案し，その概要と有効性について述べた．

質的データからのグラフ構造学習における既存研究の M.Wainwright らの手法について，ロジスティック回帰分析を用いた分析があらゆる場合において最適ではないことを述べ，決定木を用いたグラフ構造学習手法を提案することを述べた．

本手法の有効性を確かめる検証実験では，決定木における変数選択に変数減少法が有効であることを示した．学習されたグラフの妥当性については，似たような概念であっても，それぞれの特徴を正しく反映させた関係性が構築されていることを示し，グラフ全体としてもニコニコ動画内におけるタグの関係性を概ね正しく反映できていることが確認出来た．学習されたグラフは，説明変数間の関係性を反映しており，それを考慮したグラフ分析により，新たな知識を発見できる可能性が示唆された．また，本手法から得られるグラフが，相関分析によって得られるグラフよりも，高度なルールを反映していることを示し，知識発見の点で相関分析よりも優位であることを示した．

5.2　今後の課題

5.2.1　変数選択手法について

本稿では，変数減少法の有効性を示したものの，変数減少法による結果はロバストなものでないという指摘がされている．よって変数選択手法については再考の余地がある．

今後目指すべきは，Lasso 正則化を指向した変数選択手法を新たに開発することであるが，数学的な理論付け等，様々な困難が予想される．

5.2.2　変数重要度の評価について

本手法は，Gini-Importance によって変数重要度を評価しているが，これ以外にも変数重要度を評価する方法はある．例えば，説明変数の一つを隠すことで予測精度がどの程度落ちるか，といった基準から変数重要度を評価することも可能である．よって，変数重要度の評価法について，どのような評価法がどのような目的に対して有効であるのか等を検証する必要がある．

5.2.3　多重共線性の問題について

重回帰分析において，説明変数間に強い相関がある場合，結果がロバストなものとならないことが指摘されており，多重共線性の問題と呼ばれている．本手法は決定木によるものであるが，同様の問題を抱えているものと考えられる．よって，結果をより信頼度の高いものとするためにも，決定木において多重共線性の問題をどう解決するのかを考える必要がある．

参考文献

[1] Rich Caruana and Dayne Freitag. Greedy attribute selection. In *ICML*, pages 28–36. Citeseer, 1994.

[2] Tsuyoshi Idé, Aurelie C Lozano, Naoki Abe, and Yan Liu. Proximity-based anomaly detection using sparse structure learning. In *SDM*, pages 97–108, 2009.

[3] Alan J Izenman. *Modern multivariate statistical techniques: regression, classification, and manifold learning*. Springer, 2008.

[4] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, pages 1137–1145, 1995.

[5] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

[6] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[7] Martin J Wainwright, John D Lafferty, and Pradeep K Ravikumar. High-dimensional graphical model selection using ℓ_1-regularized logistic regression. In *Advances in neural information processing systems*, pages 1465–1472, 2006.

[8] 奥喜正, 本村猛能, 前鶴政和, and 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**, 44:1–14, 2004.

[9] 家村俊一郎 and 夏目徹. タンパク質のネットワーク解析から創薬へ. *Synthesiology*, 1(2):123–129, 2008.

[10] 悟, 藤重, 正夫, and 伊理. **グラフ・ネットワーク・組合せ論**. 共立出版, 2002.

[11] 林知己夫. **数量化*:* 理論と方法**. 朝倉書店, 1993.

[12] 池谷智行 and 村田剛志. 複数種ノードネットワークからのコミュニティ抽出. *2008* **年度人工知能学会全国大会** *(***第** *22* **回***)* **論文集**, 3, 2008.

[13] 浩, 元田, 周作, 津本, 高平, 山口, and 情報処理学会. **データマイニングの基礎**. オーム社, 2006.

[14] 金英子, 松尾豊, and 石塚満. Web 上の情報を用いた企業間関係の抽出. **人工知能学会論文誌**, 22(1):48–57, 2007.

[15] 鹿島久嗣. ネットワーク構造予測. **人工知能学会誌**, 22(3):344–351, 2007.

謝辞

本研究を進めるにあたり，主指導教員である北陸先端科学技術大学院大学　知識科学研究科 Dam Hieu Chi 准教授には大変お世話になりました．私はデータマイニングに関して全く無知なところから研究を始めましたが，Dam 先生によるデータマイニング・統計の勉強会，またゼミなどによる手厚い研究指導により，何とか研究活動を全うすることができました．具体的な研究活動以外にも，研究に対する姿勢や哲学，また人としてどう生きるべきか，といったことまで幅広くご指導頂き，人間として成長できたことを感じております．本当にありがとうございました．深く感謝しております．

日々の学生生活や，各種文書の校正などでは，北陸先端科学技術大学院大学　知識科学研究科　杉山　歩　助教授に大変お世話になりました．研究発表における良いスライドの作り方や，研究文書の書き方など，研究活動に必須の知識について色々と教えて頂きました．また，学生生活において有用な情報を多く教えて頂き，そのお陰で有意義な学生生活を送ることができました．深く感謝しております．

研究に詰まった時のアドバイスや，各種文書の校正などでは，北陸先端科学技術大学院大学　マテリアルサイエンス研究科　水上　卓　助教授に大変お世話になりました．杉山先生と同様に，水上先生からも研究活動に必要な知識，心構えなど様々なことを教えて頂きました．深く感謝しております．

研究活動並びに学生生活において，研究室の皆様や友人達にお世話になりました．良い研究室メンバー，友人に恵まれたことで，非常に充実した大学院生活を送る事が出来ました．深く感謝しております．

最後に，あらゆる面で学生生活を支援してくれた家族に深く感謝しております．

付録

ここでは，本研究の検証実験に用いたデータの元データについて説明する．なお，ここでの説明は国立情報学研究所情報学研究データリポジトリ[9]のニコニコデータセットに存在するデータ仕様書からの引用である．

動画情報 JSON

動画情報 JSON は下記のキーと値を持つ．

キー	値	値の型	値の例
video_id	動画ID	String	"sm9"
thread_id	スレッドID	Number	1173108780
title	動画タイトル	String	"新・豪血寺一族 -煩悩解放 - レッツゴー！陰陽師"
description	動画説明文	String	"レッツゴー！陰陽師（フルコーラスバージョン）"
thumbnail_url	サムネイル画像のURL	String	"http://tn-skr2.smilevideo.jp/smile?i=9"
upload_time	投稿日時 (ISO 8601 形式)	String	"2007-03-06T00:33:00+09:00"
length	動画再生長 (秒数)	Number	319
movie_type	動画フォーマット	String	("flv", "swf", "mp4")
size_high	高画質動画のファイルサイズ (byte)	Number	21138631
size_low	低画質動画のファイルサイズ (byte)	Number	17436492
view_counter	動画の再生数	Number	10133072
comment_counter	コメント数	Number	4056043
mylist_counter	マイリスト登録数	Number	130221
last_res_body	最近のコメントのサマリ	String	"遺影!!! 遺影☆ 遺影! 遺影☆ 列豪!! 陰陽師☆ (ﾟ∀ﾟ)o彡°どーまん！せーまん！(ﾟ∀ﾟ)o..."
tags	後述のタグ情報 JSON の配列	Array	[{ "tag": "陰陽師", "lock": 1 }, { "tag": "ゲーム", "category": 1, "lock": 1 }, ...]

図 5.1: 動画情報 JSON

[9]http://www.nii.ac.jp/cscenter/idr/index.html

タグ情報 JSON

タグ情報 JSON は下記のキーと値を持つ．ただし、lock と category キーについては、存在しない場合がある．

キー	値	値の型	値の例
`tag`	タグ文字列	String	"陰陽師"
`category`	タグがカテゴリ指定されていればキーが存在 (値は常に1)	Number	1
`lock`	タグがロックされていればキーが存在 (値は常に1)	Number	1

図 5.2: タグ情報 JSON

本研究に関する発表論文

国内会議（査読なし）

1. 川崎　隆史，鈴木　大輔，杉山　歩，Dam Hieu Chi，決定木を用いた質的データからのグラフ構造学習，情報処理学会第 76 回全国大会，発表予定 (2014/3/11)