平成27年度　修士論文

教育支援を目的とした要注意学生の推定精度改善法

(Estimation Accuracy Improvement Methods of Students Requiring Guidance for Education Support)

指導教員

舟橋　健司　准教授

伊藤　宏隆　　助教

名古屋工業大学大学院　工学研究科

博士前期課程　情報工学専攻

平成23年度入学　23417596番

西脇　雅弥

目 次

第1章　はじめに

情報通信技術の発展により，マーケティングにおける顧客情報や医療機関における臨床データなど，様々な分野で電子的にデータを扱うことが増えている．このことは，大量のデータの保持や参照速度の向上に寄与したが，近年ではそれだけでなく，このように蓄積されたデータからデータマイニング技術によって新たな知識・傾向を見出す手法が数多く報告されている．先に触れたマーケティングと医療の例では，購買履歴などから顧客が好むと推測される商品を推薦するレコメンデーションや，既存薬の効果から新薬の副作用を予測する研究 [1] があり，他にも，自律型ロボットの行動制御 [2][3] や，電力需要予測への応用 [4][5] も期待されている．少し例を挙げただけでも，データマイニングの応用分野は非常に幅広いことが分かる．

データマイニングは，大きく分けて「相関分析」,「分類」,「クラスタリング」,「外れ値検出」の 4 つの技術から成り，そのうち「分類」にカテゴライズされるベイジアンネットワークは，既知の情報から未来を予測する手法としてよく用いられる手法である．ベイジアンネットワークは，生体の神経細胞網を模して作られた人工ニューラルネットワーク [6] の派生形と見ることができ，現に脳機能をベイジアンネットワークを用いて再現しようとする試み [7] まである．

教育現場でも，このベイジアンネットワークを活用して学生に関するデータから学生一人ひとりの修学傾向を読み取り，修学指導に援用するという提案がある．具体的には，各学生の成績と授業アンケートをもとに，成績と受講態度の関係，成績と教員の授業の進め方との関係などを調査した研究 [8] や，e-Learning の学習履歴から，学習者一人ひとりに適した教育ナビゲーションの構築を目指す研究 [9] などがある．

名古屋工業大学（以下，名工大）においても，早期の修学指導を目的とした双方向型教育支援システムを構築している [10]．このシステムとして，名工大の情報基盤センターは，コースマネジメントシステム（Course Management System, CMS）と IC カードによる出欠システムを開発・運用している．CMS は Web 技術を使って e-Learning を支援するシステムである．例えば，レポートの提出管理，小テストの実施，教材の作成支援などを行うことが可能である．CMS によって，課題提出状況や小テストに対する評価が電子的にサーバに保持される仕組みが整っている．一方，IC カード出欠システムは，講義の開始前と終了後に，学生が IC カード学生証を講義室に設置されたカードリーダにかざすこと（以下，打刻）で得られる時刻をサーバに保持するシステムである．教員はこの時刻情報にアクセスして，出欠判定に利用する．

ところで，多くの大学にとって，留年・退学者の存在は大きな問題となっている．学生が留年・退学してしまう原因としては，経済的理由や病気・怪我などのやむを得ない事由，転学などの積極的な動機づけによるもの，大学生活への不適応，就職・大学院入試の失敗による計画的留年などが挙げられる．こういった学生への措置として，学生と教職員が直接向かい合って，学習面や生活面のアドバイスや相談をする指導方法が多くの大学でとられている [11]．しかし，この方法は一人の教職員に対する学生の数が多い場合，教職員側の負担が大きくなることが課題となっており [12]，結果的に充分な指導が行えない可能性がある．この負担軽減を目的に，ベイジアンネットワークを用いた確率推論によって講義の進め方を提案する研究 [13] もなされているが，実験対象の講義はただ 1 つであり，他の講義への対応は検証されていない．

そこで，学生が履修した講義すべての成績データと打刻データを用いて，留年・退学する学生を調査し「要注意学生」を定義・予測する研究 [14] が行われた．予測により指導対象者を絞り込むことで，指導にかかる時間的コストを削減している．また，予測された学生は要注意学生であるので，指導方法も判断しやすい．この研究では，Grade Point Average（以下，GPA）を用いて，要注意学生の定義を行っている．GPA が 1 年次前期または 1 年次後期において 1.0 未満の学生は．指導が必要であることは明白であるため，予測を行うことなく指導対象者とする．このことを踏まえ，GPA が 1 年次前期，1 年次後期においてともに

1.0 以上であるにも関わらず，将来的に留年・退学してしまう学生を要注意学生と定義している．予測にはベイジアンネットワークを用いており，この手法により 122 名の指導対象者を抽出し，留年・退学する学生の 81.4%を全学生 338 名を指導する場合の約 3 分の 1 の時間的コストで予測できている．これは，ベイジアンネットワークによる要注意学生の予測の有用性を示している．

一方で，この研究の問題点は，要注意学生の定義にあった．GPA による機械的な定義では，指導を必要としている学校に馴染めない学生や学業に不安がある学生と，指導をあまり必要としていない転学や計画的留年などをする学生が区別できない．これでは，本当に指導を必要としている学生を見つけることができず，留年・退学をしてしまうかもしれない．この問題に対処するため，要注意学生の定義を見直し，予測を行う研究 [15][16] がなされた．この研究では，要注意学生を「GPA が 1 年次前期，1 年次後期において共に 1.0 以上であり，なおかつ今後消極的理由により留年・退学する学生」と定義し，推定を行っている．

本研究では推定精度の更なる向上を目指し，新たな変数である「科目属性（必修・選択）を考慮した変数」,「授業クラスにおける成績の偏差値を考慮した変数」,「成績と出席の関連付けによる変数」を導入した．これにより，1 年次前期終了時，1 年次後期終了時，2 年次後期終了時の 3 つの時期において要注意学生の推定精度改善が確認できた．また，従来研究においては，過去のある年度 A, B に名工大に入学した学生のデータを一括りに扱っており，入学年度別の予測精度の検証や，学生の傾向の違いに対する汎化能力の検証は行ってこなかった．そこで，検証法として交差検証法だけではなくホールドアウト検証も採用することで，予測モデルの汎化能力を確かめた．その結果，ある年度に入学した学生のデータを用いて推定モデルを構築し，その次年度に同じ学科に入学する学生が要注意学生になるかどうかの推定を行うときの推定精度は，同年度に入学した他群の学生（例えば，他学科の学生）が要注意学生になるかどうかの推定を行うときの推定精度よりも高いという知見を得た．

本論文では，まず第 2 章でデータマイニングによる知識発見法について，第 3 章では予測に利用するデータの概要とその変換方法，推定精度改善のための新変数の生成について示す．次に第 4 章でベイジアンネットワークによる要注意学生の予測と検証について説明し，最後に第 5 章では本研究のまとめと今後の展望を述べる．なお研究を進める上で学生のデータを扱う際には，氏名や学生番号といった個人を特定できる情報は除いてある．そのため，本研究によって個人情報が侵害されることがないことをここに明記する．

第2章　知識発見の技術

データマイニングは，データベースからの知識発見（knowledge discovery in database, KDD）[17] とも呼ばれ，そのプロセスは図 2.1 で概観することができる．

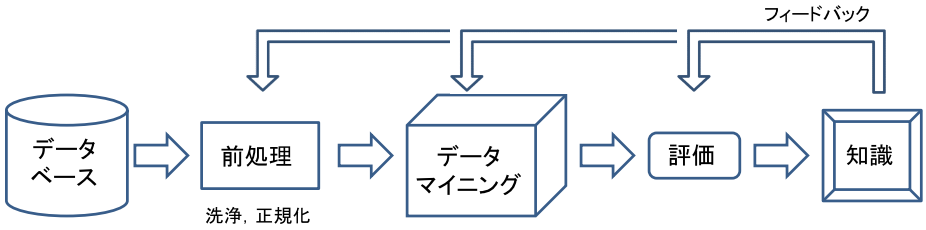


図 2.1: KDD プロセス

全体の流れは大きく「データの前処理」,「データマイニング」,「結果の評価による知識獲得」の 3 段階に分かれる．

前処理

データベースにあるデータを，データマイニングに適する形に加工するためのプロセスで，次の 3 つが主である．

洗　浄 · · · データ中の欠損値やノイズ・重複を除去する．

正規化 · · · データの単位や構造などを統一し，管理しやすくする．

選　択 · · · 良い結果を得るために，必要なデータを選び出す．

データマイニング

データマイニングと一口に言っても，その技術は大きく分けて 4 つある．

相関分析

　複数のデータ間の相関ルールを発見すること．例えば，顧客が購入した商品の中に頻繁に発生する商品の組み合わせから相関ルールを作り，他の顧客に対してそれらの商品を同時に提示するレコメンデーションなどの応用がある．

分類

　データが所属すべきクラスが既知のデータから分類器を作り，それをもとに未知のデータがどのクラスに属するかを判定する技術で，決定木やベイジアンネットワーク，サポートベクターマシン（Support Vector Machine, SVM）などがこれにあたる．分類はスパムメールの検出などに用いられる．

クラスタリング

データ間の距離をもとに，未知のデータ中で類似しているもの同士を同じグループにまとめること．クラスタリングはウォード法に代表される階層的クラスタリングと，k-means 法に代表される非階層的クラスタリングに大別され，マーケティングなどに応用される．

外れ値検出

外れ値検出手法には統計モデルや距離などを用いた方法があり，クレジットカードの不正利用検出，新薬の副作用推定などに利用される．

評価

データマイニングによって見出されたパターンを評価・解釈し，可視化することで知識を得る．

KDD プロセスは必ずしも一回の処理で終わるものではなく，結果をフィードバックして改善を試みることが一般的である．データ概要については次章で述べることとし，本章ではデータ選択手法である特徴選択と，数あるデータマイニング手法の中でクラスタリング，ベイジアンネットワークについて説明する．また，ベイジアンネットワークモデルの出力評価手法として交差検証とホールドアウト検証，分類問題について述べる．

2.1　特徴選択

特徴選択（Feature Selection）は，複数のデータの中から有用なものを選択，または合成する手法である．データの次元が大きすぎると，有用でない属性がノイズを引き起こし，データマイニングの有用性が失われる可能性がある．そこで特徴選択を行い，データを選択，または合成することで結果の向上が期待できる．選択する手法には大きく分けてフィルター法とラッパー法がある [18]．ラッパー法は学習評価結果を用いるため選択精度の面では優れているが，処理速度の面で実用的ではない．そこで本節では，フィルター法の特徴選択アルゴリズムの一つである，相関ベース特徴選択（Correlation-based Feature Selection, CFS）[19] について述べる．また，合成の代表的手法である主成分分析（Principal Component Analysis, PCA）についても触れる．

2.1.1　CFS

本研究では多くの変数を用いて予測を行うが，先述したようにデータの次元が大きいほど良い結果が得られるとは限らず，逆に悪影響をもたらす場合もある．この「次元の呪い」の影響を軽減するため，多くの変数の中から有用な変数を選ぶ必要がある．CFS はカルバック・ライブラー（KL）ダイバージェンス [20] を用いた特徴選択手法である．KL ダイバージェンスは，2 つの確率分布の距離を計測するために用いられる．ここで，確率変数 X に関する確率分布を P とする．また，X に関するエビデンス $\boldsymbol{e}$ を得たときの条件付き確率 Q とすると，P の Q に対する KL ダイバージェンスは，

$$D(P \parallel Q) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)} \tag{2.1}$$

と定義できる．式 2.1 は，言い換えれば交差エントロピー $H(P,Q)$ からエントロピー $H(P)$ を差し引いたものであり，確率変数 X に関してエビデンス $\boldsymbol{e}$ から得られるエントロピーの平均を表している．このことから，KL ダイバージェンスは情報利得（Information Gain）とも呼ばれる．なお KL ダイバージェンスは，距離の公理である対称性と三角不等式が成り立たないため，数学的な意味での距離とはいえない．

CFS は，目的変数 Z と相関が強い変数を選択する際の指標として有用である．ここで説明変数 Y になりうる変数の候補数を k, 選択された説明変数群の集合を Ω とすると，CFS は

$$CFS \quad Score(\Omega) = \frac{\sum_{i=1}^{k} SU(Y_i, Z)}{\sqrt{k + \sum_{i=1}^{k} \sum_{j \neq i, j=1}^{k} SU(Y_i, Y_j)}} \tag{2.2}$$

と定義される．式 2.2 の分子の値は目的変数と説明変数の相関の強さに比例している．また，分母の値は説明変数間の相関の強さに比例し，集合 Ω 内の冗長度を測ることができる．すなわち，CFS が最大となるような変数を選択すれば，目的変数と相関が強く，なおかつ冗長な情報を除いた変数群 $Y_i \in \Omega$ を抽出することができる．ただし，SU(Symmetrical Uncertainty) は

$$SU(Y, Z) = 2\left[\frac{D(Y \parallel Z)}{H(Y) + H(Z)}\right] \tag{2.3}$$

で定義される．

2.1.2　主成分分析

主成分分析 [21] は多変量データを合成することによって新しい変数を生成する手法で，Karhunen-Loève 変換とも呼ばれる．この新しい変数は主成分と呼ばれ，主成分に解釈を与えることによって，特徴の分析を行うことができる．ここで，主成分分析の概要を説明するために，p 次元ベクトル $\boldsymbol{x} = (x_1, x_2, \cdots, x_p)^{\mathrm{T}}$ からなる情報を，m 次元ベクトル $\boldsymbol{z} = (z_1, z_2, \cdots, z_m)^{\mathrm{T}}$ に縮約することを考える．いま，$k = 1, 2, \cdots, m$ に対して

$$\sum_{i=1}^{p} {l_{ki}}^2 = 1 \tag{2.4}$$

で定義される $\{l_{ki}\}$ を (k, i) 成分に持つ行列 $\boldsymbol{L}$ を考えたとき，

$$\boldsymbol{z} = \boldsymbol{L}\boldsymbol{x} \tag{2.5}$$

を $\boldsymbol{x}$ の主成分と呼ぶ．また，$\boldsymbol{z}$ の各成分 $z_1, z_2, \cdots, z_m$ をそれぞれ第 1 主成分，第 2 主成分，$\cdots$，第 m 主成分と呼び，第 1 主成分から順に計算される．

第 1 主成分の計算は，次のように行われる．まず，ベクトル $\boldsymbol{x}$ の情報をできる限り z_1 に縮約するため，

$$\mathrm{Var}[z_1] = \sum_{i=1}^{p} l_{1i} \sum_{j=1}^{p} l_{1j} \mathrm{Cov}[x_i, x_j] \tag{2.6}$$

を最大化する係数 $\{l_{1i}\}$ を求める．これは，ラグランジュの未定乗数 λ を用いた式

$$Q = \sum_{i=1}^{p} l_{1i} \sum_{j=1}^{p} l_{1j} \mathrm{Cov}[x_i, x_j] - \lambda \left(\sum_{i=1}^{p} {l_{1i}}^2 - 1 \right) \tag{2.7}$$

に対して，

$$\frac{\partial Q}{\partial l_{1i}} = 0 \tag{2.8}$$

を解くことで計算できる．ここで，ベクトル $\boldsymbol{x}$ の分散共分散行列 $\boldsymbol{\Sigma}$ を用いて i について式 2.8 を書き下すと，

$$(\boldsymbol{\Sigma} - \lambda \boldsymbol{I})\, \boldsymbol{l_{1i}} = \boldsymbol{0} \tag{2.9}$$

が得られる．ただし，$\boldsymbol{I}$ は p 次の単位行列である．この式において，未知数 $\{l_{1i}\}$ がすべて 0 の場合が自明な解となるが，これでは $z_1 = 0$ となり不適である．よって，$k = 1, 2, \cdots, p$ に対して係数行列 $(\boldsymbol{\Sigma} - \lambda \boldsymbol{I})$ が非正則になる $\{\lambda_k\}$ を求める必要がある．これは固有方程式

$$\det(\boldsymbol{\Sigma} - \lambda \boldsymbol{I}) = 0 \tag{2.10}$$

を解くことで計算でき，この $\{\lambda_k\}$ を $\boldsymbol{\Sigma}$ の固有値と呼ぶ．$\{\lambda_k\}$ のうち，その値が最大となる固有値に対応する固有ベクトル $\{l_{1i}\}$ が係数となるとき，$\mathrm{Var}[z_1]$ が最大となる．

また，主成分分析では各主成分同士が無相関となるように定められるので，$k = 2, \cdots, m$ に対しても $\frac{\partial Q}{\partial l_{ki}} = 0$ が成り立ち，$\{l_{ki}\}$ を行列 $\boldsymbol{\Sigma}$ の固有ベクトルとできる．このように求めた固有値を大きい順に並べて，最大固有値に対応する固有ベクトルが第 1 主成分，その次に大きい固有値に対応する固有ベクトルが第 2 主成分，$\cdots$ というように m 個の主成分を決めることができる．ただし，生成された主成分をすべて用いる必要はなく，元のデータの特徴を一定以上表現できればよいため，累積寄与率という基準を用いて必要となる主成分を採択する．第 m 主成分までの累積寄与率は，

$$C_m = \frac{\sum_{k=1}^{m} \lambda_k}{\sum_{i=1}^{p} \mathrm{Var}[x_i]} \tag{2.11}$$

で表され，一般に累積寄与率が 70% –80% 以上となるように主成分を選択する．

2.2　クラスタリング

クラスタリング [22] とは，データ以外にあらかじめ基準を設定することなく，データの集まりをいくつかのグループに分ける方法のことである．クラスタリング手法には大きく分けて 2 種類あり，一つは階層的クラスタリング，もう一つは非階層的クラスタリングと呼ばれる．階層的クラスタリングは，各データ対に類似度あるいは非類似度を定義したのち，この類似度に基づき 1 つのクラスタになるまで逐次結合するという手法である．一方で非階層的クラスタリングは，あらかじめ設定しておいたクラスタに各データを割り当てる手法をとる．本節では階層的クラスタリングの例としてウォード法，非階層的クラスタリングの例として k-means 法を取り上げ，概説する．なお，アンケートデータや販売時点（Point of Sales, POS）データなどにおける一般論では，階層的クラスタリングは質問項目のクラスタリングに向いているとされ，対して非階層的クラスタリングは回答者のクラスタリングに適しているとされる [23].

2.2.1　ウォード法

ウォード法は，ユークリッド空間の距離による非類似度に基づいた手法である．すなわち，各データを x とすると非類似度 d は

$$d(x_i, x_j) = \|x_i - x_j\|^2 \tag{2.12}$$

である．p を空間の次元とすると，クラスタ G に関する重心 $M(G) = (M^1(G), \cdots, M^p(G))$ は

$$M^j(G) = \frac{1}{\#G} \sum_{x_i \in G} x_i^j \tag{2.13}$$

で与えられる．さらに，

$$E(G) = \sum_{x_i \in G} \|x_i - M(G)\|^2 \tag{2.14}$$

とすると，$E(G) = \mathrm{Var}[G]\#G$ であり，クラスタ G を重心 $M(G)$ で代表させることにより発生する誤差を表すことができる．そこで，

$$\Delta E(G_i, G_j) = E(G_i \cup G_j) - \{E(G_i) + E(G_j)\} \tag{2.15}$$

とおき，ΔE が最小となるクラスタ G_q, G_r を結合することで，誤差の最小化が実現できる．ウォード法は外れ値に強く，実用性が高い．しかし $O(N^2)$ であり，後述の k-means 法よりも計算量が多いことが欠点である．

2.2.2　k-means 法

k-means 法も，非類似度としてユークリッド距離を用いた手法である．データ群を k 個のクラスタに分類する場合，アルゴリズムは，

1. データ群から k 個のデータを無作為に選び，各クラスタの中心とする．
2. 各データを最も近い平均値をもつクラスタに割り当てる．
3. 割り当てられたクラスタについて，平均値を更新する．

の 3 ステップで進む．ステップ 2 とステップ 3 を繰り返し，クラスタ割り当て結果が前ループと同じであれば，処理を終了する．k-means 法の長所は実装が容易で，計算量も $O(Nk)$ と高速な点である．ただし，初期値の選び方によっては適切な結果が得られないことがある．

2.3　ベイジアンネットワーク

ベイジアンネットワークは，複数の確率変数間の関係を有向グラフと条件付き確率で表した確率モデルであり，事象の予測などに用いられる．図 2.2 に例として示すのは，確率変数 $X_1, \cdots, X_5$ と条件付き確率 $P(X_3|X_1, X_2), P(X_4|X_3), P(X_5|X_3)$ および有向グラフによって定義されたベイジアンネットワークである．

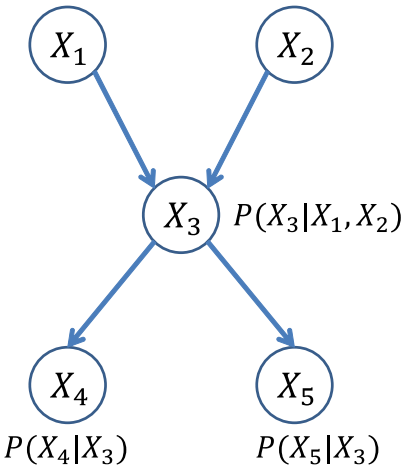


図 2.2: ベイジアンネットワークの一例

本節ではベイジアンネットワークを構成するこれらの 3 要素（確率変数，有向グラフ，条件付き確率）について説明する．

2.3.1　確率変数

確率変数とは，標本空間 Ω 中の標本点 ω を得る行為のことを指す．実データの確率変数には身長や時間など，連続値をとるものが多く存在する．ベイジアンネットワークを構築する際には，連続値データをそのままノードに与えると，状態数が膨大になり計算効率が悪化する可能性がある．そのため，ベイジアン

ネットワークを構築する前処理として，確率変数の離散化を行うことが望ましいとされている．本研究では，データマイニング手法の一つであるクラスタリングを離散化に用いる．

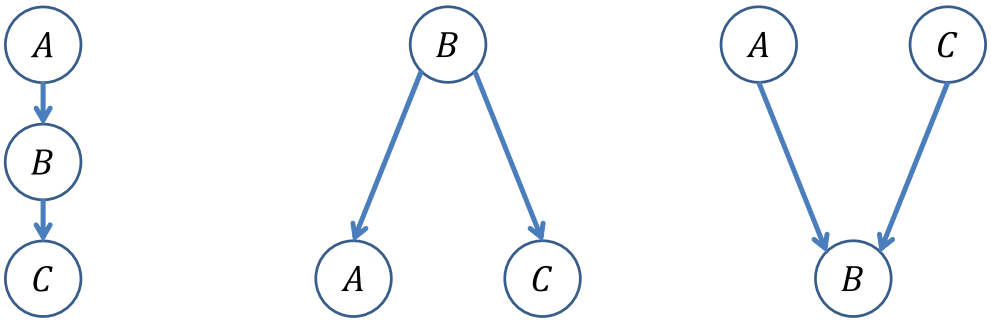
2.3.2　有向グラフ

有向グラフとは，すべてのエッジが有向エッジからなるグラフである．グラフ $G = (\mathbf{V}, \mathbf{E})$ はグラフのノード集合 $\mathbf{V} = \{V_1, \cdots, V_N\}$ と二つのノード V_i と V_j の間にエッジが存在するときのエッジ集合 $\mathbf{E} = \{E_{ij}\}$ により定義され，このエッジ集合 $\mathbf{E}$ に対して $E_{ij} \in \mathbf{E}$ かつ $E_{ji} \notin \mathbf{E}$ が成り立つとき，$E_{ij} = V_i \to V_j$ は有向エッジと呼ばれる．また，このときの V_i を V_j の親ノード，V_j を V_i の子ノードと呼ぶ．

有向グラフにおいて，始点と終点が同じノードとなるような路は循環と定義される．この循環が一つもないグラフは有向非循環グラフ（Directed Acyclic Graph，DAG）といい，ベイジアンネットワークに応用される．しかし，各確率変数の状態数を w としたときの評価にかかる計算量は $O(w^N)$ であり，応用上好ましくない．すべての場合に対して計算をすることなくベイジアンネットワークを構築するには，d 分離という概念が重要となる [24]．d 分離とは，次の条件が成り立つことをいう．

1. 線形経路，または分岐経路において B がインスタンス化されたとき．
2. 合流経路において B がインスタンス化されていないとき．

ここで，インスタンス化とは変数の状態がわかることを指す．以下，簡単のため図 2.3(a), 図 2.3(b), 図 2.3(c) のような，三つのノードからなる DAG を前提に d 分離について説明する．



(a) 線形経路　(b) 分岐経路　(c) 合流経路

図 2.3: 3 ノード間の関係

線形経路の場合

A は B を経由して，C に影響を与える．同様に，C のインスタンス化も B を通じて A に影響を与える．しかし，B がインスタンス化された場合，A と C は独立となる．このことは，乗法公式により次のように示せる．

$$P(A, C|B) = P(A|B)P(C|B) \tag{2.16}$$

分岐経路の場合

B がインスタンス化されない限り B のすべての子ノード間で情報が伝播される．一方，B がインスタンス化されると A と C は独立となり，A の情報は C を知るための新しい手がかりにはならなくなる．このことも，線形経路の場合と同じように乗法公式によって示せる．

合流経路の場合

B がインスタンス化されると，A の事象 a が起きたという情報は C の事象 c が起こる確率を減少させ，逆に a が起こらなかったという情報は c が起こる確率を増加させる．これは弁明と呼ばれる現象である．B がインスタンス化されていないときには，A と C は独立であり，A に関して何らかの情報が与えられても，その情報は C に影響を与えない．このことを，以下に示す．同時確率 $P(A,B,C)$ は

$$P(A,B,C) = P(B|A,C)P(A)P(C) \tag{2.17}$$

と表せる．ここで B が未知，すなわちインスタンス化されていないならば周辺化を行って，

$$P(A,C) = P(A)P(C) \tag{2.18}$$

が得られるので，A と C は独立である．

このように，d 分離という考え方を用いてノード間の独立性を調べることで，構造設計の高速化が可能となる．

グラフ構造の探索

すべての DAG を探索した場合には厳密解が得られる．しかし，構造数の計算式

$$f(N) = \sum_{i=1}^{N} (-1)^{i+1} \binom{N}{i} 2^{i(N-i)} f(N-i) \tag{2.19}$$

にも明らかなように，ノード数 N を少し増やすだけで DAG の数は膨大となる．そこで，計算時間の短縮のために発見的な手法がとられることが多い．ここでは，DAG の学習手法としてよく知られた K2 アルゴリズムについて述べる．K2 アルゴリズムは，$X_1 > \cdots > X_N$ のような全順序関係を前提としている．この全順序関係を時間的順序と考えれば，この手法は「結果が起こるのは原因より後」という性質をよく表しているといえる．アルゴリズムは，

1. 子ノードを定義する．
2. 子ノードごとに，全順序関係を用いて候補となる親ノード集合を探索する．
3. 子ノードと親ノードの条件付き確率を決定する．
4. 情報量基準に基づき，最適な親ノードを子ノードごとに探索する．

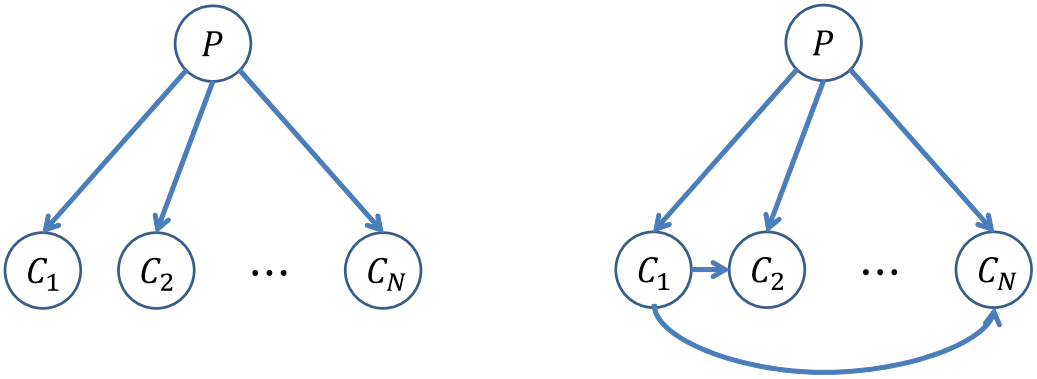
の 4 ステップで進む．このようにして得られるグラフ構造のうち，代表的なものを 2 つ紹介する．

NBC

図 2.4(a) のように，ある一つのノード P から他のすべてのノード $C_1, C_2, \cdots, C_N$ に向かってエッジが伸びているグラフ構造をもつベイジアンネットワークを単純ベイズ分類器（Naive Bayes Classifier, NBC）[25] と呼ぶ．多くの場合，目的変数を親ノード，説明変数を子ノードとする．NBC は比較的精度の高い分類結果が得られ，なおかつ説明変数間に完全な独立性を仮定した単純な構造のため高速計算が可能である．

TAN

NBC は説明変数間の完全な独立性を仮定していたが，その条件を緩和して子ノード間に木構造を許し，説明変数間の相互作用を表現できるようにした．このグラフ構造を木構造拡張化単純ベイズ分類器（Tree Augmented Naive-Bayes-Classifier, TAN）[26] と呼ぶ．TAN の一例を図 2.4(b) に示す．



(a) NBC　(b) TAN

図 2.4: ベイジアンネットワーク構造の代表的な例

情報量基準

予測モデルを選択する基準となるのが，情報量基準である．情報量基準は大きく分けて二つあり，一つは期待対数尤度を用いるもの，もう一つは予測分布を用いるものである．ここでは，前者の例として AIC を，後者の例として BIC を説明する．

AIC では，推定されたモデルの分布と真の分布との間の KL ダイバージェンスを最小化するモデルを採用する．ここで，最大対数尤度を L, パラメータ数を k とすると，AIC は

$$AIC = -2L + 2k \tag{2.20}$$

と定義される．この式の優れたところは第一項で推定モデルの精度を評価するだけでなく，パラメータ数を罰則項として加えることで，過学習への対策をしている点にある．一方，BIC の定義は，

$$BIC = -2L + k \log N \tag{2.21}$$

である．AIC と異なる点は，サンプルサイズ N が罰則項に含まれている点にある．

2.3.3　条件付き確率

エッジで結合された 2 つのノードの確率変数がとりうる値のすべての組み合わせのデータが存在する完全データの場合には，グラフ構造の探索により条件付き確率値を得ることができ，これらの値を格納する条件付き確率表（Conditional Probabilities Tables, CPT）が得られる．しかし，完全データでない場合，すなわち不完全データの場合には CPT のすべての項を埋めることができない．そこで，不完全データの場合には，

1. 仮説モデルで確率推論を行い，CPT における欠損部の推定値を確率分布として得て，これを疑似的な完全データとする．

2. 1. で得た疑似的な完全データから CPT を定め，これを新しい仮説モデルとする．

を繰り返して，CPT の計算を行う．ちなみに，ステップ 1. は E ステップ，ステップ 2. は M ステップと呼ばれ，手続き全体を EM（Expectaion Maximization）アルゴリズムと呼ぶ．

確率伝播法

このようにして得られた CPT をもとに，ベイジアンネットワークは確率推論を行う．その手順は以下の通りである．

1. 観測された変数の値 e をノードにセットする．
2. 親ノードも観測値も持たないノードに事前確率分布を与える．
3. 知りたい対象の変数 X の事後確率 $P(X|e)$ を得る．

この事後確率を求めるために用いられる手法が，確率伝播法 [27] である．この手法では観測された情報からの確率伝播，すなわち変数間の局所計算により，各変数の確率分布を更新していくという計算法がとられる．以下，簡単な例を挙げて確率伝播法を概説する．

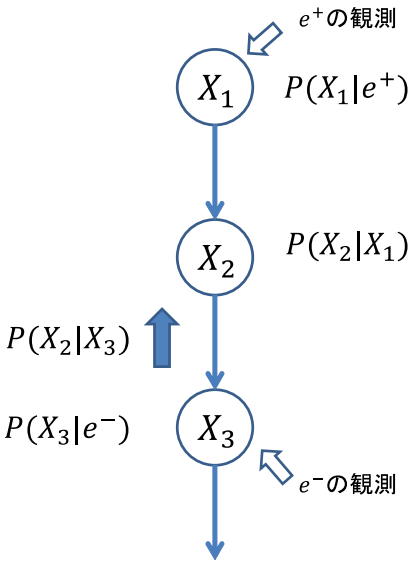


図 2.5: 確率伝播法

図 2.5 において，$P(X_2|e)$ を知りたい場合を考える．親ノードへの観測情報 e^+, 子ノードへの観測情報 e^- およびベイズの定理を用いて，$P(X_2|e)$ は次のように表される．

$$P(X_2|e) = \frac{P(e^-|X_2, e^+)P(X_2|e^+)}{P(e^-|e^+)} \tag{2.22}$$

ここで，$P(e^-|e^+)$ は X_2 に依存しない項であるから，正規化定数 α として考えることが出来るので，

$$P(X_2|e) = \alpha P(e^-|X_2)P(X_2|e^+) \tag{2.23}$$

と変形できる．親ノードから伝播する確率を $P(X_2|e^+) = \pi(X_2)$ とおくと，

$$\pi(X_2) = \sum_{X_1} P(X_2|X_1)P(X_1|e^+) \tag{2.24}$$

が成り立つ．$P(X_1|e^+)$ は次のように得る．

(i) 観測情報があるときは，その値を用いる．

(ii) 観測情報がないときは，親ノードの有無により場合分けする．

(1) 親ノードがないときは，事前確率を与える．

(2) 親ノードがあるときは，式 2.24 によって再帰的に計算する．

一方で，子ノードから伝播する確率を $P(e^-|X_2) = \lambda(X_2)$ とおくと，

$$\lambda(X_2) = \sum_{X_3} P(e^-|X_3)P(X_3|X_2) \tag{2.25}$$

となる．$P(e^-|X_3)$ は以下のように決定する．

(i) 観測情報があるときは，その値を用いる．

(ii) 観測情報がないときは，親ノードの有無により場合分けする．

(1) 子ノードがないときは，一様確率分布と仮定する．

(2) 子ノードがあるときは，式 2.25 によって再帰的に計算する．

ゆえに，式 2.24 と式 2.25 を式 2.23 に代入することで，ノード X_2 の事後確率を計算することができる．同様に，次式によって任意のノードの事後確率も求めることができる．

$$P(X_j|e) = \alpha\lambda(X_j)\pi(X_j) \tag{2.26}$$

しかし，この確率計算はグラフの構造によっては収束しないことがある．グラフ構造を 2 つの場合に分けて，説明する．計算が収束するのは，ベイジアンネットワークのエッジの向きを考慮しないグラフ構造において，すべての路がループを持たないときである．このベイジアンネットワークは単結合なネットワークと呼ばれ，この場合の確率計算は各ノードに結合するエッジ数 n に対して $O(n)$ で完了する．

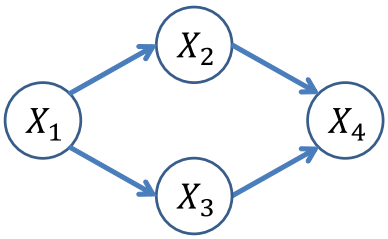
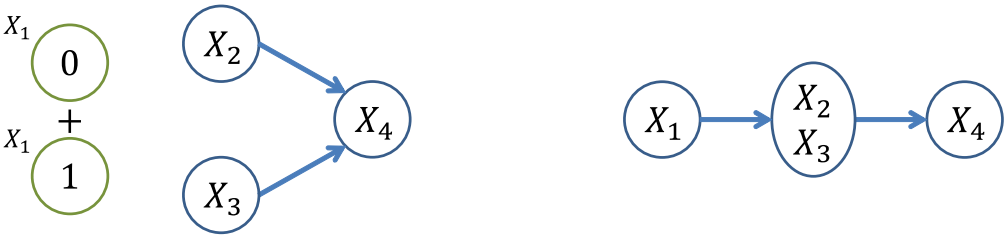


図 2.6: 複結合なベイジアンネットワークの例

逆に，図 2.6 に示すような一つでもループを持つグラフの場合，そのベイジアンネットワークは複結合なネットワークと呼ばれ，確率計算の収束性は保障されない．この問題の対処法には，コンディショニングとクラスタリングの 2 つがある．



(a) コンディショニング　(b) クラスタリング

図 2.7: 複結合なベイジアンネットワークの変換法

前者は図 2.7(a) のように，確率変数の値で場合分けしてループをなくす方法である．それに対して，後者は図 2.7(b) のようにノードをクラスタリングしてグラフを変換する手法で，代表的な手法としてジャンクションツリーアルゴリズムがある．これは，適切な親ノードの結合を繰り返し，ノードのクリークをクラスタとして生成することで，単結合な木構造からなる無向グラフに変換するアルゴリズムである．

ただし，複雑な構造を持つベイジアンネットワークの場合には，変換の計算コストが大きくなるという欠点がある．そのため，近年では Loopy Belief Propagation やサンプリング法に代表される，グラフ変換なしでそのまま計算を行うアルゴリズムの研究が進んでいる．

2.4　評価

ベイジアンネットワークの出力は，事後確率で表現される．有用な知識を得るためには，この値を何らかの方法で評価しなければならない．そこで本節では，その出力評価法として交差検証とホールドアウト検証，分類問題について概説する．

2.4.1　交差検証

交差検証は図 2.8 に示すように，モデルの精度計算に利用される手法であり，

1. 元データ（データ数 N）を均等に，k 個のブロックに分割する．
2. 1 つのブロックを選んで検証データ，残りを学習データとしてモデル構築と精度計算を行う．
3. ステップ 2 の操作を k 個のブロックが 1 回ずつ検証データとなるように，k 回繰り返す．
4. 各回で計算された精度の平均を，モデルの推定精度とする．

という手順で行われる．交差検証は元データに対して $k-1$ 倍の学習データでモデル構築を行うことができるため，元データが少ないときによく利用される手法である．また，特に $k = N$ の場合を Leave-one-out 法と呼ぶ．

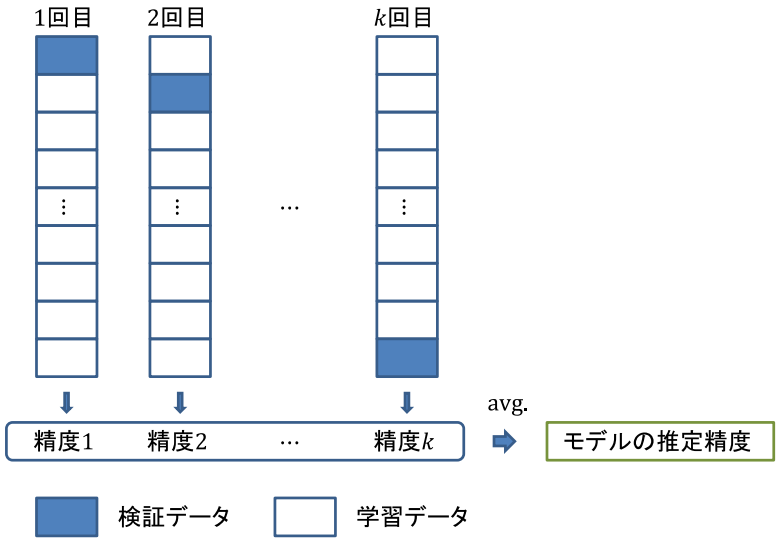


図 2.8: 交差検証

2.4.2　ホールドアウト検証

交差検証は k 回の繰り返しの中で，すべてのブロックが 1 回ずつ検証データとなるよう処理を行うのに対し，ホールドアウト検証は学習データと検証データを完全に分けて処理を行う．これにより，学習データから構築したモデルが未知な事象をどの程度予測できるのかを評価する．ホールドアウト検証によって推定される精度は図 2.8 の精度 1, 精度 2, $\cdots$, 精度 k にあたり，交差検証とは異なり平均値ではない．ゆえに，その精度はデータの分け方に強い影響を受け，構築されたモデルのロバスト性を検証するのに適しているといえる．

2.4.3　分類問題

ここでは，クラス（状態）数 $M = 2$ の 2 値分類問題を考える．2 値分類問題は，入力があるクラスに属している（正例クラス）か，あるクラスに属していない（負例クラス）かを判定する問題である．本研究では，入力としてベイジアンネットワークの出力である事後確率値を与え，評価を行う．予測モデルを用いて，データをこの 2 つのクラスに分類するときの精度を評価する方法には，分割表によるものがある．分割表を用いると 2 値分類問題は，

表 2.1: 2×2 分割表

実際 \ 予測	正例	負例
正例	TP	FN
負例	FP	TN

のように表せる．表 2.1 の TP, FP, FN, TN の意味は，

$TP \cdots$ True Positive, 実際に正例クラスに属するデータを正例と予測した件数.

$FP \cdots$ False Positive, 実際は負例クラスに属するデータを正例と予測した件数.

$FN \cdots$ False Negative, 実際は正例クラスに属するデータを負例と予測した件数.

$TN \cdots$ True Negative, 実際に負例クラスに属するデータを負例と予測した件数.

であり，この 4 つの値を用いて情報検索システムの評価指標 accuracy（分類正解率），recall（再現率），precision（適合率），および F-measure（F 値）がそれぞれ定義できる．

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{2.27}$$

$$recall = \frac{TP}{TP + FN} \tag{2.28}$$

$$precision = \frac{TP}{TP + FP} \tag{2.29}$$

$$F - measure = \frac{(\beta^2 + 1) \times recall \times precision}{recall + \beta^2 \times precision} \tag{2.30}$$

分類正解率は分類の予測が的中した割合を示し，単に正解率とも呼ばれる．再現率は実際の正例クラスに属するデータの件数に対してモデルが正例クラスと予測できた件数の割合を示し，適合率はモデルが正例クラスと予測したデータの件数の中で実際に正例クラスに属するデータの件数が占める割合を示す．この 2 つの指標はトレードオフの関係にあるため，両者を同時に評価するための F-measure（F 値）なる指標も本研究では用いている．β は，適合率の再現率に対する相対的な重要度を示すパラメータで，通常 $\beta = 1$ とする．F 値は 0 から 1 の間の値をとる recall と precision の調和平均値であり，その値が大きいほど予測モデルの性能が良いことを意味する．

第3章　データ概要とその変換

本章では本研究で用いるデータの概要と，その変換方法について述べる．データは成績データ，打刻データ，修学データの 3 種類から成る．これらのデータを要注意学生の推定に利用するため，確率変数への変換処理を行う．

3.1　データの概要

本研究では，過去のある年度 A, B に名工大へ入学した 338 名の学生（A 年度入学 171 名，B 年度入学 167 名）に関するデータを用いる．先述したようにデータは成績データ，打刻データ，修学データの 3 種であり，成績データと打刻データは各入学年度に対して 2 年分のレコードを持つ．この節では，これら 3 種類のデータについて簡単に述べる．なお，データのイメージとして挙げている表のデータは説明上作成したものであり，元のデータとは一切関係がないことを付記しておく．

3.1.1　成績データ

成績は教員が開講期ごとに決めるもので，名工大では表 3.1 のように六段階評価である．評点が 60 以上であれば合格となり，その授業の単位修得が認められるが，不可と失格の場合は認められない．不可と失格の違いは，評価可否にある．すなわち，試験を欠席した場合や，出席回数が不足している場合は評価不能として失格扱いとなる．GP は各成績評価に対して与えられるポイントであり，GPA の算出に用いられる．また代表値は，秀–可に関しては中央値を，不可と失格に関しては 0 を与える．

表 3.1: 成績

評語	評点	代表値	GP	判定
秀	90–100	95.0	4	合格
優	80– 89	84.5	3	
良	70– 79	74.5	2	
可	60– 69	64.5	1	
不可	0– 59	0.0	0	不合格
失格	評価不能	0.0	0	

表 3.1 で示した成績と暗号化された学生番号，および授業の属性を加えた 5 つの情報が本研究で用いるために与えられた成績データであり，そのレコード数は約 1 万 8 千である．

授業科目名は,「演習 1」や「専門 1」のように変更されている．これは学生の受講した授業の特定を防ぎ，所属する学科が分からないようにするための措置である．同じ授業科目名を持つレコードでも授業番号が異なるものがある．これは，当該授業が複数クラス開講されていることを示している．また，開講期は前期と後期の二つの値をとり，前期は 4 月–9 月を，後期は 10 月–3 月を指す．

3.1.2　打刻データ

打刻データは，前述した IC カード出欠システムにより得られる時刻情報のことである．本研究のために暗号化学生番号，日付（年月日），打刻時間の 3 つの情報が与えられている．レコード数は約 42 万である．表 3.2 に，打刻データのイメージを示す．

表 3.2: 打刻データ

暗号化学生番号	日付	打刻時間
12509	20160411	084548
12509	20160411	102142
⋮	⋮	⋮
93011	20170202	142531

打刻時間は秒単位まで記録されており，例えば表 3.2 の 084548 は，8 時 45 分 48 秒に打刻されたことを示している．なお，この打刻データに関連するデータとして「出欠データ」も存在する．出欠データは，打刻データと授業の開講時間データから自動生成されている．本研究では，表 3.3 に示すような形式で与えられている．

表 3.3: 出欠データ

暗号化学生番号	暗号化授業番号	日付	入室打刻	退室打刻	出欠
12509	5188745	20160411	084548	102142	出席
⋮	⋮	⋮	⋮	⋮	⋮
93011	3192231	20170202		142531	遅刻

与えられた出欠データは六つの属性を持つ．入室打刻 t_e は学生が教室へ入室したときに打刻した時刻，退室打刻 t_g は学生が教室から退室したときに打刻した時刻をそれぞれ表す．出欠判定は t_e と t_g が適切な時間内になされているかどうかにより行い，出席，遅刻，早退，欠席の 4 つに分類される．ここで，暗号化学生番号と暗号化授業番号は成績データと共通のものであるため，出欠状況と成績の関連付けができることを明記しておく．

3.1.3　修学データ

名工大では，1 年次から 3 年次までに各授業を履修し，4 年次に研究室配属が行われる．しかし，3 年次までに一定の単位を取得しなければ，4 年次生でも卒業研究に着手することは認められない．また，通常の修業年限は 4 年であるが，所定の単位を修得しなければ卒業研究に着手できてもその年度に卒業することはできない．表 3.4, 表 3.5, 表 3.6 は，学生が卒業研究に着手するまでの年数分布と卒業するまでの年数分布，および退学理由の分布をそれぞれ示している．

表 3.4: 卒業研究に着手するまでの年数分布

	3 年	4 年	5 年	6 年	未着手	退学	計
A 年度入学生	145	10	2	3	5	6	171
B 年度入学生	138	13	2	0	6	8	167
計	283	23	4	3	11	14	338

表 3.5: 卒業までの年数分布

	4 年	5 年	6 年	在学中	退学	計
A 年度入学生	134	19	3	8	7	171
B 年度入学生	134	12	0	10	11	167
計	268	31	3	18	18	338

表 3.6: 退学理由

	一身上	就職	転学科	他大学受験	授業料未納	経済的理由	不明	計
A 年度入学生	2	2	1	0	2	0	0	7
B 年度入学生	4	1	0	2	2	1	1	11
計	6	3	1	2	4	1	1	18

表 3.4 において，未着手は卒業研究に着手していないことを意味し，退学は卒業研究に着手する前に退学したことを意味している．また表 3.5 において，在学中は卒業・退学をしていない状態を意味し，退学は退学したことを意味している．

3.2　データの変換

成績データ，打刻・出欠データおよび修学データは，この形式のままでは分析に不適なため，変換をして変数を生成する必要がある．そこで，本節と次節で，これらのデータの変換について述べる．本節では，従来用いてきた変数の生成について，次節では，要注意学生の推定精度を改善することを目的とした新変数の生成について説明する．

3.2.1　成績データの変換

本研究は教育支援を目的としたものであるため，各学生の傾向をつかむことが肝要である．このことを鑑み，成績データを学生ごとの成績，獲得成績数に変換する．成績指標としては，客観的な評価法として近年普及が進んでいる GPA を採用した．GPA は多くの教育機関で採用されており，名工大もその一例である．履修登録した講義 l の集合を L とすると，GPA は表 3.1 で示した GP 値 $GP(l)$ と，講義ごとに決められている単位数 c_l を用いて，

$$GPA = \frac{\sum_l^L GP(l) \times c_l}{\sum_l^L c_l} \tag{3.1}$$

と表される．例えば，学生が履修したすべての講義で秀を獲得すれば GPA は 4.0 で，逆に履修したすべての講義で不合格となれば GPA は 0.0 となる．このように，学生ごとの GPA 値を比較することによって，学生の修学傾向を評価することが可能である．本研究では，一般に GPA の意味で使われる開講期別の GPA だけでなく，講義区分別の GPA も用いている．講義区分は大学が定めているものであり，本研究ではそれに倣って表 3.7 に示す 6 つの区分で GPA を算出した．

また，各成績評価の獲得数も年次・開講期ごとに算出した．例えばある学生の GPA が 2.0 でも，すべての講義で良を獲得した場合と，秀と不合格を同数ずつ獲得した場合が考えられるからである．成績評価は表 3.1 のように 6 段階であるため，各開講期において 6 つの変数が生成される．

表 3.7: 講義区分

区分名	意味
理系基礎	数学や理科など，自然科学に分類される科目に対する区分
外国語	外国語教育に分類される科目に対する区分
人間社会	人文科学，社会科学に分類される科目に対する区分
体育	体育教育に分類される科目に対する区分
専門	専門教育に分類される科目に対する区分
その他	上記区分に該当しない科目に対する区分

3.2.2　打刻・出欠データの変換

打刻・出欠データの変換については，新変数の扱いとなるため，次節を参照されたい．

3.2.3　修学データの変換

教育支援への実用化を考えたとき「一見した限りでは優秀であるが，将来的に修学状況が悪化し留年・退学してしまう学生」を推定により見つけ出すことが重要である．例えば，GPA が 0.0 に近い学生に対して修学指導が必要であることは誰にでも判断できるが，GPA が 2.0 付近の学生に対して教育指導が必要かどうかは判断が難しい．そこで，このように要指導か否かの判断が難しい学生を推定するため，修学データと成績データを用いてこれらの学生に明確な定義を与える．まず，各学生の 1 年次前期・1 年次後期における GPA と，留年・退学の関係は表 3.8 の通りであった．

表 3.8: 1 年次 GPA と留年・退学の関係

GPA	1 年前期			1 年後期		
	人数	留年・退学者	割合	人数	留年・退学者	割合
[0.0, 0.5)	5	5	100%	13	13	100%
[0.5, 1.0)	8	6	75%	12	12	100%
[1.0, 1.5)	12	8	67%	32	15	47%
[1.5, 2.0)	47	22	47%	68	14	21%
[2.0, 2.5)	106	16	15%	95	10	11%
[2.5, 3.0)	99	10	10%	69	2	3%
[3.0, 3.5)	53	3	6%	42	4	10%
[3.5, 4.0)	8	0	0%	7	0	0%
計	338	70		338	70	

全学生 338 名のうち留年・退学した学生は 70 名であり，これは表 3.5 を使うことでも，$338 - 268 = 70$ により計算できる．GPA の区間別に留年・退学者の割合を見ると，GPA が高くなるに従ってその割合は少なくなっていることが分かる．また 1 年次前期，または 1 年次後期において GPA が 1.0 未満の学生のほとんどが留年・退学をしており，1 年次後期に限れば GPA1.0 未満の学生は全員，留年・退学をしている．このことから 1 年次前期，または 1 年次後期における GPA がどちらか一方でも 1.0 未満である学生は，留年・退学する可能性が高いと推測することができる．そこで，1 年次前期，または 1 年次後期における GPA がどちらか一方でも 1.0 未満である学生は，修学指導が必要であることが明らかであると考え，本研究の推定対象者から除外する．これに該当する学生は 1 年次前期において 13 名，1 年次後期において 25 名である

が，その重複 7 名を除くと 31 名である．またそのうち，留年・退学者は 29 名である．以上をベン図で表現すると，図 3.1 のようになる．

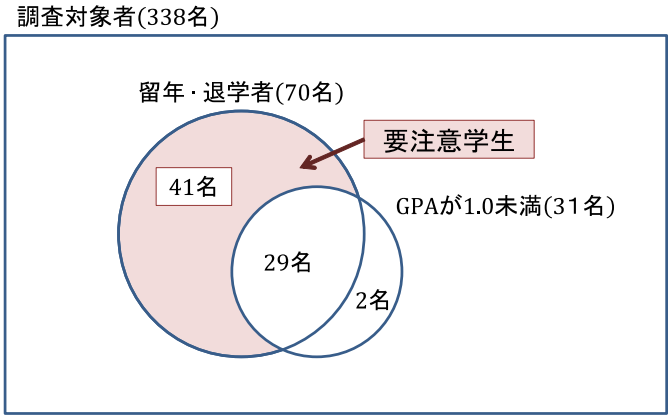


図 3.1: 従来研究 [14] における要注意学生の定義

この図は従来研究 [14] における要注意学生の定義を表している．つまり推定すべき学生は，「1 年次前期・1 年次後期における GPA がともに 1.0 以上であり，将来的に留年・退学する学生」である．具体的には，$338 - 31 = 307$ 名の学生から 41 名の要注意学生を推定することが，この研究における目的であった．しかし，GPA による機械的な定義だけでは，大学院浪人や就職留年といった計画的に留年を選択する学生，転学のために退学した学生など教職員による修学指導を必要としていないと考えられる学生も要注意学生に含めてしまう．本来発見したい学生は学校に馴染めない学生や学業に不安がある学生であり，彼らにこそ修学指導が必要である．

そこで，従来研究 [15][16] では，このような学生を発見するために定義の改良を行っている．表 3.4, 表 3.5 を比較すると，$283 - 268 = 15$ 名の学生が 3 年で卒業研究に着手できているにも関わらず，4 年間で卒業できていないことが分かる．この 15 名のうち，1 年次前期・1 年次後期において，ともに GPA が 1.0 以上であった学生だけを抜き出して，表 3.5 を書き直すと表 3.9 が得られる．

表 3.9: 3 年で卒業研究に着手できた学生の卒業までの年数分布

	5 年	6 年	在学中	退学	計
A 年度入学生	7	0	0	1	8
B 年度入学生	0	0	1	3	4
計	7	0	1	4	12

表 3.9 を見ると，5 年で卒業した学生が 7 名，在学中の学生が 1 名，退学した学生が 4 名であることが分かる．5 年で卒業した学生や在学中の学生 11 名は，大学院浪人や就職留年をしたと考え，要注意学生に含めないこととする．残りの学生 4 名は順調に 4 年次まで進級できたにも関わらず，退学をしている．全学生 338 名から GPA フィルタで除外した 307 名の中には，この 4 名を含めて退学者が 9 名存在する．この 9 名の退学者について，退学理由と在学年数，卒業研究着手までの年数をまとめたものが表 3.10 である．また参考資料として，退学者の退学理由について平成 24 年に文部科学省が行った調査 [28] の結果を図 3.2 に示す．

図 3.2 を見ると，退学理由として，その他を除き最も高い割合を占めたのが経済的理由である．平成 19 年度における調査と比較すると，その割合は他の退学理由よりも増加しており，年々増加傾向にあると推測される．要注意学生の推定をするとき，その対象となるべき学生は学業不振を理由に退学をする学生である．そのため，退学理由として最も多いのは経済的理由であるという現状の中，すべての退学者を一律に要注意学生の予測対象とすることは問題があると考えられる．

表 3.10: 退学者に関するデータ

番号	退学理由	在学年数	卒業研究着手まで
1	転学科	0 年	未着手
2	他大学受験	1 年	未着手
3	他大学受験	2 年	未着手
4	授業料未納	2 年	未着手
5	一身上	3 年	3 年
6	就職	4 年	3 年
7	経済的理由	4 年	3 年
8	授業料未納	5 年	未着手
9	不明	不明	3 年

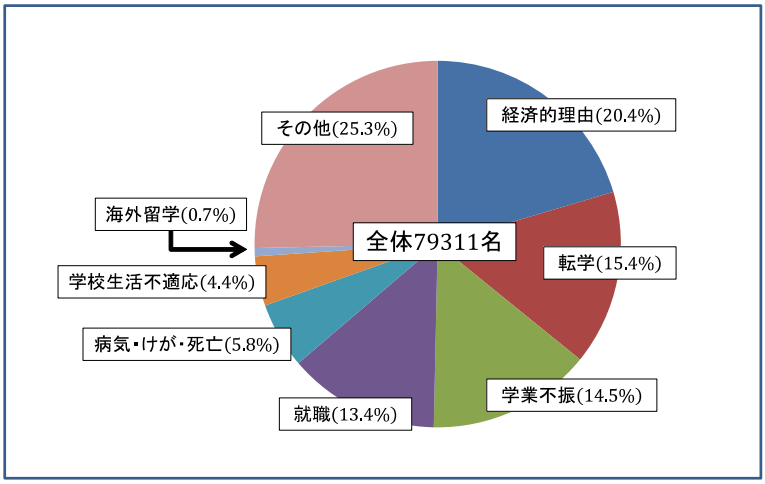


図 3.2: 平成 24 年度における退学者の退学理由

ところで，表 3.10 を見ると，番号 1–4 の学生 4 名は在学年数が 3 年に満たず，そのうち 3 名は積極的理由（転学科，他大学受験）である．また，番号 7 の学生は経済的理由による退学である．このように退学者の調査をすると，これら 5 名の学生は，真に修学指導が必要だと考えられる学生とはいえない．そこで，これら 5 名の学生はデータから除外することとする．以上をまとめると，修正した要注意学生の定義は次のようになる．

要注意学生

1 年次前期・1 年次後期における GPA がともに 1.0 以上であるにも関わらず，将来的に留年・退学する学生．ただし，3 年で卒業研究に着手した学生は対象外，在学年数が 3 年に満たない退学者と，経済的理由による退学者のデータは除外する．

従来研究では，この再定義によって，より修学指導やアドバイスを必要としている学生を 41 名から 25 名に，調査対象を 307 名から 302 名に絞り込むことができるため，精度の向上が期待できるとしている．

3.3　推定精度改善のための新変数生成

前述した通り，本節では要注意学生の推定精度を改善することを目的とした新変数の生成について説明する．生成した変数は 3 種類で，1 つ目は科目属性（必修・選択）を考慮した変数，2 つ目は授業クラスにおける成績の偏差値を考慮した変数，3 つ目は成績と出席の関連付けによる変数である．

科目属性を考慮した変数

従来手法では，講義が必修科目か選択科目かの属性を無視して変数生成を行っていた．しかし,「必修科目で不合格判定を受けた講義の数が多いと留年しやすい」という予想を立てた場合，必修科目かどうかを考慮することは要注意学生の推定において良い結果を期待できると考えられる．そこで，新しい変数として次の変数を生成する．

表 3.11: 科目属性を考慮した変数

変数名	意味
必修科目不合格数	履修した必修科目において不可または失格の評価をうけた数

成績の偏差値を考慮した変数

GPA と成績評価の獲得数は絶対的な値であり，講義難易度を反映した数値とはいえない．そのため，これらの評価値のみを学生の成績値として扱うことは，要注意学生の推定において不都合であると考えられる．そこで，難しい講義で成績評価 G をうけた学生と，易しい講義で同じ成績評価 G をうけた学生を分けて考えることで，推定精度の向上を図りたい．その手法として，授業クラスにおける偏差値を変数として生成する．これにより，クラスの平均点が低い講義，例えば評価の厳しい先生の担当する講義や内容そのものが難しい講義で良い成績を修めた学生をより高く評価し，逆にクラスの平均点が高い講義で悪い成績を取った学生を低く評価することができる．具体的には，表 3.1 の代表値から各講義クラスごとの偏差値 s_l を計算し，その値域によって 3 つの集合 $S_{low} = \{s_l | s_l < 40\}, S_{mid} = \{s_l | 40 \leq s_l < 60\}, S_{high} = \{s_l | 60 \leq s_l\}$ を生成する．これらの集合を用いて，表 3.12 の変数を定義する．

表 3.12: クラス内偏差値から生成される変数

変数名	意味
$\#(S_{low})$	クラス内偏差値が 40 未満の履修講義数
$\#(S_{mid})$	クラス内偏差値が 40 以上 60 未満の履修講義数
$\#(S_{high})$	クラス内偏差値が 60 以上の履修講義数

成績と出席の関連付けによる変数

従来研究において，打刻・出席データから生成している変数は月別の打刻回数・出席回数であり，授業成績との関連付けは行っていない．そこで，成績と出席を別個に考えるのではなく，統合した変数を用いることで推定精度の向上を図りたい．成績と出席を関連付けるために出欠データからクラス別の出席数（出席，遅刻，早退の総計）を計算する．ただし，同じ名前をもつ講義でも，クラスにより担当教員や教室が異なる．出席をとらない教員や，IC カード出欠システムのカードリーダがない教室で授業がある場合を考慮すると，単にクラス別の出席数を変数とするのは整合性に疑問が残る．以上の理由から，成績の偏差値と同じように，他の履修生と比較してどの程度出席していたかを知るため，クラス内における出席数の偏差値 $s_{l,att}$ を計算することで正規化を行う．そして，$S_{l,att} < 40$ となる履修講義数を変数とする．ただし，出席数は高々15 回であり標準偏差 σ_l が小さくなることを考慮して，$\sigma_l < 1$ となるデータは除外した．また，教員は出席数が一定回数に満たない学生に対して $e_l = X$, すなわち失格評価を与えるが，何らかの理由で他の評価が与えられたデータが存在する．以上から，打刻・出席データより，

$$S_{att} = \{l | (s_{l,att} < 40) \wedge (1 \leq \sigma_l) \wedge (e_l \neq X)\} \tag{3.2}$$

なる集合を定義し，$\#(S_{att})$, すなわち「出席数の偏差値が 40 未満，標準偏差が 1 以上，かつ失格評価を受けていない履修講義数」を変数として生成した．

3.4　データ変換の総括

本研究では，ベイジアンネットワークを用いて要注意学生の推定を行うが，この要注意学生の推定には，データの変換により得られた変数を用いる．ここでは本節のまとめとして推定に用いる変数，いわゆる説明変数と，説明変数によって予測される要注意学生，すなわち目的変数をそれぞれ表 3.13 と表 3.14 に示す．説明変数は成績データと打刻・出席データから，目的変数は修学データから生成されている．

表 3.13: 説明変数

番号	変数名	意味
1	1 年次前期理系基礎 GPA	1 年次前期における理系基礎科目の GPA
2	1 年次前期外国語 GPA	1 年次前期における外国語科目の GPA
3	1 年次前期人間社会 GPA	1 年次前期における人間社会科目の GPA
4	1 年次前期体育 GPA	1 年次前期における体育科目の GPA
5	1 年次前期専門 GPA	1 年次前期における専門科目の GPA
6	1 年次前期その他 GPA	1 年次前期におけるその他科目の GPA
7	1 年次前期秀獲得数	1 年次前期における秀評価の数
8	1 年次前期優獲得数	1 年次前期における優評価の数
9	1 年次前期良獲得数	1 年次前期における良評価の数
10	1 年次前期可獲得数	1 年次前期における可評価の数
11	1 年次前期不可獲得数	1 年次前期における不可評価の数
12	1 年次前期失格獲得数	1 年次前期における失格評価の数
13	1 年次前期必修科目不合格	1 年次前期の必修科目で不可・失格評価を受けた数
14	1 年次前期 $\#(S_{low})$	1 年次前期においてクラス内偏差値が 40 未満の科目数
15	1 年次前期 $\#(S_{mid})$	1 年次前期においてクラス内偏差値が 40 以上 60 未満の科目数
16	1 年次前期 $\#(S_{high})$	1 年次前期においてクラス内偏差値が 60 以上の科目数
17	1 年次前期 $\#(S_{att})$	1 年次前期に $(s_{l,att} < 40) \wedge (1 \leq \sigma_l) \wedge (e_l \neq X)$ となる科目数
18	1 年次後期理系基礎 GPA	1 年次後期における理系基礎科目の GPA
19	1 年次後期外国語 GPA	1 年次後期における外国語科目の GPA
20	1 年次後期人間社会 GPA	1 年次後期における人間社会科目の GPA
21	1 年次後期体育 GPA	1 年次後期における体育科目の GPA
22	1 年次後期専門 GPA	1 年次後期における専門科目の GPA
23	1 年次後期その他 GPA	1 年次後期におけるその他科目の GPA
24	1 年次後期秀獲得数	1 年次後期における秀評価の数
25	1 年次後期優獲得数	1 年次後期における優評価の数
26	1 年次後期良獲得数	1 年次後期における良評価の数
27	1 年次後期可獲得数	1 年次後期における可評価の数
28	1 年次後期不可獲得数	1 年次後期における不可評価の数
29	1 年次後期失格獲得数	1 年次後期における失格評価の数
30	1 年次後期必修科目不合格	1 年次後期の必修科目で不可・失格評価を受けた数
31	1 年次後期 $\#(S_{low})$	1 年次後期においてクラス内偏差値が 40 未満の科目数
32	1 年次後期 $\#(S_{mid})$	1 年次後期においてクラス内偏差値が 40 以上 60 未満の科目数

次ページに続く

前ページからの続き

番号	変数名	意味
33	1 年次後期 $\#(S_{high})$	1 年次後期においてクラス内偏差値が 60 以上の科目数
34	1 年次後期 $\#(S_{att})$	1 年次後期に $(s_{l,att} < 40) \wedge (1 \leq \sigma_l) \wedge (e_l \neq X)$ となる科目数
35	2 年次前期理系基礎 GPA	2 年次前期における理系基礎科目の GPA
36	2 年次前期外国語 GPA	2 年次前期における外国語科目の GPA
37	2 年次前期人間社会 GPA	2 年次前期における人間社会科目の GPA
38	2 年次前期体育 GPA	2 年次前期における体育科目の GPA
39	2 年次前期専門 GPA	2 年次前期における専門科目の GPA
40	2 年次前期その他 GPA	2 年次前期におけるその他科目の GPA
41	2 年次前期秀獲得数	2 年次前期における秀評価の数
42	2 年次前期優獲得数	2 年次前期における優評価の数
43	2 年次前期良獲得数	2 年次前期における良評価の数
44	2 年次前期可獲得数	2 年次前期における可評価の数
45	2 年次前期不可獲得数	2 年次前期における不可評価の数
46	2 年次前期失格獲得数	2 年次前期における失格評価の数
47	2 年次前期必修科目不合格	2 年次前期の必修科目で不可・失格評価を受けた数
48	2 年次前期 $\#(S_{low})$	2 年次前期においてクラス内偏差値が 40 未満の科目数
49	2 年次前期 $\#(S_{mid})$	2 年次前期においてクラス内偏差値が 40 以上 60 未満の科目数
50	2 年次前期 $\#(S_{high})$	2 年次前期においてクラス内偏差値が 60 以上の科目数
51	2 年次前期 $\#(S_{att})$	2 年次前期に $(s_{l,att} < 40) \wedge (1 \leq \sigma_l) \wedge (e_l \neq X)$ となる科目数
52	2 年次後期理系基礎 GPA	2 年次後期における理系基礎科目の GPA
53	2 年次後期外国語 GPA	2 年次後期における外国語科目の GPA
54	2 年次後期人間社会 GPA	2 年次後期における人間社会科目の GPA
55	2 年次後期体育 GPA	2 年次後期における体育科目の GPA
56	2 年次後期専門 GPA	2 年次後期における専門科目の GPA
57	2 年次後期その他 GPA	2 年次後期におけるその他科目の GPA
58	2 年次後期秀獲得数	2 年次後期における秀評価の数
59	2 年次後期優獲得数	2 年次後期における優評価の数
60	2 年次後期良獲得数	2 年次後期における良評価の数
61	2 年次後期可獲得数	2 年次後期における可評価の数
62	2 年次後期不可獲得数	2 年次後期における不可評価の数
63	2 年次後期失格獲得数	2 年次後期における失格評価の数
64	2 年次後期必修科目不合格	2 年次後期の必修科目で不可・失格評価を受けた数
65	2 年次後期 $\#(S_{low})$	2 年次後期においてクラス内偏差値が 40 未満の科目数
66	2 年次後期 $\#(S_{mid})$	2 年次後期においてクラス内偏差値が 40 以上 60 未満の科目数
67	2 年次後期 $\#(S_{high})$	2 年次後期においてクラス内偏差値が 60 以上の科目数
68	2 年次後期 $\#(S_{att})$	2 年次後期に $(s_{l,att} < 40) \wedge (1 \leq \sigma_l) \wedge (e_l \neq X)$ となる科目数

以上

表 3.14: 目的変数

変数名	意味
要注意学生	要注意学生かどうか（TRUE, FALSE）

第4章　検証実験

本章では，前章の方法で生成した変数を入力とするベイジアンネットワークを用いて要注意学生の推定を行う．ベイジアンネットワークに与える変数は CFS によって選択し，ウォード法を用いてクラスタ数が 4 になるように離散化する．また，ベイジアンネットワークの構造は NBC，モデル精度の計算は交差検証とホールドアウト検証で行う．モデル評価は，2 値分類問題における正例クラスを要注意学生とみなして正解率，再現率，適合率，F 値を計算することにより行う．

なお，本研究における推定とは，学生データを用いて将来的に学生が要注意学生になるかどうかの未来予測を指すが，用いるデータはすべて過去のデータであり，未来予測といえども検証データを将来のデータと見なした検証実験であることに注意されたい．しかし，将来的に要注意学生になると予測される学生に対して早期の修学指導を行い，そのような学生を支援するシステムの実用化を目指す上で，どのような学生が将来的に留年・退学してしまうのかを実験を通して調査することは重要なポイントになると考えられる．

4.1　実験概要

本節では実験の概要について述べる．

4.1.1　実験環境

変数生成にあたってはレコード形式のデータを容易に扱える点から Microsoft 社の Excel 2010 を，CFS による特徴選択に関しては Waikato 大学で開発された Weka[29] を, 変数の離散化，検証には NTT データ数理システム社の VMS(Visual Mining Studio)[30], BAYONET[31] をそれぞれ利用した．

4.1.2　説明変数の選択と離散化

モデルの精度は説明変数によって変化するため，どの説明変数を用いてモデルを構築するかが重要となる．例えば，表 3.13 における番号 1–17 の説明変数だけを用いれば，1 年次前期が終了した段階での要注意学生の推定を行うことができるため，要注意学生の早期発見が期待できるが，2 年次後期までの説明変数を用いたときに比べて推定精度が劣ると考えられる．特に，2 年次のデータを用いる場合は説明変数が多くなるが，変数が多いほど良い推定結果が得られるとは限らず，逆にノイズとなる可能性もある．そこで，本研究では CFS を用いて変数選択を行い，冗長な変数の削除を行っている．変数選択の結果は，章末の表 4.18, 表 4.19, 表 4.20, 表 4.21 に示した．これらの表において,「1」は CFS によって当該の変数が選択されたことを示している．

また，表 3.13 のすべての変数は連続型であり，ベイジアンネットワークに与える入力には不適である．そのため，本研究ではウォード法を用いて各説明変数をクラスタ数 4 となるように離散化している．

4.1.3　検証と評価

従来研究においては A 年度に入学した学生，B 年度に入学した学生すべての推定対象学生のデータを用いてモデル構築を行い，Leave-one-out 法により検証を行っていた．しかし，この方法では各学習データに

対する推定精度を平均するため，未知のデータに対するモデルの汎化能力を確かめることができない．このことを確かめるため，A 年度に入学した学生のデータで学習したモデルから B 年度に入学した学生で要注意学生となる者を推定する，逆に B 年度に入学した学生のデータで学習したモデルから A 年度に入学した学生で要注意学生となる者を推定する，というホールドアウト検証を行う．ホールドアウト検証とは，学習データと検証データを完全に分けて行う検証法のことを指す．他にも，同一年度に入学した学生の中でも将来的に要注意学生となる者の傾向が異なる可能性がある．そこで，各入学年度の学生をランダムに 3 分割（A-1, A-2, A-3 と B-1, B-2, B-3）し，例えばデータセット A-1 により学習したモデルで A-2 中の要注意学生を推定する，というホールドアウト検証を行う．

推定結果 $P(S_{rg}|e)$ は，要注意学生かどうかを示すノードの確率変数を S_{rg} とすると，次のように表される．

$$P(S_{rg}|e) = \begin{cases} TRUE & P(S_{rg}|e) \geq p_{th} \\ FALSE & \text{otherwise} \end{cases} \tag{4.1}$$

ここで，p_{th} は閾値を示す．閾値は，より柔軟な推定を行うため，50%, 30%, 事前確率（全体に対する実際の留年退学者の割合）の 3 パターン用意した．最後に，この推定結果と真値を用いて 2 値分類問題を解くことで正解率，再現率，適合率，F 値の計算を行う．将来的に実用化を目指すにあたって，要注意学生を発見すること，すなわち再現率の高いモデルを採用することは大切なことである．しかし，本研究では，教職員の修学指導にかかる負担も減らすべく，指導の対象となる学生をより絞り込むことができる，適合率の高いモデルも重視したい．そこで，この 2 つの評価指標を同時に扱うことができる F 値が高いモデルを，本研究における良い推定モデルと位置付けることとした．

また，本研究では推定を行う時期による推定精度の違いを確かめるため，図 4.1 に示すように 1 年次前期まで，1 年次後期まで，2 年次前期まで，2 年次後期までの 4 つの時期で得られる説明変数を用いた推定を行う．表 3.13 の変数番号でいえば「1 年次前期まで」は番号 1–17 に,「1 年次後期まで」は番号 1–34 に,「2 年次前期まで」は番号 1–51 に，そして「2 年次前期まで」は番号 1–68 にそれぞれ該当する．

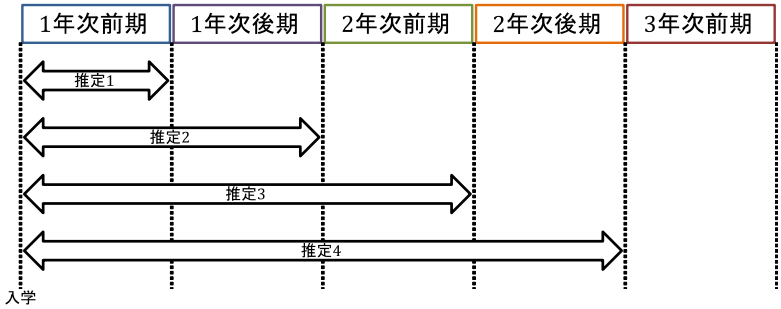


図 4.1: 推定時期

4.2　Leave-one-out 法による検証

本節では，推定の時期を 4 つに，データセットを 3 つに分けた上で，Leave-one-out 法による検証を行った結果を従来手法と比較して示す．4 つの時期とは先述した通り「1 年次前期まで」,「1 年次後期まで」,「2 年次前期まで」,「2 年次後期まで」であり，3 つのデータセットとは「A 年度入学の学生と B 年度入学の学生に関するデータ」,「A 年度入学の学生に関するデータ」,「B 年度入学の学生に関するデータ」のことを指し，以降は便宜上「A+B」,「A」,「B」とそれぞれ表すことがある．時期ごとの要注意学生推定精度を確認するために，表 3.13 の変数 1–17, 1–34, 1–51, 1–68 から CFS により選択した変数を入力とする NBC を構築し，そのモデルをもとに推定した結果を表 4.1, 表 4.2, 表 4.3, 表 4.4 に示す．さらに，実際に構築された推定モデルの例として，1 年次前期終了時に「A+B」を用いて構築された NBC を図 4.2 に示す.

1 年次前期までのデータを用いた推定

表 4.1: 推定精度（1 年次前期終了時点）

データ	閾値	正解率			再現率			適合率			F 値
		対象	的中	割合	対象	的中	割合	対象	的中	割合	
A+B	50%	302	265	88%	25	10	40%	32	10	31%	0.351
	30%	302	260	86%	25	11	44%	39	11	28%	0.344
	8.3%	302	236	78%	25	13	52%	67	13	19%	0.283
A	50%	154	137	89%	13	4	31%	12	4	33%	0.320
	30%	154	134	87%	13	4	31%	15	4	27%	0.286
	8.4%	154	130	84%	13	8	62%	27	8	30%	0.400
B	50%	148	137	93%	12	2	17%	3	2	67%	0.267
	30%	148	129	87%	12	4	33%	15	4	27%	0.296
	8.1%	148	102	69%	12	6	50%	46	6	13%	0.207

1 年次後期までのデータを用いた推定

表 4.2: 推定精度（1 年次後期終了時点）

データ	閾値	正解率			再現率			適合率			F 値
		対象	的中	割合	対象	的中	割合	対象	的中	割合	
A+B	50%	302	264	87%	25	13	52%	39	13	33%	0.406
	30%	302	263	87%	25	16	64%	46	16	35%	0.451
	8.3%	302	253	84%	25	18	72%	60	18	30%	0.424
A	50%	154	139	90%	13	7	54%	16	7	44%	0.483
	30%	154	139	90%	13	7	54%	16	7	44%	0.483
	8.4%	154	140	91%	13	9	69%	19	9	47%	0.563
B	50%	148	136	92%	12	6	50%	12	6	50%	0.500
	30%	148	130	88%	12	6	50%	18	6	33%	0.400
	8.1%	148	123	83%	12	7	58%	27	7	26%	0.359

2 年次前期までのデータを用いた推定

表 4.3: 推定精度（2 年次前期終了時点）

データ	閾値	正解率			再現率			適合率			F 値
		対象	的中	割合	対象	的中	割合	対象	的中	割合	
A+B	50%	302	270	89%	25	15	60%	37	15	41%	0.484
	30%	302	269	89%	25	16	64%	40	16	40%	0.492
	8.3%	302	262	87%	25	19	76%	53	19	36%	0.487
A	50%	154	140	91%	13	8	62%	17	8	47%	0.533
	30%	154	140	91%	13	9	69%	19	9	47%	0.563
	8.4%	154	133	86%	13	9	69%	26	9	35%	0.462
B	50%	148	135	91%	12	3	25%	7	3	43%	0.316
	30%	148	134	91%	12	4	33%	10	4	40%	0.364
	8.1%	148	130	88%	12	8	67%	22	8	36%	0.471

2 年次後期までのデータを用いた推定

表 4.4: 推定精度（2 年次後期終了時点）

データ	閾値	正解率			再現率			適合率			F 値
		対象	的中	割合	対象	的中	割合	対象	的中	割合	
A+B	50%	302	281	93%	25	20	80%	36	20	56%	0.656
	30%	302	275	91%	25	20	80%	42	20	48%	0.597
	8.3%	302	268	89%	25	20	80%	49	20	41%	0.541
A	50%	154	143	93%	13	10	77%	18	10	56%	0.645
	30%	154	141	92%	13	10	77%	20	10	50%	0.606
	8.4%	154	139	90%	13	10	77%	22	10	45%	0.571
B	50%	148	131	89%	12	6	50%	17	6	35%	0.414
	30%	148	132	89%	12	7	58%	18	7	39%	0.467
	8.1%	148	126	85%	12	8	67%	26	8	31%	0.421

考察

まず，従来手法で得られた推定精度（F 値）との比較結果を，表 4.5 に示す．従来研究では，データセット「A+B」に対する検証を行っているため，提案手法の比較対象もデータセット「A+B」における推定結果である．2 年次前期終了時の推定以外は精度の改善がみられる．本提案のために生成した変数はどの時期においても選択されており，要注意学生の推定において有用な変数であるといえる．

次に，データセット別の結果の相違について述べる．データセット「A+B」,「A」に関しては，推定の時期が遅くなるに従って F 値が高い値を示すが，データセット「B」に関しては 1 年次後期終了段階における推定精度が最高値を示した．また，同一推定時期における F 値は，データセット「B」よりも「A」のほうが高い．つまり，表 3.13 にある説明変数は「B」における要注意学生よりも「A」における要注意学生の説

明に適していることになる．選択された変数を詳しく見ると，データセット「A」からは，成績評価「優」「可」の獲得数や各クラスにおいて平均的な成績を修めた講義が多いことを示す $\#(S_{mid})$ が選択されている．すなわち，A 年度入学生の傾向としては成績が普通，もしくは良くても将来的に留年や退学をしてしまう学生が多いということが分かる．このことから，この 2 年度分のデータに限っていえば，1 年次 GPA によるフィルタで除外されないが，普通以上の成績でもない準要注意学生を発見する仕組みが必要になると考えられる．

表 4.5: 従来手法との比較

推定時期	従来手法		提案手法
1 年次前期終了時	0.348	$<$	0.351
1 年次後期終了時	0.394	$<$	0.451
2 年次前期終了時	0.554	$>$	0.492
2 年次後期終了時	0.600	$<$	0.656

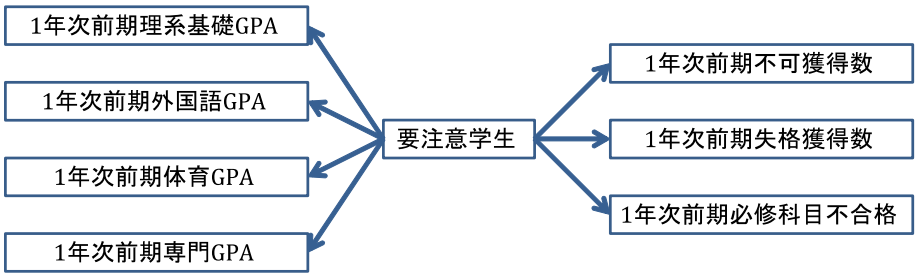


図 4.2: 「A+B」において，1 年次前期終了時に構築される推定モデル

4.3　ホールドアウト法による検証 1

本節と次節では，ホールドアウト検証を行った結果を示す．これは，年度が変わると授業のカリキュラムや担当教員が変更される可能性があることや，同一年度の入学生でも将来的に要注意学生となる学生の傾向が異なる場合があることを考慮し，予測モデルがその変化，すなわち未知の学生群にどの程度対応できるかを調べることを目的としている．この目的を達成するため，本節と次節ではデータセットの扱い方を変更する．

本節では，入学年度の違いに対する汎化能力を確認するために，A 年度に入学した学生データから構築したベイジアンネットワークによって B 年度における要注意学生をどの程度推定できるかを調査する．また，逆に B 年度に入学した学生データから構築したベイジアンネットワークによって A 年度における要注意学生をどの程度推定できるかを調査する．

また，推定の時期は，図 4.1 で示した 4 つの時期に分ける．すなわち，表 3.13 の変数番号 1–17, 1–34, 1–51, 1–68 から CFS により選択した変数を入力とするベイジアンネットワークを構築する．結果は，表 4.6, 表 4.7, 表 4.8, 表 4.9 に示す．

1 年次前期までのデータを用いた推定

表 4.6: 推定精度（1 年次前期終了時点）

学習	検証	閾値	正解率			再現率			適合率			F 値
			対象	的中	割合	対象	的中	割合	対象	的中	割合	
A	B	50%	148	128	86%	12	3	25%	14	3	21%	0.231
		30%	148	127	86%	12	5	42%	19	5	26%	0.323
		8.1%	148	119	80%	12	7	58%	31	7	23%	0.326
B	A	50%	154	142	92%	13	1	8%	1	1	100%	0.143
		30%	154	138	90%	13	3	23%	9	3	33%	0.273
		8.4%	154	134	87%	13	3	23%	13	3	23%	0.231

1 年次後期までのデータを用いた推定

表 4.7: 推定精度（1 年次後期終了時点）

学習	検証	閾値	正解率			再現率			適合率			F 値
			対象	的中	割合	対象	的中	割合	対象	的中	割合	
A	B	50%	148	128	86%	12	7	58%	22	7	32%	0.412
		30%	148	122	82%	12	8	67%	30	8	27%	0.381
		8.1%	148	116	78%	12	9	75%	38	9	24%	0.360
B	A	50%	154	136	88%	13	6	46%	17	6	35%	0.400
		30%	154	130	84%	13	8	62%	27	8	30%	0.400
		8.4%	154	117	76%	13	8	62%	40	8	20%	0.302

2 年次前期までのデータを用いた推定

表 4.8: 推定精度（2 年次前期終了時点）

学習	検証	閾値	正解率			再現率			適合率			F 値
			対象	的中	割合	対象	的中	割合	対象	的中	割合	
A	B	50%	148	134	91%	12	5	42%	12	5	42%	0.417
		30%	148	130	88%	12	6	50%	18	6	33%	0.400
		8.1%	148	119	80%	12	7	58%	31	7	23%	0.326
B	A	50%	154	144	94%	13	6	46%	9	6	67%	0.545
		30%	154	143	93%	13	7	54%	12	7	58%	0.560
		8.4%	154	138	90%	13	11	85%	25	11	44%	0.579

2 年次後期までのデータを用いた推定

表 4.9: 推定精度（2 年次後期終了時点）

学習	検証	閾値	正解率			再現率			適合率			F 値
			対象	的中	割合	対象	的中	割合	対象	的中	割合	
A	B	50%	148	128	86%	12	9	75%	26	9	35%	0.474
		30%	148	125	84%	12	9	75%	29	9	31%	0.439
		8.1%	148	120	81%	12	9	75%	34	9	26%	0.391
B	A	50%	154	144	94%	13	9	69%	15	9	60%	0.643
		30%	154	144	94%	13	10	77%	17	10	59%	0.667
		8.4%	154	138	90%	13	11	85%	25	11	44%	0.579

考察

「A」,「B」に対して Leave-one-out 法を適用した場合と比較するため，表 4.1, 表 4.2, 表 4.3, 表 4.4 と表 4.6, 表 4.7, 表 4.8, 表 4.9 から各時期において F 値が最大となったものを抜き出すと，次の表 4.10 を得る．

表 4.10: 推定精度の比較

推定時期	データセット A			データセット B		
	Leave-one-out		ホールドアウト	Leave-one-out		ホールドアウト
1 年次前期終了時	0.400	>	0.326	0.296	>	0.273
1 年次後期終了時	0.565	>	0.412	0.500	>	0.400
2 年次前期終了時	0.563	>	0.417	0.471	<	0.579
2 年次後期終了時	0.645	>	0.474	0.467	<	0.667

不等号は，8 対の F 値の大小関係を示しており，うち 6 対において Leave-one-out 法による検証結果が上回っている．一般に，推定精度はホールドアウト法よりも交差検証法のほうが勝ることが多いため，2 対では逆の検証結果となったことになる，このことから，入学年度の違いに対する推定モデルの汎化能力にはまだ課題が残っているものの，希望の持てる結果となったといえる．

4.4　ホールドアウト法による検証 2

本節では前節で述べた通り，同一年度に入学した学生間の傾向の違いに対する要注意学生推定モデルの汎化能力を確認するために，A 年度入学の学生に関するデータを A-1, A-2, A-3 の 3 つに，B 年度入学の学生に関するデータを B-1, B-2, B-3 の 3 つに分割し，一つのデータセットから構築したベイジアンネットワークによって他のデータセットの要注意学生をどの程度推定できるかを調査する．具体的には，入学年度ごとに閉じたホールドアウト検証を行うため，推定の時期別に 12 通りの実験結果を得る．推定の時期は，図 4.1 で示した 4 つの時期に分ける．すなわち，表 3.13 の変数番号 1–17, 1–34, 1–51, 1–68 から CFS により選択した変数を入力とするベイジアンネットワークを構築する．結果は，表 4.12, 表 4.13, 表 4.14, 表 4.15 に示す．なお，データセットを分割した結果は表 4.11 の通りである．

また，表 4.12, 表 4.13 において，F 値が inf や 0 となっている箇所がある．これは，分割後のデータセットにおける要注意学生の数が少ないことに起因している．そのため，適切なデータセットの分割数とそのサイズについては検討の余地がある．

表 4.11: 検証 2 で用いるデータセット

データセット	学生数	要注意学生数
A-1	51	5
A-2	52	2
A-3	51	6
B-1	50	5
B-2	49	3
B-3	49	4

表 4.12: 推定精度（1 年次前期終了時点）

学習	検証	閾値	正解率			再現率			適合率			F 値
			対象	的中	割合	対象	的中	割合	対象	的中	割合	
A-1	A-2	50%	52	46	88%	2	1	50%	6	1	17%	0.250
		30%	52	43	83%	2	1	50%	9	1	11%	0.182
		3.8%	52	34	65%	2	1	50%	18	1	6%	0.100
	A-3	50%	51	41	80%	6	0	0%	4	0	0%	inf
		30%	51	41	80%	6	2	33%	8	2	25%	0.286
		11.2%	51	38	75%	6	2	33%	11	2	18%	0.235
A-2	A-3	50%	51	46	90%	6	1	17%	1	1	100%	0.286
		30%	51	46	90%	6	1	17%	1	1	100%	0.286
		11.2%	51	44	86%	6	1	17%	3	1	33%	0.222
	A-1	50%	51	45	88%	5	0	0%	1	0	0%	inf
		30%	51	45	88%	5	0	0%	1	0	0%	inf
		9.8%	51	42	82%	5	1	20%	6	1	17%	0.182
A-3	A-1	50%	51	46	90%	5	2	40%	4	2	50%	0.444
		30%	51	45	88%	5	2	40%	5	2	40%	0.400
		9.8%	51	41	80%	5	2	40%	9	2	22%	0.286
	A-2	50%	52	45	87%	2	1	50%	7	1	14%	0.222
		30%	52	43	83%	2	1	50%	9	1	11%	0.182
		3.8%	52	31	60%	2	2	100%	23	2	9%	0.160
B-1	B-2	50%	49	45	92%	3	0	0%	1	0	0%	inf
		30%	49	44	90%	3	0	0%	2	0	0%	inf
		6.1%	49	22	45%	3	2	67%	28	2	7%	0.129
	B-3	50%	49	43	88%	4	0	0%	2	0	0%	inf
		30%	49	43	88%	4	1	25%	4	1	25%	0.250
		8.2%	49	31	63%	4	2	50%	18	2	11%	0.182
B-2	B-3	50%	49	40	82%	4	0	0%	5	0	0%	inf
		30%	49	33	67%	4	1	25%	14	1	7%	0.111
		8.2%	49	32	65%	4	3	75%	19	3	16%	0.261
	B-1	50%	50	37	74%	5	0	0%	8	0	0%	inf
		30%	50	34	68%	5	0	0%	11	0	0%	inf
		10%	50	31	62%	5	1	20%	16	1	6%	0.095
B-3	B-1	50%	50	41	82%	5	1	20%	6	1	17%	0.182
		30%	50	31	62%	5	2	40%	18	2	11%	0.174
		10%	50	31	62%	5	2	40%	18	2	11%	0.174
	B-2	50%	49	44	90%	3	1	33%	4	1	25%	0.286
		30%	49	39	80%	3	2	67%	11	2	18%	0.286
		6.1%	49	39	80%	3	2	67%	11	2	18%	0.286

表 4.13: 推定精度（1 年次後期終了時点）

学習	検証	閾値	正解率			再現率			適合率			F 値
			対象	的中	割合	対象	的中	割合	対象	的中	割合	
A-1	A-2	50%	52	45	87%	2	1	50%	7	1	14%	0.222
		30%	52	44	85%	2	1	50%	8	1	13%	0.200
		3.8%	52	39	75%	2	1	50%	13	1	8%	0.133
	A-3	50%	51	43	84%	6	2	33%	6	2	33%	0.333
		30%	51	40	78%	6	2	33%	9	2	22%	0.267
		11.2%	51	41	80%	6	3	50%	10	3	30%	0.375
A-2	A-3	50%	51	45	88%	6	0	0%	0	0	inf	0.000
		30%	51	45	88%	6	0	0%	0	0	inf	0.000
		11.2%	51	46	90%	6	1	17%	1	1	100%	0.286
	A-1	50%	51	45	88%	5	0	0%	1	0	0%	inf
		30%	51	44	86%	5	0	0%	2	0	0%	inf
		9.8%	51	43	84%	5	0	0%	3	0	0%	inf
A-3	A-1	50%	51	45	88%	5	2	40%	5	2	40%	0.400
		30%	51	46	90%	5	3	60%	6	3	50%	0.545
		9.8%	51	45	88%	5	3	60%	7	3	43%	0.500
	A-2	50%	52	42	81%	2	0	0%	8	0	0%	inf
		30%	52	40	77%	2	0	0%	10	0	0%	inf
		3.8%	52	35	67%	2	2	100%	19	2	11%	0.190
B-1	B-2	50%	49	43	88%	3	0	0%	3	0	0%	inf
		30%	49	43	88%	3	1	33%	5	1	20%	0.250
		6.1%	49	31	63%	3	1	33%	17	1	6%	0.100
	B-3	50%	49	40	82%	4	0	0%	5	0	0%	inf
		30%	49	36	73%	4	0	0%	9	0	0%	inf
		8.2%	49	29	59%	4	3	75%	22	3	14%	0.231
B-2	B-3	50%	49	43	88%	4	2	50%	6	2	33%	0.400
		30%	49	42	86%	4	2	50%	7	2	29%	0.364
		8.2%	49	26	53%	4	2	50%	23	2	9%	0.148
	B-1	50%	50	42	84%	5	0	0%	3	0	0%	inf
		30%	50	40	80%	5	2	40%	9	2	22%	0.286
		10%	50	20	40%	5	4	80%	33	4	12%	0.211
B-3	B-1	50%	50	41	82%	5	1	20%	6	1	17%	0.182
		30%	50	31	62%	5	2	40%	18	2	11%	0.174
		10%	50	31	62%	5	2	40%	18	2	11%	0.174
	B-2	50%	49	44	90%	3	1	33%	4	1	25%	0.286
		30%	49	39	80%	3	2	67%	11	2	18%	0.286
		6.1%	49	39	80%	3	2	67%	11	2	18%	0.286

表 4.14: 推定精度（2 年次前期終了時点）

学習	検証	閾値	正解率			再現率			適合率			F 値
			対象	的中	割合	対象	的中	割合	対象	的中	割合	
A-1	A-2	50%	52	46	88%	2	2	100%	8	2	25%	0.400
		30%	52	44	85%	2	2	100%	10	2	20%	0.333
		3.8%	52	34	65%	2	2	100%	20	2	10%	0.182
	A-3	50%	51	45	88%	6	3	50%	6	3	50%	0.500
		30%	51	44	86%	6	3	50%	7	3	43%	0.462
		11.2%	51	43	84%	6	4	67%	10	4	40%	0.500
A-2	A-3	50%	51	46	90%	6	2	33%	3	2	67%	0.444
		30%	51	46	90%	6	2	33%	3	2	67%	0.444
		11.2%	51	46	90%	6	2	33%	3	2	67%	0.444
	A-1	50%	51	47	92%	5	2	40%	3	2	67%	0.500
		30%	51	45	88%	5	2	40%	5	2	40%	0.400
		9.8%	51	45	88%	5	2	40%	5	2	40%	0.400
A-3	A-1	50%	51	49	96%	5	4	80%	5	4	80%	0.800
		30%	51	49	96%	5	4	80%	5	4	80%	0.800
		9.8%	51	45	88%	5	4	80%	9	4	44%	0.571
	A-2	50%	52	47	90%	2	2	100%	7	2	29%	0.444
		30%	52	46	88%	2	2	100%	8	2	25%	0.400
		3.8%	52	40	77%	2	2	100%	14	2	14%	0.250
B-1	B-2	50%	49	38	78%	3	1	33%	10	1	10%	0.154
		30%	49	37	76%	3	1	33%	11	1	9%	0.143
		6.1%	49	31	63%	3	2	67%	19	2	11%	0.182
	B-3	50%	49	45	92%	4	2	50%	4	2	50%	0.500
		30%	49	43	88%	4	2	50%	6	2	33%	0.400
		8.2%	49	41	84%	4	3	75%	10	3	30%	0.429
B-2	B-3	50%	49	46	94%	4	1	25%	1	1	100%	0.400
		30%	49	41	84%	4	2	50%	8	2	25%	0.333
		8.2%	49	30	61%	4	2	50%	19	2	11%	0.174
	B-1	50%	50	42	84%	5	2	40%	7	2	29%	0.333
		30%	50	42	84%	5	2	40%	7	2	29%	0.333
		10%	50	28	56%	5	4	80%	25	4	16%	0.267
B-3	B-1	50%	50	38	76%	5	2	40%	11	2	18%	0.250
		30%	50	38	76%	5	2	40%	11	2	18%	0.250
		10%	50	31	62%	5	5	100%	24	5	21%	0.345
	B-2	50%	49	40	82%	3	2	67%	10	2	20%	0.308
		30%	49	39	80%	3	2	67%	11	2	18%	0.286
		6.1%	49	38	78%	3	2	67%	12	2	17%	0.267

表 4.15: 推定精度（2 年次後期終了時点）

学習	検証	閾値	正解率			再現率			適合率			F 値
			対象	的中	割合	対象	的中	割合	対象	的中	割合	
A-1	A-2	50%	52	44	85%	2	2	100%	10	2	20%	0.333
		30%	52	43	83%	2	2	100%	11	2	18%	0.308
		3.8%	52	36	69%	2	2	100%	18	2	11%	0.200
	A-3	50%	51	45	88%	6	5	83%	10	5	50%	0.625
		30%	51	45	88%	6	5	83%	10	5	50%	0.625
		11.2%	51	44	86%	6	5	83%	11	5	45%	0.588
A-2	A-3	50%	51	46	90%	6	2	33%	3	2	67%	0.444
		30%	51	46	90%	6	2	33%	3	2	67%	0.444
		11.2%	51	46	90%	6	3	50%	5	3	60%	0.545
	A-1	50%	51	47	92%	5	2	40%	3	2	67%	0.500
		30%	51	45	88%	5	2	40%	5	2	40%	0.400
		9.8%	51	42	82%	5	2	40%	8	2	25%	0.308
A-3	A-1	50%	51	49	96%	5	4	80%	5	4	80%	0.800
		30%	51	49	96%	5	4	80%	5	4	80%	0.800
		9.8%	51	48	94%	5	4	80%	6	4	67%	0.727
	A-2	50%	52	46	88%	2	2	100%	8	2	25%	0.400
		30%	52	45	87%	2	2	100%	9	2	22%	0.364
		3.8%	52	42	81%	2	2	100%	12	2	17%	0.286
B-1	B-2	50%	49	41	84%	3	2	67%	9	2	22%	0.333
		30%	49	41	84%	3	3	100%	11	3	27%	0.429
		6.1%	49	33	67%	3	3	100%	19	3	16%	0.273
	B-3	50%	49	44	90%	4	2	50%	5	2	40%	0.444
		30%	49	44	90%	4	3	75%	7	3	43%	0.545
		8.2%	49	43	88%	4	4	100%	10	4	40%	0.571
B-2	B-3	50%	49	44	90%	4	1	25%	3	1	33%	0.286
		30%	49	43	88%	4	1	25%	4	1	25%	0.250
		8.2%	49	40	82%	4	2	50%	9	2	22%	0.308
	B-1	50%	50	44	88%	5	1	20%	3	1	33%	0.250
		30%	50	41	82%	5	1	20%	6	1	17%	0.182
		10%	50	40	80%	5	3	60%	11	3	27%	0.375
B-3	B-1	50%	50	37	74%	5	2	40%	12	2	17%	0.235
		30%	50	38	76%	5	4	80%	15	4	27%	0.400
		10%	50	33	66%	5	5	100%	22	5	23%	0.370
	B-2	50%	49	42	86%	3	2	67%	8	2	25%	0.364
		30%	49	39	80%	3	2	67%	11	2	18%	0.286
		6.1%	49	36	73%	3	2	67%	14	2	14%	0.235

表 4.16 にそれぞれの時期において推定精度が最も良くなったときの F 値を示す.

表 4.16: 推定精度まとめ

(a) A 年度，1 年次前期まで

学習 \ 検証	A-1	A-2	A-3
A-1		0.250	0.286
A-2	0.182		0.286
A-3	0.444	0.222	

(b) B 年度，1 年次前期まで

学習 \ 検証	B-1	B-2	B-3
B-1		0.129	0.250
B-2	0.095		0.261
B-3	0.182	0.286	

(c) A 年度，1 年次後期まで

学習 \ 検証	A-1	A-2	A-3
A-1		0.222	0.375
A-2	inf		0.286
A-3	0.545	0.190	

(d) B 年度，1 年次後期まで

学習 \ 検証	B-1	B-2	B-3
B-1		0.250	0.231
B-2	0.286		0.400
B-3	0.182	0.286	

(e) A 年度，2 年次前期まで

学習 \ 検証	A-1	A-2	A-3
A-1		0.400	0.500
A-2	0.500		0.444
A-3	0.800	0.444	

(f) B 年度，2 年次前期まで

学習 \ 検証	B-1	B-2	B-3
B-1		0.182	0.500
B-2	0.333		0.400
B-3	0.345	0.308	

(g) A 年度，2 年次後期まで

学習 \ 検証	A-1	A-2	A-3
A-1		0.333	0.625
A-2	0.500		0.545
A-3	0.800	0.400	

(h) B 年度，2 年次後期まで

学習 \ 検証	B-1	B-2	B-3
B-1		0.429	0.571
B-2	0.375		0.308
B-3	0.400	0.364	

考察

表 4.16(a)–表 4.16(h) の各表における 6 つの F 値の平均と，Leave-one-out 法で得られた最も良い推定精度とを比較すると，表 4.17 のようになる．

表 4.17: 推定精度の比較

推定時期	データセット A			データセット B		
	Leave-one-out		ホールドアウト	Leave-one-out		ホールドアウト
1 年次前期終了時	0.400	＞	0.278	0.296	＞	0.200
1 年次後期終了時	0.563	＞	0.270	0.500	＞	0.272
2 年次前期終了時	0.563	＞	0.515	0.471	＞	0.345
2 年次後期終了時	0.645	＞	0.534	0.467	＞	0.408

どの時期においても，Leave-one-out 法による推定精度がホールドアウト法によるものを上回った．先述したように，この結果は一般的な傾向に合致するものである．検証 1 の結果とあわせて考えると，ある年度に入学した学生のデータを用いて推定モデルを構築し，その次年度に同じ学科に入学する学生が要注意学生になるかどうかの推定を行うときの推定精度は，同年度に入学した他群の学生（例えば，他学科の学生）が要注意学生になるかどうかの推定を行うときの推定精度よりも信頼性があるといえる．ただし，この推定精度はデータの選び方に大きく影響を受けるため，今後十分検討すべき箇所でもある．

4.5　特徴選択結果

本研究では，推定の時期ごとに CFS を用いた特徴選択を行っている．すなわち，表 3.13 の変数 1–17, 1–34, 1–51, 1–68 について，各データセット別に特徴選択を行い，結果として表 4.18, 表 4.19, 表 4.20, 表 4.21 を得た．各表中の「1」は当該の変数が選択されたことを,「計」はその回数の合計をそれぞれ示しており，少なくとも 1 回以上選択された変数のみ記載している．4 つの推定時期を通して，よく選択されている変数は「不可・失格」の獲得数であり，成績の悪さや修学意識の低さが要注意学生の推定に与える影響の大きさが数学的に示されたといえる．

1 年次前期までのデータ

1 年次前期（表 4.18）において，選択された回数が最も多い変数は不可の獲得数，次点が失格の獲得数となった．この 2 つはどちらも単位の修得が認められない不合格評価の数であり，学業不振により留年・退学してしまう学生を推定するにおいて有効な説明変数であることは直感的に理解しやすい．また，これらの採択数は「必修科目の不合格数」の採択数を上回っている．これは，要注意学生を推定するという目的のもとでは，不合格であった科目の属性が必修科目か選択科目かということよりも，単位を落としたという事実のほうが有用であることを示唆している．

加えて，理系基礎科目・専門科目に関する GPA と $\#(S_{att})$ も各 4 回選択されている．入学して間もない時期にも関わらず，出席数が周囲に比べて少ない学生が将来的に要注意学生となるのも,「不合格評価の数」の場合と同じように理解しやすい．

表 4.18: CFS 結果（1 年次前期まで）

番号	変数名	データセット									計
		A+B	A	A-1	A-2	A-3	B	B-1	B-2	B-3	
1	1 年次前期理系基礎 GPA	1	1			1				1	4
2	1 年次前期外国語 GPA	1	1								2
3	1 年次前期人間社会 GPA		1								1
4	1 年次前期体育 GPA	1	1			1					3
5	1 年次前期専門 GPA	1	1	1				1			4
7	1 年次前期秀獲得数					1					1
8	1 年次前期優獲得数			1		1			1		3
11	1 年次前期不可獲得数	1	1	1		1	1	1	1		7
12	1 年次前期失格獲得数	1			1	1	1			1	5
13	1 年次前期必修科目不合格	1		1					1		3
14	1 年次前期 $\#(S_{low})$			1							1
15	1 年次前期 $\#(S_{mid})$		1	1				1			3
16	1 年次前期 $\#(S_{high})$								1		1
17	1 年次前期 $\#(S_{att})$			1		1		1	1		4

1 年次後期までのデータ

表 4.19 に示すように，1 年次後期までのデータに関して CFS により選択された回数が最も多い変数は 1 年次後期における失格の獲得数・専門科目の GPA であった．また，次点で 1 年次前期における不可・失格の数，1 年次後期の人間社会科目に関する GPA・$\#(S_{low})$ が続いた．1 年次前期と同じように「専門科目の GPA」が多く選択されているのは，工業大学である名工大特有の特徴だと推測される．専門科目の出来不出来には勉強の積み重ねがものを言うため，1 年次の専門科目でつまづいてしまうと 2 年次以降に開講される専門科目の授業内容についていくことが難しくなる．成績データの全レコードに対する専門科目のレコード数の割合は，1 年次では約 28%と，2 年次の約 64%と比べて少ないにも関わらずこのような結果になったのは，このことを如実に表しているといえる．

表 4.19: CFS 結果（1 年次後期まで）

番号	変数名	データセット									計
		A+B	A	A-1	A-2	A-3	B	B-1	B-2	B-3	
1	1 年次前期理系基礎 GPA					1				1	2
2	1 年次前期外国語 GPA		1								1
4	1 年次前期体育 GPA		1								1
5	1 年次前期専門 GPA		1					1			2
8	1 年次前期優獲得数			1							1
11	1 年次前期不可獲得数	1	1	1			1				4
12	1 年次前期失格獲得数				1	1	1			1	4
13	1 年次前期必修科目不合格			1							1
14	1 年次前期 $\#(S_{low})$			1							1
15	1 年次前期 $\#(S_{mid})$		1	1				1			3
17	1 年次前期 $\#(S_{att})$			1		1		1			3
19	1 年次後期外国語 GPA	1	1								2
20	1 年次後期人間社会 GPA	1	1				1	1			4
21	1 年次後期体育 GPA								1		1
22	1 年次後期専門 GPA	1	1	1			1	1			5
23	1 年次後期その他 GPA				1						1
24	1 年次後期秀獲得数					1					1
25	1 年次後期優獲得数		1						1		2
26	1 年次後期良獲得数							1			1
27	1 年次後期可獲得数		1	1		1					3
28	1 年次後期不可獲得数	1									1
29	1 年次後期失格獲得数	1	1			1	1		1		5
30	1 年次後期必修科目不合格	1	1	1							3
31	1 年次後期 $\#(S_{low})$	1	1			1	1				4
32	1 年次後期 $\#(S_{mid})$			1							1
33	1 年次後期 $\#(S_{high})$	1	1								2
34	1 年次後期 $\#(S_{att})$	1	1			1					3

2 年次前期までのデータ

2 年次前期までのデータを用いたとき，各データセットから比較的多く選択された変数は，回数が多い順に「2 年次前期の失格評価数」,「1 年次後期の失格評価数」,「2 年次前期の不可評価数」,「2 年次前期 $\#(S_{att})$」,「1 年次前期の失格評価数」$\cdots$ となっている．1 年次までのデータと同様，不可・失格評価に関する変数が目立っている．

表 4.20: CFS 結果（2 年次前期まで）

番号	変数名	データセット									計
		A+B	A	A-1	A-2	A-3	B	B-1	B-2	B-3	
1	1 年次前期理系基礎 GPA									1	1
5	1 年次前期専門 GPA							1			1
11	1 年次前期不可獲得数			1							1
12	1 年次前期失格獲得数				1	1				1	3
13	1 年次前期必修科目不合格									1	1
15	1 年次前期 $\#(S_{mid})$			1							1
20	1 年次後期人間社会 GPA							1			1
21	1 年次後期体育 GPA								1		1
22	1 年次後期専門 GPA			1							1
23	1 年次後期その他 GPA				1						1
24	1 年次後期秀獲得数					1					1
25	1 年次後期優獲得数								1		1
29	1 年次後期失格獲得数	1	1			1	1		1	1	6
30	1 年次後期必修科目不合格	1									1
32	1 年次後期 $\#(S_{mid})$			1						1	2
34	1 年次後期 $\#(S_{att})$					1					1
36	2 年次前期外国語 GPA								1		1
45	2 年次前期不可獲得数	1			1	1	1	1			5
46	2 年次前期失格獲得数	1	1	1	1	1	1	1			7
47	2 年次前期必修科目不合格		1							1	2
48	2 年次前期 $\#(S_{low})$	1						1			2
49	2 年次前期 $\#(S_{mid})$				1			1			2
51	2 年次前期 $\#(S_{att})$	1	1	1		1					4

2 年次後期までのデータ

2 年次終了段階でのデータを用いた場合，それぞれのデータセットから比較的多く選択されている変数は，その回数が多い順に「2 年次前期の失格評価数」,「1 年次後期の失格評価数」,「2 年次後期の失格評価数」,「2 年次後期 $\#(S_{low})$」,「2 年次前期の不可評価数」,「2 年次後期の外国語 GPA」$\cdots$ であり，2 年を通して「不可・失格」評価の数が要注意学生の推定をするための重要な説明変数となっていることが分かる．

表 4.21: CFS 結果（2 年次後期まで）

番号	変数名	データセット									計
		A+B	A	A-1	A-2	A-3	B	B-1	B-2	B-3	
1	1 年次前期理系基礎 GPA									1	1
11	1 年次前期不可獲得数			1							1
12	1 年次前期失格獲得数				1					1	2
13	1 年次前期必修科目不合格									1	1
15	1 年次前期 $\#(S_{mid})$			1							1
16	1 年次前期 $\#(S_{high})$								1		1
20	1 年次後期人間社会 GPA							1			1
21	1 年次後期体育 GPA								1		1
23	1 年次後期その他 GPA				1						1
24	1 年次後期秀獲得数					1					1
29	1 年次後期失格獲得数	1	1			1	1		1	1	6
32	1 年次後期 $\#(S_{mid})$									1	1
34	1 年次後期 $\#(S_{att})$					1					1
45	2 年次前期不可獲得数	1			1		1	1			4
46	2 年次前期失格獲得数	1	1	1	1	1	1	1			7
47	2 年次前期必修科目不合格		1							1	2
49	2 年次前期 $\#(S_{mid})$				1						1
51	2 年次前期 $\#(S_{att})$		1	1		1					3
53	2 年次後期外国語 GPA	1				1	1	1			4
56	2 年次後期専門 GPA	1	1	1							3
59	2 年次後期優獲得数		1								1
60	2 年次後期良獲得数									1	1
62	2 年次後期不可獲得数			1							1
63	2 年次後期失格獲得数	1	1		1	1	1		1		6
64	2 年次後期必修科目不合格		1			1	1				3
65	2 年次後期 $\#(S_{low})$	1	1	1		1		1			5
66	2 年次後期 $\#(S_{mid})$		1			1		1			3
68	2 年次後期 $\#(S_{att})$	1	1			1					3

第5章　むすび

本研究では，過去のある年度 A, B に名工大へ入学した学生 338 名の成績データ，打刻・出欠データにより学習したベイジアンネットワークを用いて，将来的に要注意学生となる学生の推定・検証を行った．ベイジアンネットワークに与える説明変数には GPA 値，各成績評価の獲得数のほかに，各学生が履修しているクラスにおける相対的な成績と出席数を反映するための「クラス内偏差値」と「出席数の偏差値」，および科目の属性（必修科目・選択科目）が与える影響を確認するための「必修科目で不合格となった数」を導入した．また，結果の検証法として交差検証だけでなくホールドアウト検証を用いることで，構築したモデルの汎化能力を確認した．第 2 章ではデータマイニングの要素技術である特徴選択，クラスタリング，ベイジアンネットワーク，評価法について述べ，第 3 章ではデータの概要を説明したのち，要注意学生の定義と推定のための変数生成法に触れた．そして第 4 章では過去のデータの一部を未知のデータとして扱うことで要注意学生推定の検証実験を行った．

その結果，従来研究との比較実験では，2 年次前期を除いて推定精度の改善が確認できた．本研究で生成した説明変数「クラス内偏差値」,「出席数の偏差値」および「必修科目で不合格となった数」は，それぞれの時期において選択されており，要注意学生の推定において有用な変数であるといえる．ベイジアンネットワークを用いた要注意学生の推定システムを，将来的に教育支援に実用化するにあたって対処すべき課題は，推定モデルの汎化能力である．未知のデータに対応することができなければ，発見すべき要注意学生の見逃しや，逆に指導を与えるべきでない学生の呼び出しが起こりうる．そこで，従来研究から用いている交差検証（Leave-one-out）法の他にホールドアウト法による検証を取り入れることで，推定モデルの汎化能力を確かめた．その結果，入学年度の違いに対する汎化能力は，同一年度に入学した学生間の傾向の違いに対する汎化能力よりも高いという知見を得た．すなわち，ある年度 X に入学した学生のデータを用いて推定モデルを構築し，その次年度 X+1 に同じ学科に入学する学生が要注意学生になるかどうかの推定を行うときの推定精度は，同年度 X に入学した他群の学生（例えば，他学科の学生）が要注意学生になるかどうかの推定を行うときの推定精度よりも高いといえる．

今後の課題としては，扱うデータを拡張しての実験が挙げられる．現在扱っているデータは名工大のある学科に入学した学生のものであり，本システムが他の学科や他の大学・教育機関における要注意学生の推定に援用できるかを確かめるには，実際にそれらのデータを用いた検証実験を行うことが望ましいと考えられる．また，他のデータマイニング手法を用いることも考える必要がある．例えば，本研究では用いなかった相関ルール分析を行うことで，要注意学生となる学生の傾向に新たな発見があるかもしれない．そして将来的には，これらの課題を解決しシステムの実用化を目指したい．

謝辞

本研究を進めるにあたって，日頃から多大な御尽力を頂き，ご指導を賜りました名古屋工業大学　舟橋健司　准教授，伊藤宏隆　助教に心から感謝致します．また，本研究の実験データの提供元である，出欠システム及びコースマネジメントシステムの開発に尽力されました，名古屋工業大学　情報基盤センター長　松尾啓志　教授，内匠逸　教授，情報基盤センター教職員の皆様に深く感謝致します．最後に，本研究に多大な御協力を頂きました舟橋研究室諸氏に心から感謝致します．

参考文献

[1] 江谷典子：“創薬データマイニングにおける副作用予測モデルの提案”，情報処理学会研究報告，Vol.160, No.27, pp.1–6, 2014.

[2] 稲邑哲也：“ロボティクスにおけるベイジアンネットの応用”，人工知能学会誌，Vol.17, No.5, pp.546–552, 2002.

[3] 坂本佳愛，岡田将吾，西田豊明：“時系列マルチモーダルデータマイニングを用いたロボットの撮影行動則の獲得”，人工知能学会全国大会論文集，1G3-4, 2011.

[4] 市川昌宏，向井政貴，西尾信彦：“家庭内生活パターンを考慮した電力需要予測手法”，情報処理学会研究報告，Vol.150, No.17, pp.1–5, 2012.

[5] 北栄輔，武井健悟：“ベイジアンネットワークを用いた電力需要予測について”，日本機械学会計算力学講演会講演論文集，No.27, pp.348–349, 2014.

[6] B. Kermanshahi: “ニューラルネットワークの設計と応用”，昭晃堂，1999.

[7] 一杉裕志：“大脳皮質とベイジアンネット”，日本ロボット学会誌，Vol.29, No.5, pp.412–415, 2011.

[8] 原圭司，高橋健一，上田祐彰：“ベイジアンネットワークを用いた授業アンケートからの学生行動モデルの構築と考察”，情報処理学会論文誌，Vol.51, No.4, pp.1215–1226, 2010.

[9] 寿真田崇志，松本哲也，大西昇：“e-Learning におけるベイジアンネットワークを用いた学習者特性の推定”，電子情報通信学会技術研究報告，Vol.106, No.583, pp.203–208, 2007.

[10] 伊藤宏隆，舟橋健司，中野智文，内匠逸，大貫徹：“名古屋工業大学における Moodle の構築と運用”，メディア教育研究，Vol.4, No.2, pp.15–21, 2008.

[11] 佐藤和彦：“修学指導支援のための学生の質的傾向を可視化する手法の検討”，電子情報通信学会技術研究報告，Vol.110, No.334, pp.25–28, 2010.

[12] 加藤利康：“授業支援システムにおける学習分析の展開”，情報処理学会研究報告，Vol.124, No.23, pp.1–7, 2014.

[13] 福島潤一郎，藤原祥隆，前田康成：“確率的推論を基礎とする学習成績マップを利用した対面教育適応化法”，電子情報通信学会技術研究報告，Vol.108, No.470, pp.139–143, 2009.

[14] 伊藤圭佑，舟橋健司，伊藤宏隆：“データマイニングによる『要注意学生』の発見に関する研究”，平成 25 年度名古屋工業大学修士論文，2013.

[15] 平田大智，舟橋健司，伊藤宏隆：“ベイジアンネットワークによる要注意学生の半期毎の発見精度に関する検証実験”，平成 26 年度名古屋工業大学卒業研究論文，2014.

[16] 稲垣諒，舟橋健司，伊藤宏隆：“変数を見直したベイジアンネットワークによる要注意学生の発見手法に関する研究”，平成 26 年度名古屋工業大学卒業研究論文，2014.

[17] 石川博，新見礼彦，白石陽，横山昌平：“データマイニングと集合知”，共立出版，2012.

[18] 元田浩，鷲尾隆：“機械学習とデータマイニング”，人工知能学会誌，Vol.12, No.4, pp.505–512, 1997.

[19] S. Duan, S. Babu: ”Processing Forecasting Queries”, *Proceedings of the 33rd International Conference on Very Large Data Bases*, pp.711–722, 2007.

[20] 植野真臣：“ベイジアンネットワーク”，コロナ社，2013.

[21] 奥野忠一，芳賀敏郎，矢島敬二，奥野千恵子，橋本茂司，古河陽子：“続　多変量解析法”，日科技連出版社，1976.

[22] 宮本定明：“クラスター分析入門”，森北出版，1999.

[23] 天辰次郎，高田徹：“アンケートデータを用いた各種グルーピング方法の比較”，NTT データ数理システムユーザーコンファレンス発表資料，2003.

[24] 繁桝算男，植野真臣，本村陽一：“ベイジアンネットワーク概説”，培風館，2006.

[25] N. Friedman, D. Geiger, M. Goldszmidt: “Bayesian Network Classifiers”, *Machine Learning*, Vol.29, No.2, pp.131–163, 1997.

[26] N. Friedman, M. Goldszmidt: “Building Classifiers using Bayesian Networks”, *Proceedings of the 13th National Conference on Artificial Intelligence*, pp.1277–1284, 1996.

[27] 本村陽一，岩崎弘利：“ベイジアンネットワーク技術”，東京電機大学出版局，2006.

[28] 文部科学省：“学生の中途退学や休学等の状況について”，報道発表，2014.

[29] Machine Learning Group at the University of Waikato: “Weka”, http://www.cs.waikato.ac.nz/ml/weka/, 2016 年 1 月 20 日更新，2016 年 1 月 20 日参照.

[30] NTT データ数理システム：“Visual Mining Studio”, https://www.msi.co.jp/vmstudio/, 2016 年 1 月 20 日更新，2016 年 1 月 20 日参照.

[31] NTT データ数理システム：“BAYONET”, http://www.msi.co.jp/bayonet/, 2016 年 1 月 20 日更新，2016 年 1 月 20 日参照.

発表論文リスト

口頭発表

1. 西脇雅弥，伊藤宏隆，舟橋健司，松尾啓志，内匠逸：“変数縮約したベイジアンネットワークによる要注意学生抽出法”，平成 27 年度電気・電子・情報関係学会東海支部連合大会講演論文集，A3-3，2015.

2. M. Nishiwaki, H. Itoh, K. Funahashi: “A Method of Identifying Students Who Require Guidance Using Bayesian Network”, *Proceedings of IEEE 4th Global Conference on Consumer Electronics*, pp.283–287, 2015.