平成 27 年度 修士論文

# 学生の将来の成績レベル予測の向上及び特徴分析に関する研究

(Study about Improvement of Forecasting Future Students' Academic Level and Analyzing Students' Feature)

指導教員
黒柳　奨　准教授
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　創成シミュレーション工学専攻
平成 26 年度入学　26413505 番

伊藤 雄真

# 目 次

# 第1章　はじめに

名古屋工業大学では, 早期の修学指導を目的とし, コースマネジメントシステム（以下 CMS と呼ぶ）と IC カード出欠管理システムを連携した教員と学生間における双方向型教育支援システムの構築を行っている [1]. CMS は情報技術やインターネットを使って教育指導を支援するシステムであり, 教材の作成支援や課題の提出管理, 小テストの実施, 学生の受講管理などを Web 上で行う機能を持っている．また IC カード出欠管理システムは, IC カード化された学生証を講義室に設置された IC カードリーダにかざすこと（以下この行為を打刻と呼ぶ）により, そのときの打刻情報（ID, 打刻時間, 打刻 IC カードリーダ番号）をリアルタイムで管理サーバに送信し, 蓄積する機能を持っている．教員はこれらのシステムを用いることで, 自分の受け持つ学生の課題結果やレポートの状況, 小テストの結果, 出席率などの学習データにより総合的な成績評価が可能となる．

近年 ICT（情報通信技術）[2][3] の進展により生成・収集・蓄積等ができるデータ（ビッグデータ）が注目されている．ビッグデータでは例えばある通販サイトの購入履歴であったり動画配信サイトの音声や映像，位置情報や電車の乗車履歴など様々な分野のデータが収集され, それらを分析し個々の需要に応じたサービスを提供することで業務の効率化などにつなげることができる [4]．教育の分野においても同様にデータを収集し分析するといった事例は珍しいものではなくなっており，e-learning を用いている大学や企業などが増えている [5]．e-learning とは情報技術によるコミニュケーションネットワーク等を活用した主体的な学習であり, 精度の高い教育データを収集できる．先に述べた CMS や IC カード出欠管理システムも e-learning システムの 1 種であり，名古屋工業大学でもこのシステムで得られたデータを用いてデータマイニングによる分析や予測を行う研究をしている．

研究の背景として，学習指導において教員 1 人あたりの受け持つ学生数が多くなると教員の負担が多くなってしまうという問題が挙げられる．さらに近年の社会の変化に伴い, 学生の目的意識や興味が多様化していることから同一の入学試験を経て入学してきたとしても授業についていけず成績不振者となる学生が現れたり，授業に満足できずふきこぼれてしまう学生が現れてしまうといった問題が発生する．このような学生達に対してできるだけ早く指導を行うことが理想的であるが, 教員の負担が多くなってしまうと学生毎の理解度や学習意欲を把握することが困難となり，そういった学生達を早期に発見することが難しい．従来ではこのような学生の指導は成績が出た後で行っており, 授業への出席率が極端に悪い学生の場合手遅れになってしまっていた．そこでこのような問題の解決方法として名古屋工業大学では CMS や IC カード出欠管理システムによって得られたデータを用いることにより, 先ほど述べたような出席率が極端に悪い学生などを早期に発見し，担当部局に連絡したり, データマイニングを用いることで早期にリスク予知し対応することにより成績不振学生を見つけるといったことを行っている [6].

過去の関連研究 [7][8] において, ある一つの授業における学生の出欠状況や課題提出状況が成績に影響を与えることが証明されており，それらを利用することによって成績予測が可能であることも証明されている．また予測結果を学生に伝えることができれば, 現在置かれている状況を把握させ学習意欲の向上や学習に対する姿勢を改善できるのではないかということを期待し，CMS を用いて Web 上で実装可能であるニューラルネットワーク手法を用いて予測を行う研究 [9] も行われている．またニューラルネットワークでは計算過程を知ることができないことから説得力に欠けてしまうことを危惧し, 出力結果の直観的なわかりやすさと計算過程を参照できるベイジアンネットワークに着目した研究 [10][11] も行われている．これらの研究では学習指導が必要な学生を「要注意学生」と定義し，CMS や IC カード出欠管理システムから得られた成績データ・打刻データを用いてベイジアンネットワークを構築することにより要注意学生を予測・発見し指導対象者を絞ることで教員にかかる負担を少なくすることを目的としている．

本研究では将来の学生の成績レベルを予測した際に打刻による特徴を見つけることを期待しており, また特徴を見つけたとしても予測がしっかりできていなければ正確であるとは言えないので予測的中精度の向上を目指した．今まで予測に用いていたデータは A 年度入学者のデータだけだったが, データ数が少ないとたとえ予測的中精度が高い値を示したとしても他のデータで試した時に同じ精度が出るかわからないため,B 年度入学者のデータを加えて同様の予測を行うことにより確かめた．また更に C 年度入学者のデータを足すことで,2 学年度分のデータを学習データとして, 余りの 1 学年度を評価データとして予測を行うといったことが確認できるため, 考えられる各パターンがどうなるかについても検証を行った．評価データを用いない場合の予測に関しては leave one out 法を用いて検証し, モデルの信頼度がある程度あるものとした．その過程で従来, 予測時期を 2 年前期終了時点にしていたが,1 年次終了時点で予測を行うことにした．これにより 1 学期分のデータが少なくなることで, その分予測的中精度は下がってしまったが, できるだけ早く予測し学生に対して指導を行いたいという点を考慮したほうが良いと考えたためこの時期で予測することにした．また打刻データだけでなく出欠データも一緒に用いることにした．ここで打刻データと出欠データの違いについて説明する．IC カード出欠管理システムによって得られるデータとして打刻データと出欠データがある．打刻データにはある学生がある講義を受ける際にその時の打刻情報がすべて記録されている．出欠データは打刻データによって得られた情報から自動的に出席時と退席時の打刻情報を抽出し出欠状況が生成されている．しかし打刻データにおいて例えばある授業の出席時に複数回打刻するとその時の打刻情報がすべて出席記録として記録されてしまうため余分なデータを含んでしまう問題がある．また IC カード出欠管理システム側の問題として, 講義の開始時あるいは終了時には打刻有効範囲というものが設定されておりその有効範囲内で打刻しなければ, 出欠データにおいて出席扱いされなかったり遅刻早退扱いになってしまう．これらの問題から出欠データを打刻データにより補正を行うことで従来用いていた打刻データより正確なデータとして用いることができるため本研究でもこの補正したデータを使い予測することにした．この補正したデータを用いて先ほど述べた要注意学生を見つける研究も行われている [12]．また従来研究では用いていなかった新たな変数を作成することで更なる予測精度の向上を目指した．本研究で用いる手法としてはベイジアンネットワークを採用している．ベイジアンネットワークは専門的な知識を元に各変数の因

果関係を矢印に見立てた有向グラフで表現し変数間の因果関係を確認することができる．これにより学生の成績がどの変数と因果関係にあるのかを確認することができ，成績レベルを予測した際の特徴として調べることができる．

本論文では, 第 2 章において本研究で用いる手法の概要を述べ, 第 3 章において B 年度入学者のデータを追加し予測的中精度の向上を図った結果を述べる. また第 4 章においてさらに C 年度入学者のデータを追加し 3 章で出した結果の検証及び 3 学年度分のデータより考えられる組み合わせによる予測の結果を述べ, 第 5 章にて本研究の結論と今後の課題を述べる. なお, 本研究に用いる学習データには学生を特定できる情報は一切含まれておらず仮の番号により管理されているため, プライバシーが侵害されることはないことをここに付記する

# 第2章　本研究で用いる手法の理論

本研究では成績予測の手法としてデータマイニング手法の一つであるベイジアンネットワークを採用している. また予測に用いるデータに関しても属性選択したり, 離散化を行っている. 本章では本研究で用いている手法の概要について記す.

## 2.1　ベイジアンネットワーク

ベイジアンネットワーク [13][14][15] は確率変数・有向グラフ構造・条件付き確率の集合によって定義されており，複数の確率変数の間の定性的な依存関係をグラフ構造によって表し，個々の変数の間の定量的な関係を条件付き確率で表した確率モデルである. 最適なベイジアンネットワークモデルを作成するには最適な条件付き確率の推定, 確率変数の選択，有向グラフの獲得が重要となる.

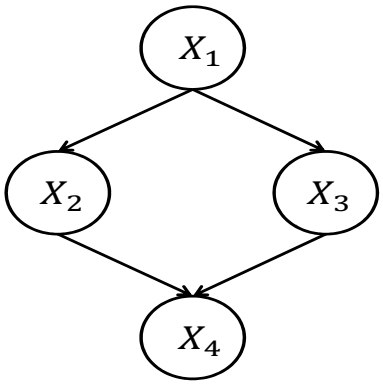


図 2.1: ベイジアンネットワークの例

図 2.1 はベイジアンネットワークの例である. 確率変数 $X_1$ と $X_2$ の関係に着目すると, 条件付き依存性を $X_1 \to X_2$ と表しており, リンク先であるノード ($X_2$) は子ノード、リンク元であるノード ($X_1$) は親ノードとして扱われる. 親ノードが複数ある場合, 子ノード $X_2$ の親ノードの集合を $P_a(X_2)$ と書くと、$X_2$ と $P_a(X_2)$ の間の依存関係は $P(X_2|P_a(X_2))$ という条件付き確率により定量的に表せる. 図 2.1 全体より 4 個の確率変数 $X_1, \cdots, X_4$ のそれぞれを子ノードとして同様に考えた場合, すべての確率変数の同時分布 $P(X_1, \cdots, X_4)$ は

$$\begin{aligned} P(X_1, \cdots, X_4) &= P(X_1|P_a(X_1))P(X_2|P_a(X_2))\cdots P(X_4|P_a(X_4)) \\ &= P(X_1)P(X_2|X_1)P(X_3|X_1)P(X_4|X_2, X_3) \end{aligned} \tag{2.1}$$

と表せる. すべての変数の確率分布は, 同時分布を計算することによって得られるのでベイジアンネットワークはこれを用いることで求めることができる. ベイジアンネットワークによる確率的推論は以下の手順で行われる.

1. 観測された変数の値 $e$（エピデンス）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

この計算により観測情報 $e = X_2 = 1, X_3 = 1$ から事後確率 $P(X_4|e)$ を得ることができる.

事後確率を求めるために, 変数間の局所計算を繰り返しながら確率をネットワーク中に伝搬することにより各変数の確率分布を更新していく計算法を確率伝搬法という. 確率伝搬法の説明のために今回はある複雑な構造を持つモデルの一部分として次の単純なモデルを取り出しその部分における確率計算を考える. 図 2.2 のように $X \to Y \to Z$ の間に依存関係が

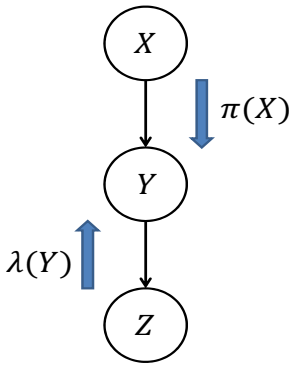


図 2.2: 単純な構造における確率伝搬

あり条件付き確率が与えられているものとする. 計算しようとしているノードを $Y$ とし, 親ノードに与えられる観測情報を $e^+$, 子ノードに与えられる観測情報を $e-$ とする. また $e^+$ と $e^-$ は $Y$ を固定したときに条件付き独立になるため、$\alpha = \frac{1}{P(e^+|e^-)}$ を $Y$ の値によらない正規化された値とすれば求めたい事後確率 $P(Y|e)$ はベイズの定理により

$$P(Y|e) = \alpha P(e^-|Y)P(Y|e^+) \tag{2.2}$$

となる. ここで図 2.2 のように親ノードからの寄与確率値を $P(Y|e^+) = \pi(Y)$ とおくと

$$\pi(Y) = \sum_X P(Y|X)P(X|e^+) \tag{2.3}$$

のように変形できる. このときノード $X$ に親ノードがいない場合は予め用意された事前確率を与え, 観測情報が与えられている場合, その値は決定できる. ノード $X$ に入力がなく, かつノー

ド $X$ に親ノードが存在するとき式 (2.3) を再帰的に適用することによりその値を求めることができる. 同様にノード $Z$ についても考える. 子ノードからの寄与確率値を $P(e^-|Y) = \lambda(Y)$ と置くと

$$\lambda\ (Y) = \sum_Z P(e^-|Y, Z)P(Z|Y) \tag{2.4}$$

となる. 観測情報 $e^-$ は $Y$ の値に関係なく独立であることから

$$\lambda\ (Y) = \sum_Z P(e^-|Z)P(Z|Y) \tag{2.5}$$

ここで $P(Z|Y)$ は条件付き確率表として与えられいることから $P(e^-|Z) = \lambda\ (Z)$ は観測情報が与えられているとき値が決定できる. また, 観測情報がなくそのノードが子ノードを持たない下端のノードの場合は、無情報であることから一様分布確率として $Z$ のいかなる状態について等しい値とする. また, ノードが子ノードを持つ場合, 式 (2.5) を再帰的に適用することで最終的に下端のノードの値を求めることができる. これらを利用することでノード $Y$ の事後確率を求めることができる. よってグラフ構造内のすべてのパスがループを持たないとき任意のノードの事後確率を局所的に求めることができる. これにより複雑な構造を持つモデルであったとしても確率伝搬法に基づいた計算方法により導出できる.

### 2.1.1　最適なベイジアンネットワークモデルの構築

先ほど述べたようにベイジアンネットワークのモデルは有向グラフにより形成される. よってベイジアンネットワークで予測を行った場合, 有向グラフの構造によって予測結果が影響される. ここでベイジアンネットワークによるモデルの構築の際に使われる代表的なグラフ構造について説明する.

#### Naive Bayes

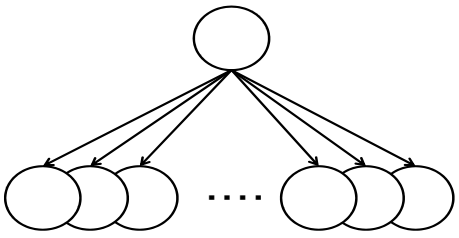

図 2.3: Naive Bayes の例

NaiveBayes [16] はある 1 つのノード（親ノード）から他の全てのノード（子ノード）に向かって矢印が伸びているグラフ構造を持つベイジアンネットワークのことである. またこの時, 親ノードは 1 つのみで子ノード同士はつながっていない. NaiveBayes は Bayes の定理及び「属性はどれも区別なく同様に重要である, またクラス値が与えられれば条件付き独立

である」という仮定により基づく確率モデルの 1 つであり，設計も仮定も非常にシンプルなものであるにもかかわらず実世界において期待したよりもずっとうまく分類できることがわかっている．テキストマイニングとして使われることが多くスパムメールの判別手法として採用されている．

**Tree Augmented Naive Bayes(TAN)**

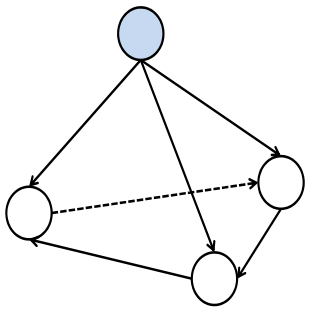

図 2.4: Tree Augmented Network の例

TAN は Naive Bayes と似た基本的構造を有している．Naive Bayes と違う点は目的変数以外の変数がもう一つの変数を親ノードとして持っている点である．条件付き相互情報量により親となるノードが決められる．目的変数を $C$, その他の 2 つの変数を $X, Y$ としたとき目的変数 $C$ が与えられたという条件付き相互情報量は

$$I(X;Y|C) = \sum_{x,y,c} P(x,y,c) log \frac{P(x,y|c)}{P(x|c)P(y|c)} \tag{2.6}$$

で求めることができる．このときの $I(X;Y|C)$ が最大となる $X, Y$ を求めることにより TAN のグラフを決定することができる．本研究では TAN 構造を採用した．

### 2.1.2　確率モデルの評価指標

最適なモデルを構築するにはグラフ構造を見直すだけでなく確率モデルを評価する指標も重要となる．この評価指標は情報基準量とも呼ばれており，代表として赤池情報量基準（Akaike's Information Criterion:以下 AIC と記す）とベイズ情報量基準（Bayesian Information Criterion:以下 BIC と記す）がある．ここではその 2 つについて説明する．

**AIC**

ベイジアンネットワークではほとんどの場合においてデータを基にモデルが構築されるため, 構築に用いるデータにより適合度合（尤度）が変わる．尤度はモデルへの当てはまりの良さを表しており尤度が大きければより観測データとして適していると評価できる．しか

し尤度を最大化してしまうとパラメータ数が増加してしまう傾向を持っておりこのパラメータ数が多いとモデルが複雑になることが分かっている．そのため AIC ではパラメータ数をペナルティと見なしており, あるモデルにおける尤度を $L$, パラメータ数を $k$ と置くと指標値 AIC は下記の式 (2.7) で求めることができる．

$$AIC = -2\log L + 2k \tag{2.7}$$

AIC は小さいほど優秀なモデルであると言うことができ, 式 (2.7) より尤度 $L$ が大きいほど第 1 項の値は小さくなり, またパラメータ数 $k$ が小さいほど第 2 項の値は小さくなるので,AIC の値は小さくなる．

**BIC**

BIC も AIC 同様パラメータ数が多くなるとモデルとしての当てはまり具合は良くなるがその分複雑なモデルとなってしまう．違いは AIC がパラメータ数をペナルティとしていたのに対し,BIC では回帰モデルが多くの項を含みすぎることに対してペナルティを課す部分である．尤度を $L$, 標本数を $n$, パラメータ数を $k$ と表すと指標値 BIC は下記の式 (2.8) で求めることができる．

$$AIC = -2\log L + k\log n \tag{2.8}$$

式 (2.8) によりペナルティ部分である第 2 項もサンプルサイズの関数として表されるので,AIC よりもシンプルなモデルが選ばれやすくなる．本研究では BIC による評価指標を用いて予測を行うことにした．

### 2.1.3　有向グラフの学習

ネットワーク構造のグラフの決定方法としていくつかのアルゴリズムが報告されており, 有向グラフにおいても変数間の因果関係を調べるために最適な構造の決定が必要とされる．ここでは全探索アルゴリズム及び K2 アルゴリズム, また山登り法について説明する．

#### 全探索アルゴリズム

全探索アルゴリズムは最も単純な探索手法であり，全てのグラフ構造を想定しその中で最も良好な構造を選択する．そのため最適な結果が得られるが, 全てのグラフ構造を想定するため確率変数の数に応じて計算時間が指数関数的に増加指定しまう欠点を持っている．例えば確率変数が 3 つのとき, 考えられるネットワーク構造は向きを含め考えると 25 通りとなる．次に確率変数が 5 つのとき, 同様に考えた場合ネットワーク構造は 29281 通り存在することとなる．このように確率変数がたった 2 つ増えただけでネットワーク構造は急増してしまう．この問題は Non-deterministic Polynomial(NP) 問題と呼ばれている．

**K2 アルゴリズム**

K2 アルゴリズムは現在良く知られているベイジアンネットワークの構造学習アルゴリズムであり, 変数間の親子関係を表した全順序関係を用いることでネットワーク構造の数を少なくし探索にかかる時間を減らすことができるアルゴリズムである．このアルゴリズムは以下のことを繰り返している．

1. 各ノード間について親ノードになりえる候補を限定する
2. ある子ノードを 1 つ選び候補となる親ノードを 1 つずつ加えてグラフを作る
3. そのグラフのもとでパラメータを決定し評価する
4. 評価が高くなった時だけ親ノードとして採用する
5. 親ノードとして加える候補がなくなるか加えても評価が高くならなくなったら他の子ノードに移る

**山登り法**

山登り法はゴールの位置を山頂にたとえて、常にゴールに近づく方向 (山を登る方向) へ探索を進めるアルゴリズムである。ヒューリスティック探索を実装する場合、何らかの方法でゴールへの距離 (コスト) を求める必要がある．この時の値を「評価値」といい，評価値を計算する関数を「評価関数 (evaluation function) 」と呼ぶ．評価関数の極致を探索することができ，最も代表的な局所探索法として知られている．本研究では山登り法を採用し, できるだけ良い結果となるものを見つける．

### 2.1.4　検証方法

本研究では予測の検証方法として leave one out 法とホールドアウト法の 2 手法を採用している．その 2 つの手法について簡単に説明する．

**leave one out 法**

標本から 1 つの事例を取り出して評価データとし、残りを学習データとする。そして全事例が 1 回は評価となるように検定を繰り返す検証方法のことを leave one out 法という．

**ホールドアウト法**

標本を 2 分割し，一方を学習データとしてモデル構成に使い，そこで得たモデルを残りのデータ（評価データ）に対しては適用し，モデルの良さを検証する方法をホールドアウト法と呼ぶ．簡単に判別分析ができるが, 標本サイズが大きい必要がある．狭義にはこれをクロスバリデーションという．

## 2.2　データの整理

データマイニングでは大量のデータから有用な知識や情報を見つけることを目的としているため, 従来の統計解析やデータ解析と比べてデータ量が多いことや, データの種類が多いことが特徴として挙げられる．そのためデータマイニングに使うデータとして余分なデータを多く含んでいることが可能性としてはあり得る．またデータマイニングで用いる手法によってはデータをその手法に合った形に直して使用しなければ良い結果が得られないこともあり得る．そこで本研究では最適な結果を出すためにデータの整理を行った．データについての詳しい説明は第 3 章で述べるがここではデータの整理に用いた手法について説明する．

### 2.2.1　属性選択

k 個の特徴量（属性）のベクトルで記述された対象に機械学習を適用させる場合,k 個の特徴量をすべて利用せずその中で有用なものだけ選び出すことを属性選択（特徴選択）と呼ぶ．目的としては「目的変数と無関係な特徴量を使わないことで予測精度を向上させる」ことや「学習された関数を, 定性的に解釈しやすい」などが挙げられる．属性選択をする際には何かしらの指標を設けなければならない．その指標の 1 つとして情報利得（Information gain）が挙げられる．情報利得とは「ある素性が出現したか否か」という情報がクラスに関するあいまいさをどれくらい減少させるかを表したものであり, 確率変数を C とするとエントロピー $H(C)$ は

$$H(C) = -\sum P(c)logP(c) \tag{2.9}$$

で表される.「ある素性が出現した」ことが分かっている場合の条件付きエントロピーは

$$H(C|X_w = 1) = -\sum_c P(c|X_w = 1)logP(c|X_w = 1) \tag{2.10}$$

で表される. また「ある素性が出現しなかった」ことが分かっている場合の条件付き確率は

$$H(C|X_w = 0) = -\sum_c P(c|X_w = 0)logP(c|X_w = 0) \tag{2.11}$$

で表される. 以上より素性 $w$ の情報利得 $IG(w)$ は次のように定義される.

$$IG(w) = H(C) - (P(X_w = 1)H(C|X_w = 0) + P(X_w = 0)H(C|X_w = 0)) \tag{2.12}$$

情報利得を用いた変数の選択は属性数の多い変数で有効であることが多いという特徴を持っている．例えば決定木学習において根ルートから葉ルートまでのパス長の総和を最小にする決定木を作る問題は NP 問題として知られているが, ある程度実用的なサイズの決定木を構築するのであれば再帰的アルゴリズムによって構築できる．この時再帰的アルゴリズムで問題となるのはデータ集合を部分集合に分割するための属性を与えられた属性群からどのようにして選定するかということであり, その指標として情報利得が採用されている．

### 2.2.2　離散化

データには大きく分けて質的データと量的データの 2 種がある．質的データとはデータが数値ではなく, いくつかの項目のいずれかに属しているか否かという形で与えられているものを指し, 量的データとは統計による観測結果が数値により記録されているものを指す．一般的に量的データのような連続した値を持つ情報を解析することは困難である．そのため連続した値を持つデータに対して非連続的な数値に置き換えることを離散化すると言い, これにより近似的な計算を比較的簡単に算出することが可能になる．本研究で用いるデータにおいてもある変数に関して識別数が多いものがあり, そのまま用いるには難しいデータに対しては離散化を行うことで近似的なデータとして使用している．

# 第3章　出欠状況を考慮した将来の学生の成績レベル予測及び特徴分析

従来, 名古屋工業大学の A 年度に入学したある学科の学生 171 人の成績データと打刻データを用いて予測を行っていた．しかしデータのサンプル数が少ないと, 予測結果が良くても本当に有用であるとは言い難い．そこで名古屋工業大学の B 年度に入学した同じ学科の学生 167 人分のデータを加えて予測を行うことにでどのような結果になるのか検証した．またこのときデータを加えるにあたって下記の変更を行った．

1. 予測時期を 2 年前期終了時点から 1 年次終了時点に変更
2. 出欠データの追加

## 予測時期の変更

まず予測時期について説明する．従来研究では 2 年前期終了時点から予測を行い 2 年終了時点の総合 GPA から成績レベルを設定しベイジアンネットワークによる予測を行っていた．この予測では約 8 割の予測的中精度で予測できているが,2 年前期終了時点までのデータから予測を行うのでこの時点である程度成績結果が決まってしまっている．そのため半年後の 2 年終了時点における成績レベルは 2 年前期終了時点における成績結果とそう変りないことが想像できる．また予測時期は 2 年終了時点に設定しているが, これは 3 年次終了時点や卒業時を予測対象として設定すると予測期間が長いため学生が危機感を持たない可能性があるといった点を考慮しているため適切であると考えられる．よって予測を 1 年次終了時点にすることで 1 年後の結果を知ることができ, また学生に対しても早く予測することで学生が落ちこぼれないように早くに対処できることが期待できる．以上のことから予測時期は 2 年前期終了時点から 1 年次終了時点に変更した．

## 出欠データの追加

IC カード出欠管理システムからは打刻データと出欠データを得ることができ, 出欠データは打刻データを元に作成されている．従来, 打刻に関するデータとして打刻データしか用いていなかったが本研究では出欠データも一緒に用いることにした．このとき打刻データと出欠データに関しては第 1 章でも述べたが, 打刻データにおいてある授業の出席時に複数回打刻するとその時の打刻情報がすべて出席記録として記録されてしまうため余分なデータを含んでしまうといった問題があったり, 出欠データにおいて入室時間あるいは退室時間には打

刻有効範囲というものが設定されておりその有効範囲時間内で打刻しなければ, 出席扱いされなかったり遅刻早退扱いになってしまう．そのため例えば授業開始前に早く来すぎてしまい打刻データとしては記録されているのに出欠データでは入室時間の有効範囲時間内で打刻しなかったため出席扱いされていないといった場合があったり, 授業が予定終了時刻より早く終わってしまったため, 退室として打刻した記録が打刻データは記録されているが出欠データでは記録されていないなどの問題がある．このまま予測を行っても元データに不備があったとすると正確なことが言えない恐れがあるため, 出欠データにおいて打刻データから自動的に生成されていない箇所を打刻データにおいて出欠データに使われていない部分のデータを用いて補正してやることにより, 元の出欠データよりも正確なデータを作成し, それを用いることである程度正確なデータとして取り扱う．以上より出欠データを追加することで従来で使っていた打刻データより正確な打刻情報を記録しているデータ（以下出欠補正データと呼ぶ）を用いることで予測的中精度の向上を図る．

## 3.1　用いたデータの概要

本章ではデータとして, ある学科における 2 学年度分のデータを用いており, それぞれのデータを A 年度データ,B 年度データと呼ぶことにする．A 年度データから学生 171 人,B 年度データからは学生 167 人のデータが記録されており, 計 338 人分のデータを用いて予測を行った．また本研究では予測に用いるデータの種類として打刻データ, 出欠データ, 成績データを用いている．

### 打刻データ

本研究で与えられた打刻データは「暗号化学生番号, 教室, 打刻日, 打刻時間」が 1 レコードとなっており,A 年度は約 11 万 2 千レコード,B 年度は約 10 万 6 千レコード,C 年度は約 11 万 7 千レコードのデータが記録されている．ただし, 学生番号は個人が特定されないようあらかじめ暗号化されており, それぞれの学年度に関して学生が所属している学科は同じだがクラスが特定されていないため授業構成により打刻時間帯もさまざまであることを言及しておく．また, 教室についてはすべての打刻に対応しているわけではなく, 記載されていない場合もある. 取り扱うデータ範囲として, 授業に関係の無い打刻 (休日の打刻など) を使用しないためにまず休日に打刻しているデータは無視する. また前期は 4 月〜7 月, 後期は 10 月〜1 月のデータ部分を用いることでクラスによるデータのずれを考えないこととする．.

### 出欠データ

本研究で与えられた出欠データは「暗号化学生番号, 暗号化授業番号, 教室, 打刻した日付（年/月/日）, 入室時間, 退室時間」が 1 レコードとなっており,A 年度は約 7 万 3 千レコード,B 年度は約 7 万 2 千レコード,C 年度は約 7 万 5 千レコードが記録されている．出欠データは打刻データから自動的に生成される．このデータも打刻データ同様学生個人が特定され

るようなことはなく, クラスに関する問題や取り扱うデータの範囲に関して一緒であることを言及しておく.

### 成績データ

本研究で与えられた成績データは「暗号化学生番号, 成績, 暗号化授業番号, 単位, 科目名, 学期」が 1 レコードとなっており,A,B,C 年度それぞれ約 5 千レコードが記録されている．このデータも打刻データ・出欠データ同様のデータに関することが言え, また成績データにおける科目名は実際の授業名ではなく,「専門 1」や「演習 1」のように講義が特定できないようにされている．これも学生個人が推定されないようにする措置であり, 具体的な講義内容はわからないが講義の分野についてはわかるように記録されている．成績データにおける「成績」は秀・優・良・可・不可・失格の 6 つの評価で分けられており, 最も優れている評価が秀でそこから順番に成績評価が低くなり不可・失格に関しては単位が取得できなかったことを表している．不可と失格の違いは, 課題提出やテストを受けていながら単位取得条件を満たすことができなかった場合は成績が不可となり, そのほかの要因（例えば受講登録だけしておいて実際は講義に出席しない, 課題を出さない, テストを受けないなど）の場合は評価ができないため失格となる.

ベイジアンネットワークからモデルを構築するにはある程度の量を有するデータから学習しなければならない. 成績データや出欠補正データに記録されているデータをそのまま用いても情報量に乏しいため満足できるモデルを作成することができない. そこでデータの拡張を行うことで情報量を増やすことにする.

## 3.2　データの拡張

### 出欠補正データによる変数

出欠補正データについて今のままでは「暗号化学生番号, 暗号化授業番号, 教室, 打刻した日付, 入室時間, 退室時間」の 6 変数しかなく, このままで用いても最適なモデルを作成することは難しい. そこで出欠補正データから予測に使えそうな新たな変数を作成する.

まず打刻した日付から前期 4 月〜7 月, 後期 10 月〜1 月の各月毎の打刻を特定することができる. 入室時間・退室時間に記録されているデータの数より打刻回数を求めることができるので, これより前期後期分の月毎打刻回数を変数として用いる．また文献 [17] より下記の式で求められる $D_I$ を用いて不登校学生の早期発見を行っている．$D_I$ はある学生の 2 週間毎の打刻数を比べたときの出席率を表しており,$D_A, D_B$ はある学生における直近 2 週間の打刻回数, さらにその 2 週間前の打刻回数を,$\bar{D_A}, \bar{D_B}$ は全学生における直近 2 週間の打刻回数の平均, さらにその 2 週間前の打刻回数の平均を表している.

$$D_I = \frac{D_A/\bar{D_A}}{D_B/\bar{D_B}} \tag{3.1}$$

$D_I$ が 0.5 を下回り $D_B$ が 4 以上の時, 不登校が疑われる学生として取り扱われる．本研究でもこの $D_I$ に着目した．$D_I$ は 4 週間毎に値を求めることができ前期 4 か月, 後期 4 か月の期間があるため,$D_I$ は A 年度,B 年度それぞれにおいて 13 個の値を導出することができた．これより各 $D_I$ の値が 0.5 以下になっている数を変数として設定した．また $D_I$ の値が 0.5 付近の値の数についても変数として設定し成績レベル予測に関係するのか検証した．そのため $D_I$ の値が 0.5〜0.6 の間にある数,0.6〜0.7 の間にある数による変数も作成した．また各 $D_I$ の値を比べたとき最大差についても考える．この差は最も打刻していた時期と最も打刻していない時期の差を表しており, 平均的な打刻をしている場合 $D_I$ の値は 1 となるので差が 1 以上あるかどうかを変数として設定し成績レベル予測に関係するか検証した．以上より出欠補正データから拡張した変数を全て表 3.1 に列挙しておく．

表 3.1: 出欠補正データより拡張した説明変数

| 番号 | 変数名 | 意味 |
|---|---|---|
| 1 | 1 年 4 月打刻回数 | 1 年 4 月に打刻した回数 |
| 2 | 1 年 5 月打刻回数 | 1 年 5 月に打刻した回数 |
| 3 | 1 年 6 月打刻回数 | 1 年 6 月に打刻した回数 |
| 4 | 1 年 7 月打刻回数 | 1 年 7 月に打刻した回数 |
| 5 | 1 年 10 月打刻回数 | 1 年 10 月に打刻した回数 |
| 6 | 1 年 11 月打刻回数 | 1 年 11 月に打刻した回数 |
| 7 | 1 年 12 月打刻回数 | 1 年 12 月に打刻した回数 |
| 8 | 1 年 1 月打刻回数 | 1 年 1 月に打刻した回数 |
| 9 | 前期 $D_I \leq 0.5$ の出現率 | 前期における各 $D_I$ が 0.5 を下回っている確率 |
| 10 | 前期 $0.5 < D_I \leq 0.6$ の出現率 | 前期における各 $D_I$ が 0.5〜0.6 の間にある確率 |
| 11 | 前期 $0.6 < D_I \leq 0.7$ の出現率 | 前期における各 $D_I$ が 0.6〜0.7 の間にある確率 |
| 12 | 後期 $D_I \leq 0.5$ の出現率 | 後期における各 $D_I$ が 0.5 を下回っている確率 |
| 13 | 後期 $0.5 < D_I \leq 0.6$ の出現率 | 後期における各 $D_I$ が 0.5〜0.6 の間にある確率 |
| 14 | 後期 $0.6 < D_I \leq 0.7$ の出現率 | 後期における各 $D_I$ が 0.6〜0.7 の間にある確率 |
| 15 | 前期各 $D_I$ 差 | 前期における各 $D_I$ の最大差が 1 以上あるかどうか |
| 16 | 後期各 $D_I$ 差 | 後期における各 $D_I$ の最大差が 1 以上あるかどうか |

出欠補正データからは 16 個の変数を用いて予測を行う．なお従来研究では打刻回数による変数, 欠席回数による変数, 各曜日における打刻回数の平均及び分散の変数を用いて予測を

行っていたが, 欠席回数による変数と各曜日における打刻回数の平均及び分散の変数は予測にあまり関係しなかったため, 本研究では打刻回数による変数と新たな変数として $D_I$ による変数を用いて予測を行っている.

### 成績データによる変数

本研究では成績の指標として GPA を用いている．GPA は各成績評価である秀・優・良・可・不可・失格にそれぞれ 4 点・3 点・2 点・1 点・0 点・0 点の得点を割り振り, 講義毎に決められている単位数を用いて下記の式 (3.2) で求めることができる.

$$GPA = \frac{4 * \text{秀の取得単位数} + 3 * \text{優の取得単位数} + 2 * \text{良の取得単位数} + 1 * \text{可の取得単位数}}{\text{総履修登録単位数（不可・失格の単位数も含む）}} \tag{3.2}$$

成績データより変数として前期, 後期における分野毎の GPA を設定している．その時の変数を表 3.2 に列挙しておく.

表 3.2: 成績データより拡張した説明変数

| 番号 | 変数名 | 意味 |
|---|---|---|
| 17 | 1 年前期 | 1 年次の前期に受講した講義の GPA 値 |
| 18 | 1 年前期英語 | 1 年次の前期に受講した英語教科の GPA 値 |
| 19 | 1 年前期人文 | 1 年次の前期に受講した「人間文化」に分類される講義の GPA 値 |
| 20 | 1 年前期体育 | 1 年次の前期に受講した体育教科の GPA 値 |
| 21 | 1 年前期専門 | 1 年次の前期に受講した専門科目の講義の GPA 値 |
| 22 | 1 年前期理科 | 1 年次の前期に受講した理科系の講義の GPA 値 |
| 23 | 1 年前期数学 | 1 年次の前期に受講した数学系の講義の GPA 値 |
| 24 | 1 年後期 | 1 年次の後期に受講した講義の GPA 値 |
| 25 | 1 年後期英語 | 1 年次の後期に受講した英語教科の GPA 値 |
| 26 | 1 年後期人文 | 1 年次の後期に受講した「人間文化」に分類される講義の GPA 値 |
| 27 | 1 年後期体育 | 1 年次の後期に受講した体育教科の GPA 値 |
| 28 | 1 年後期専門 | 1 年次の後期に受講した専門科目の講義の GPA 値 |
| 29 | 1 年後期理科 | 1 年次の後期に受講した理科系の講義の GPA 値 |
| 30 | 1 年後期数学 | 1 年次の後期に受講した数学系の講義の GPA 値 |

成績データからは 14 個の変数を用いて予測を行う．成績データに関しては従来研究で用いていた変数と同様である．以上より予測に用いる説明変数として出欠補正データから 16 個, 成績データから 14 個の計 30 個の変数を用意して予測を行う．

### 目的変数

目的変数には 2 年終了時点における総合 GPA を離散化したものを設定している．簡易な離散化の方法として等間隔による離散化手法と等頻度による離散化手法の 2 つが挙げられるが, 本研究では成績レベル毎に離散化したいことから等間隔による離散化を行った．また名古屋工業大学では成績は「秀, 優, 良, 可, 不可」の 5 段階で表されているため, 本研究では GPA の取りうる値 0〜4 の範囲を等間隔に 5 等分したものを成績レベル (S,A,B,C,D) とした．このとき S は成績が優秀であることを表しており A, B と進むにつれて成績が悪くなっていることを表している. 表 3.3 に目的変数を 5 段階に分けた詳細について列挙する．

表 3.3: 目的変数として設定した成績レベルについて

| 番号 | 離散化 | GPA 数値範囲 |
|---|---|---|
| 31 | S | 3.2 以上 |
| | A | 2.4〜3.2 |
| | B | 1.6〜2.4 |
| | C | 0.8〜1.6 |
| | D | 0.8 以下 |

## 3.3　ベイジアンネットワークによる予測結果

本研究では成績レベルを予測した際に打刻による特徴を見つけ出せないかということを検証するために成績予測手法としてベイジアンネットワークを採用している. それによりモデルを構築しその構築過程において打刻による変数が確認できれば, 成績レベル予測に関わる特徴として捉えることができる．しかし, 予測が正しくなければモデルから打刻に関するデータの特徴を見付けることができたとしてもモデル自体が正しくないために見付けた特徴が有用であるとは言い難い. そこでまず成績を予測するにあたり予測の精度がどの程度のものか検証してからモデルより特徴分析を行っている．本研究ではデータマイニングの手法としてベイジアンネットワークを利用するためにフリーソフトの Weka [18] を用いている．

従来研究では 2 年前期終了時点から予測していたが, 従来用いていた変数から 1 年次終了時点のデータによる変数部分のみを用いてベイジアンネットワークによる予測を行ったとき, 一番予測的中精度が良かった結果を表 3.4 に示す．ただしモデルの検証方法は leave one out 法を用いている．

表 3.4: ベイジアンネットワークによる予測結果

| 実際＼予測 | S | A | B | C | D |
|---|---|---|---|---|---|
| S | 7 | 7 | 0 | 0 | 0 |
| A | 2 | 106 | 14 | 0 | 0 |
| B | 0 | 24 | 109 | 11 | 2 |
| C | 0 | 0 | 5 | 36 | 1 |
| D | 0 | 0 | 2 | 4 | 8 |

表 3.4 より予測的中精度は約 78.7 %であった．表 3.4 は行が離散化で分けられた成績レベルを持つ学生数, 列が予測された成績レベルを持つ学生数を表している．以後表 3.4 のような形式の表は同様に書かれているものとする．

次に本研究部分である表 3.1,3.2,3.3 で述べた変数を用いてベイジアンネットワークによる予測を行った．ここで最適なベイジアンネットワークモデルを作成するために予測を行う前にあらかじめ情報利得による変数選択を行った．その結果表 3.5 の変数によるモデルが一番良い結果を示すことが分かった．モデルの検証方法として leave one out 法を用いている．

表 3.5: 情報利得により取捨選択された変数

| 番号 | 変数名 |
|---|---|
| 6 | 1 年 11 月打刻回数 |
| 7 | 1 年 12 月打刻回数 |
| 8 | 1 年 1 月打刻回数 |
| 16 | 後期各 $D_I$ 差 |
| 17 | 1 年前期 |
| 18 | 1 年前期英語 |
| 19 | 1 年前期人文 |
| 21 | 1 年前期専門 |
| 22 | 1 年前期理科 |
| 23 | 1 年前期数学 |
| 24 | 1 年後期 |
| 25 | 1 年後期英語 |
| 26 | 1 年後期人文 |
| 28 | 1 年後期専門 |
| 29 | 1 年後期理科 |
| 30 | 1 年後期数学 |

表 3.5 の変数を用いたとき, 表 3.6 の結果が得られた．この表より予測的中精度は約 80.47 %であることがわかる．

表 3.6: ベイジアンネットワークによる予測結果

| 実際＼予測 | S | A | B | C | D |
|---|---|---|---|---|---|
| S | 7 | 7 | 0 | 0 | 0 |
| A | 2 | 107 | 13 | 0 | 0 |
| B | 0 | 18 | 116 | 11 | 1 |
| C | 0 | 0 | 8 | 32 | 2 |
| D | 0 | 0 | 1 | 3 | 10 |

従来の表 3.4 の結果と本研究による表 3.6 の結果を比べたとき, 成績レベル C の予測に関しては従来結果のほうが予測的中精度が高いが, 成績レベル A,B,D に関しては本研究のほうが結果が良くなっていることが確認でき, 全体的な予測的中精度に関しても向上していることがわかる．よって新たに追加した $D_I$ による変数は目的変数の予測に有効であることがわかる．

## 3.4　ベイジアンネットワークモデルによる特徴分析

表 3.6 のときベイジアンネットワークモデルは図 3.1 のように構築されており目的変数が 11 月打刻回数と 12 月打刻回数の変数から有向グラフが引かれていることが確認できる．よって成績レベルは 11 月打刻回数と 12 月打刻回数により確率的に表現できることが分かった．このとき確率結果がどうなっているかを表 3.7 に示す．この表は各変数の状態において各成績レベルがどれだけ発生しやすいかを表したものとなっており, 以後このような形式の表は同様に考えるものとする．

表 3.7: A,B 年度データのモデルによる成績レベル表現

| 11 月打刻回数 | 12 月打刻回数 | S | A | B | C | D |
|---|---|---|---|---|---|---|
| $\leq 59.5$ | $\leq 44.5$ | 1.1 % | 11.8 % | 11.8 % | 44.1 % | 31.2 % |
| $\leq 59.5$ | $44.5 <$ | 4.3 % | 4.3 % | 56.5 % | 30.4 % | 4.3 % |
| $59.5 <$ | $\leq 44.5$ | 3.4 % | 17.2 % | 51.7 % | 24.1 % | 3.4 % |
| $59.5 <$ | $44.5 <$ | 5.3 % | 41.9 % | 46.6 % | 6 % | 0.2 % |

表 3.7 より 11 月打刻回数が 59.5 より大きく 12 月打刻回数が 44.5 より大きいと成績レベル A,B と予測される確率が高いことから成績が平均以上になりやすいという傾向がみられる．また 11 月打刻回数が 59.5 以下で 12 月打刻回数が 44.5 以下だと成績レベル C,D と予測され

る確率が高いことから成績が悪くなりやすい傾向がみられる．ただこのとき成績レベル A,B に関しても約 1 割程度の確率で予測されてしまうことからこのモデルを他のデータで試した時に同程度の予測的中精度が得られるか確かめる必要がある．

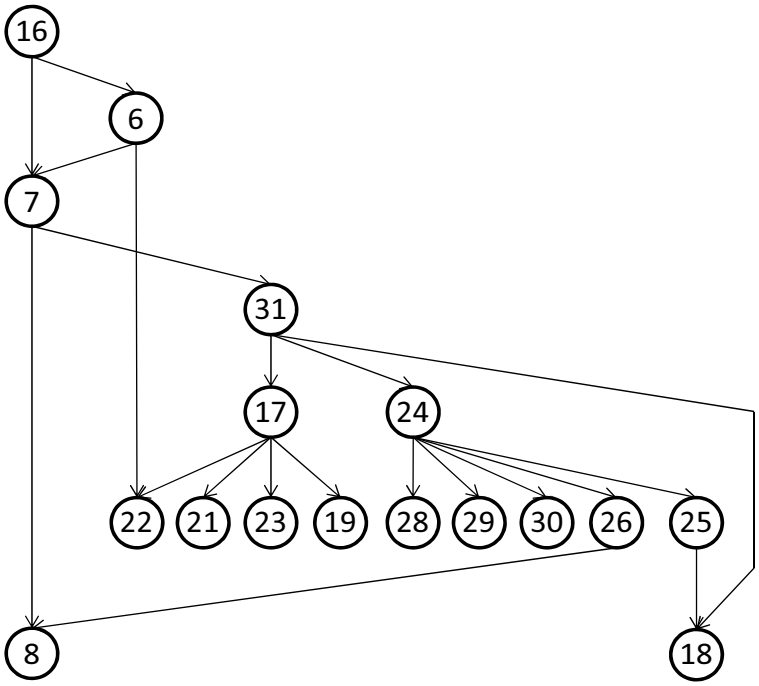


図 3.1: A,B 年度データによるベイジアンネットワークモデル

# 第4章　変数の改善による予測の検証

第3章ではA,B年度に記録された2学年度分のデータを用いて成績レベルを予測し従来研究よりも良い結果を確認できたが, まだ他データで予測的中精度が同程度得られるか不安な点がみられた．そこで本章では新たにC年度に名古屋工業大学に入学してきた学生のデータを用いることで第3章の結果の検証を行うことにした．C年度データには学生166人分のデータが記録されており, データの拡張による変数の作成は第3章で行った処理と同様に行うものとする．

## 4.1　第3章で述べた変数を使用したleave one out法による検証

A,B,C年度データを合わせて3学年度分のデータとして扱い, 第3章で述べた変数による予測を行った．ベイジアンネットワーク手法で予測した際, 最も高い予測的中精度を表す結果を表4.1に示す．ただしモデルの検証はleave one out法で行う．

表 4.1: 第3章で用いた変数による予測結果

| 実際＼予測 | S | A | B | C | D |
|---|---|---|---|---|---|
| S | 10 | 11 | 0 | 0 | 0 |
| A | 1 | 156 | 23 | 0 | 0 |
| B | 0 | 30 | 170 | 15 | 2 |
| C | 0 | 0 | 8 | 56 | 2 |
| D | 0 | 0 | 0 | 7 | 13 |

表4.1より予測的中精度は約80.36％となった．第3章よりA,B年度のデータで予測した時の結果が約80.47％であることから,C年度データを追加すると予測的中精度はほんの少し下がってしまうがほぼ同等の精度を保っていることが分かった．

## 4.2　第3章で述べた変数を使用したホールドアウト法による検証

次に第3章で構築したベイジアンネットワークモデルについてホールドアウト法による検証を行う．新たに加えたC年度データを評価データとして用いた．その時の結果を表4.2に示す.

表 4.2: ホールドアウト法による第 3 章モデルの予測結果

| 実際＼予測 | S | A | B | C | D |
|---|---|---|---|---|---|
| S | 3 | 4 | 0 | 0 | 0 |
| A | 0 | 52 | 6 | 0 | 0 |
| B | 0 | 13 | 53 | 5 | 0 |
| C | 0 | 0 | 6 | 17 | 1 |
| D | 0 | 0 | 0 | 3 | 3 |

表 4.2 より予測的中精度は約 77.11 %となることがわかった．また成績レベル A の予測的中精度は高いが, 成績レベル S,D の予測的中精度が約 5 割程度しかないことが確認できる．

## 4.3　3 学年度分のデータによる予測の検証

3 章で述べたモデルでは成績レベル予測を行うことで打刻による特徴は発見できるが, 成績レベル S,D を予測するにはこのモデルはあまり有効でないと思われる．これは年度毎にカリキュラム変更であったり授業時間の変更などで同じ授業を受けていたとしてもデータとしてはずれが生じる可能性があることなどが原因として考えられる．理想としてはどの年度においても共通の特徴を見つけ出したいが, 実際にのような結果になるかはわからない．そこで本章では, データが 3 学年度分であることから下記のパターンで予測を行い, それぞれのパターンにおける特徴を見ることにした．このとき予測に用いるデータは 1 年次終了時点で得られるデータを使用し, 予測対象は 2 年終了時点における総合の成績レベルであることを改めて確認しておく．

- A,B,C 年度のデータを用いて leave one out 法で予測
- A,B 年度のデータを学習データ,C 年度のデータを評価データとしたホールドアウト法で予測
- B,C 年度のデータを学習データ,A 年度のデータを評価データとしたホールドアウト法で予測
- A,C 年度のデータを学習データ,B 年度のデータを評価データとしたホールドアウト法で予測

### 4.3.1　変数の変更・追加

更なる予測的中精度向上を目指し, 先に述べた各パターンの予測を行う前に第 3 章で用いていた変数の変更や追加を行うことにした．先に変更・追加を行い最終的に予測で用いた変数について表 4.3 に載せる．

表 4.3: 予測に用いる変数

| 番号 | 変数名 | 意味 |
|---|---|---|
| 1 | 1 年 4 月打刻回数 | 1 年 4 月に打刻した回数 |
| 2 | 1 年 5 月打刻回数 | 1 年 5 月に打刻した回数 |
| 3 | 1 年 6 月打刻回数 | 1 年 6 月に打刻した回数 |
| 4 | 1 年 7 月打刻回数 | 1 年 7 月に打刻した回数 |
| 5 | 1 年 10 月打刻回数 | 1 年 10 月に打刻した回数 |
| 6 | 1 年 11 月打刻回数 | 1 年 11 月に打刻した回数 |
| 7 | 1 年 12 月打刻回数 | 1 年 12 月に打刻した回数 |
| 8 | 1 年 1 月打刻回数 | 1 年 1 月に打刻した回数 |
| 9 | 前期 $D_I \leq 0.5$ の出現率 | 前期における各 $D_I$ が 0.5 を下回っている確率 |
| 10 | 前期 $0.5 < D_I \leq 0.8$ の出現率 | 前期における各 $D_I$ が 0.5〜0.8 の間にある確率 |
| 11 | 前期 $0.8 < D_I \leq 1.2$ の出現率 | 前期における各 $D_I$ が 0.8〜1.2 の間にある確率 |
| 12 | 後期 $D_I \leq 0.5$ の出現率 | 後期における各 $D_I$ が 0.5 を下回っている確率 |
| 13 | 後期 $0.5 < D_I \leq 0.8$ の出現率 | 後期における各 $D_I$ が 0.5〜0.8 の間にある確率 |
| 14 | 後期 $0.8 < D_I \leq 1.2$ の出現率 | 後期における各 $D_I$ が 0.8〜1.2 の間にある確率 |
| 15 | 前期各 $D_I$ 差 | 前期における各 $D_I$ の最大差が 1 以上あるかどうか |
| 16 | 後期各 $D_I$ 差 | 後期における各 $D_I$ の最大差が 1 以上あるかどうか |
| 17 | 教養 1（平均） | 教科（教養 1）における授業開始時間 と入室時間との秒差の平均 |
| 18 | 教養 2（平均） | 教科（教養 2）における授業開始時間と入室時間との秒差の平均 |
| 19 | 教養 3（平均） | 教科（教養 3）における授業開始時間と入室時間との秒差の平均 |
| 20 | 教養 4（平均） | 教科（教養 4）における授業開始時間と入室時間との秒差の平均 |
| 21 | 教養 5（平均） | 教科（教養 5）における授業開始時間と入室時間との秒差の平均 |
| 22 | 教養 6（平均） | 教科（教養 6）における授業開始時間と入室時間との秒差の平均 |
| 23 | 教養 7（平均） | 教科（教養 7）における授業開始時間と入室時間との秒差の平均 |

| | | |
|---|---|---|
| 24 | 教養 8（平均） | 教科（教養 8）における授業開始時間と入室時間との秒差の平均 |
| 25 | 教養 9（平均） | 教科（教養 9）における授業開始時間と入室時間との秒差の平均 |
| 26 | 教養 10（平均） | 教科（教養 10）における授業開始時間と入室時間との秒差の平均 |
| 27 | 教養 11（平均） | 教科（教養 11）における授業開始時間と入室時間との秒差の平均 |
| 28 | 専門 1（平均） | 教科（専門 1）における授業開始時間と入室時間との秒差の平均 |
| 29 | 専門 2（平均） | 教科（専門 2）における授業開始時間と入室時間との秒差の平均 |
| 30 | 専門 3（平均） | 教科（専門 3）における授業開始時間と入室時間との秒差の平均 |
| 31 | 専門 4（平均） | 教科（専門 4）における授業開始時間と入室時間との秒差の平均 |
| 32 | 専門 5（平均） | 教科（専門 5）における授業開始時間と入室時間との秒差の平均 |
| 33 | 専門 6（平均） | 教科（専門 6）における授業開始時間と入室時間との秒差の平均 |
| 34 | 1 年前期 | 1 年次の前期に受講した講義の GPA 値 |
| 35 | 1 年前期英語 | 1 年次の前期に受講した英語教科の GPA 値 |
| 36 | 1 年前期人文 | 1 年次の前期に受講した「人間文化」に分類される講義の GPA 値 |
| 37 | 1 年前期体育 | 1 年次の前期に受講した体育教科の GPA 値 |
| 38 | 1 年前期専門 | 1 年次の前期に受講した専門科目の講義の GPA 値 |
| 39 | 1 年前期理科 | 1 年次の前期に受講した理科系の講義の GPA 値 |
| 40 | 1 年前期数学 | 1 年次の前期に受講した数学系の講義の GPA 値 |
| 41 | 1 年後期 | 1 年次の後期に受講した講義の GPA 値 |
| 42 | 1 年後期英語 | 1 年次の後期に受講した英語教科の GPA 値 |
| 43 | 1 年後期人文 | 1 年次の後期に受講した「人間文化」に分類される講義の GPA 値 |
| 44 | 1 年後期体育 | 1 年次の後期に受講した体育教科の GPA 値 |
| 45 | 1 年後期専門 | 1 年次の後期に受講した専門科目の講義の GPA 値 |

| | | |
|---|---|---|
| 46 | 1 年後期理科 | 1 年次の後期に受講した理科系の講義の GPA 値 |
| 47 | 1 年後期数学 | 1 年次の後期に受講した数学系の講義の GPA 値 |

変数番号 1 から 33 までが出欠補正データより拡張した変数, 変数番号 34 から 47 までのが成績データより拡張した変数を表している．変更部分に関して変数番号 10,11,13,14 の値の範囲を変更した．3 章では $D_I$ の値が 0.5 付近に着目して変数を作っていたが,$D_I$ の値が 1 付近の場合, 全体と比べて平均的な打刻をしていることから, 値が 0.8〜1.2 の間にある数を変数として設定し, 先に出てきた 2 つの範囲の間である 0.5〜0.8 の値をもつ $D_I$ の数についても変数として設定した．また追加部分に関して変数番号 17 から 33 を追加した．これらの変数は授業開始時間と入室時間の差（授業開始何分前に打刻したのかまたは何分後に打刻したのかなど）に関して授業毎に求め秒数表示したものであり, 授業の出席時間帯が成績レベル予測に関係するのか確認するために変数として設定した．このとき変数として扱う授業は単位取得必須の授業かつ出欠補正データが記録されている部分に限定した．

以上より表 4.3 の変数を用いて各パターンにおける予測を行う．目的変数である成績レベルは 3 章で述べたものと同様の変数を用いるが, 新たに変数を追加したため目的変数の変数番号は 31 番から 48 番に変更していることを記述しておく．

### 4.3.2　ABC 年度データ

#### A,B,C 年度データによる予測結果

3 学年度分, 計 504 名のデータより表 4.3 の変数を用いて予測を行う．まず情報利得による変数選択を行ったときの結果を表 4.4 に示す．ただし表に載っている全ての変数がベイジアンネットワークで予測する際に用いられるわけではないことを述べておく．

表 4.4: 情報利得により取捨選択された変数 (A,B,C 年度データ)

| 番号 | 変数名 |
|---|---|
| 3 | 1 年 6 月打刻回数 |
| 4 | 1 年 7 月打刻回数 |
| 5 | 1 年 10 月打刻回数 |
| 6 | 1 年 11 月打刻回数 |
| 7 | 1 年 12 月打刻回数 |
| 8 | 1 年 1 月打刻回数 |
| 9 | 前期 $D_I \leq 0.5$ の出現率 |
| 10 | 前期 $0.5 < D_I \leq 0.8$ の出現率 |

| | |
|---|---|
| 11 | 前期 $0.8 < D_I \leq 1.2$ の出現率 |
| 12 | 後期 $D_I \leq 0.5$ の出現率 |
| 13 | 後期 $0.5 < D_I \leq 0.8$ の出現率 |
| 14 | 後期 $0.8 < D_I \leq 1.2$ の出現率 |
| 15 | 前期各 $D_I$ 差 |
| 16 | 後期各 $D_I$ 差 |
| 26 | 教養 10（平均） |
| 31 | 専門 4（平均） |
| 32 | 専門 5（平均） |
| 34 | 1 年前期 |
| 35 | 1 年前期英語 |
| 36 | 1 年前期人文 |
| 37 | 1 年前期体育 |
| 38 | 1 年前期専門 |
| 39 | 1 年前期理科 |
| 40 | 1 年前期数学 |
| 41 | 1 年後期 |
| 42 | 1 年後期英語 |
| 43 | 1 年後期人文 |
| 44 | 1 年後期体育 |
| 45 | 1 年後期専門 |
| 46 | 1 年後期理科 |
| 47 | 1 年後期数学 |

3 学年度分のデータで予測を行うとき, 表 4.4 に書かれている変数が目的変数を予測するのに有効な変数であることが分かった．ベイジアンネットワークを用いて予測した際,leave one out 法による検証の結果を表 4.5 に示す.

表 4.5: ベイジアンネットワークによる予測結果 (A,B,C 年度データ)

| 実際＼予測 | S | A | B | C | D |
|---|---|---|---|---|---|
| S | 10 | 11 | 0 | 0 | 0 |
| A | 1 | 156 | 23 | 1 | 0 |
| B | 0 | 28 | 172 | 16 | 1 |
| C | 0 | 0 | 7 | 55 | 4 |
| D | 0 | 0 | 0 | 6 | 14 |

表 4.5 よりベイジアンネットワークによる予測的中精度は約 80.75 %になった．3 学年度分のデータにおける予測は表 4.4 による変数を用いるときが一番良い予測的中精度を示しており,C 年度データを増やしても予測的中精度は約 8 割を保てていることが確認できた．だが表 4.5 より成績レベル A,B,C の学生は約 8 割程度の確率で予測可能であるが成績レベル S の学生は約半分しか予測できていないことが分かった．また表 4.1 の結果と比べてわずかながら予測的中精度が向上していることがわかる．よって変数の変更・追加が前の変数よりもうまく機能していることが確認できる．

### A,B,C 年度データによる特徴分析

3 学年度分のデータで予測を行ったとき, 構築されたベイジアンネットワークモデルを図 4.1 に示す.

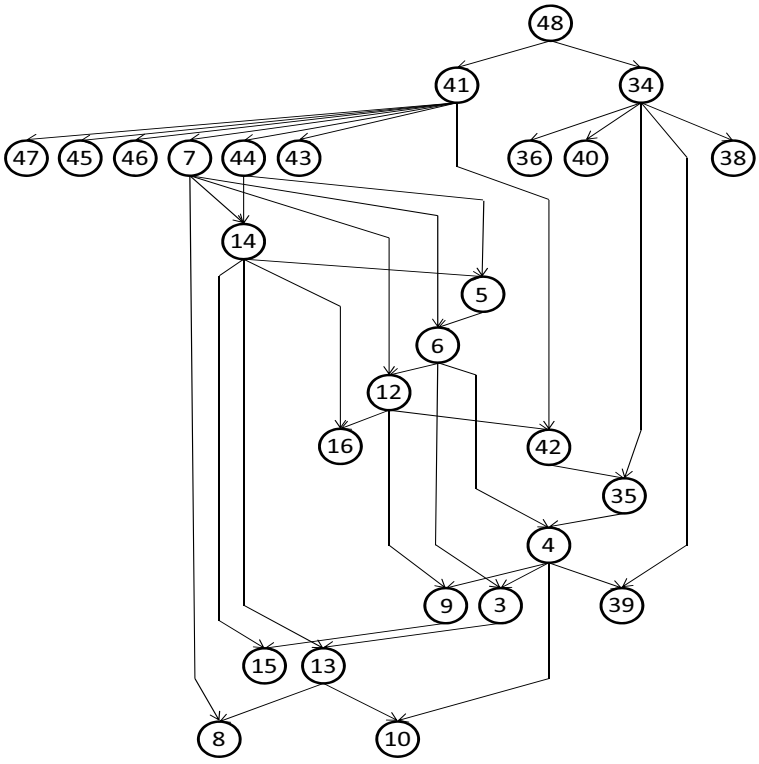


図 4.1: A,B,C 年度データによるベイジアンネットワークモデル

図 4.1 より目的変数から 1 年前期 GPA,1 年後期 GPA の変数に向かって有向グラフが引かれていることがわかる．そのため図 4.1 からは目的変数を表現できるような説明変数による特徴を見つけることはできなかった.

### 4.3.3　A,B 年度データ

#### A,B 年度データによる予測結果

A 年度 171 人,B 年度 167 人の計 338 人のデータを用いて予測を行う．情報利得により選択された変数は表 4.6 のようになった．

表 4.6: 情報利得により取捨選択された変数 (A,B 年度データ)

| 番号 | 変数名 |
|---|---|
| 2 | 1 年 5 月打刻回数 |
| 3 | 1 年 6 月打刻回数 |
| 4 | 1 年 7 月打刻回数 |
| 5 | 1 年 10 月打刻回数 |
| 6 | 1 年 11 月打刻回数 |
| 7 | 1 年 12 月打刻回数 |
| 8 | 1 年 1 月打刻回数 |
| 9 | 前期 $D_I \leq 0.5$ の出現率 |
| 10 | 前期 $0.5 < D_I \leq 0.8$ の出現率 |
| 11 | 前期 $0.8 < D_I \leq 1.2$ の出現率 |
| 12 | 後期 $D_I \leq 0.5$ の出現率 |
| 13 | 後期 $0.5 < D_I \leq 0.8$ の出現率 |
| 14 | 後期 $0.8 < D_I \leq 1.2$ の出現率 |
| 15 | 前期各 $D_I$ 差 |
| 16 | 後期各 $D_I$ 差 |
| 34 | 1 年前期 |
| 35 | 1 年前期英語 |
| 36 | 1 年前期人文 |
| 37 | 1 年前期体育 |
| 38 | 1 年前期専門 |
| 39 | 1 年前期理科 |
| 40 | 1 年前期数学 |
| 41 | 1 年後期 |
| 42 | 1 年後期英語 |
| 43 | 1 年後期人文 |
| 44 | 1 年後期体育 |
| 45 | 1 年後期専門 |
| 46 | 1 年後期理科 |
| 47 | 1 年後期数学 |

A,B 年度のデータで予測を行うとき, 表 4.6 に書かれている変数が目的変数を予測するのに有効な変数であることが分かった．このときまず A,B 年度のデータで予測した際,leave one out 法による検証の結果を表 4.7 に示す．

表 4.7: ベイジアンネットワークによる予測結果 (leave one out 法)

| 実際＼予測 | S | A | B | C | D |
|---|---|---|---|---|---|
| S | 7 | 7 | 0 | 0 | 0 |
| A | 2 | 106 | 14 | 1 | 0 |
| B | 0 | 19 | 114 | 11 | 2 |
| C | 0 | 0 | 9 | 30 | 3 |
| D | 0 | 0 | 0 | 2 | 12 |

表 4.7 よりベイジアンネットワークによる予測的中精度は約 79.59 %となった．また成績レベル D の学生の予測的中精度が非常に高いことが確認できる．また A,B 年度を学習データとして C 年度を評価データとして用いた場合どうなるか検証した．このときの予測結果を表 4.8 に示す．

表 4.8: ベイジアンネットワークによる予測結果 (ホールドアウト法)

| 実際＼予測 | S | A | B | C | D |
|---|---|---|---|---|---|
| S | 3 | 4 | 0 | 0 | 0 |
| A | 0 | 53 | 5 | 0 | 0 |
| B | 0 | 13 | 51 | 7 | 0 |
| C | 0 | 0 | 3 | 19 | 2 |
| D | 0 | 0 | 0 | 3 | 3 |

表 4.8 よりベイジアンネットワークによる予測的中精度は約 77.71 %となっている．leave one out 法で検証したときは成績レベル D の予測的中精度が非常に高かったがホールドアウト法では予測的中精度が 5 割程度に落ちてしまうことが確認できた．

次に leave one out 法において, 情報利得によって選択された変数から有効性の高い変数を順に選択していき予測的中精度が最も良くなったときの変数を表 4.9 に, そのときの結果を表 4.10 に示す．

表 4.9: 予測的中精度が最良の時の変数 (A,B 年度データ)

| 番号 | 変数名 |
|---|---|
| 5 | 1 年 10 月打刻回数 |

| | |
|---|---|
| 6 | 1 年 11 月打刻回数 |
| 7 | 1 年 12 月打刻回数 |
| 8 | 1 年 1 月打刻回数 |
| 14 | 後期 $0.8 < D_I \leq 1.2$ の出現率 |
| 16 | 後期各 $D_I$ 差 |
| 34 | 1 年前期 |
| 35 | 1 年前期英語 |
| 36 | 1 年前期人文 |
| 38 | 1 年前期専門 |
| 39 | 1 年前期理科 |
| 40 | 1 年前期数学 |
| 41 | 1 年後期 |
| 42 | 1 年後期英語 |
| 43 | 1 年後期人文 |
| 45 | 1 年後期専門 |
| 46 | 1 年後期理科 |
| 47 | 1 年後期数学 |

表 4.10: ベイジアンネットワークによる予測結果 (leave one out 法, 情報利得)

| 実際＼予測 | S | A | B | C | D |
|---|---|---|---|---|---|
| S | 7 | 7 | 0 | 0 | 0 |
| A | 2 | 106 | 14 | 0 | 0 |
| B | 0 | 17 | 118 | 10 | 1 |
| C | 0 | 0 | 9 | 31 | 2 |
| D | 0 | 0 | 0 | 3 | 11 |

表 4.10 よりベイジアンネットワークによる予測的中精度は約 80.77 %となっていることがわかり, 表 4.7 に比べて予測的中精度が向上していることがわかる．またこの結果からも第 3 章で求めた予測的中精度より高いことが確認できるため, 本章で変更・追加した変数が機能していることがわかる．

このモデルにおいて C 年度を評価データとしたときの予測結果を表 4.11 に示す

表 4.11: ベイジアンネットワークによる予測結果 (ホールドアウト法, 情報利得)

| 実際＼予測 | S | A | B | C | D |
|---|---|---|---|---|---|
| S | 3 | 4 | 0 | 0 | 0 |
| A | 0 | 52 | 6 | 0 | 0 |
| B | 0 | 10 | 55 | 6 | 0 |
| C | 0 | 0 | 4 | 19 | 1 |
| D | 0 | 0 | 0 | 3 | 3 |

表 4.11 よりベイジアンネットワークによる予測的中精度は約 79.52 %となった．表 4.8 と比べて説明変数を減らして予測を行っても予測的中精度が向上しているため A,B 年度のデータで予測した場合, この時のベイジアンネットワークモデルが最適であることが分かった．しかし変数を選んでモデルを作成した場合においても, 成績レベル D に関してはやはり約半分程度しか予測できないことがわかり成績レベル S の予測的中精度もあまり高くない．

### A,B 年度データによる特徴分析

A,B 年度（2 学年度分）のデータで予測を行ったとき, 情報利得で選択された変数から有効性の高い変数を順に選択したときの結果が最も良かったので, そのときに構築されたベイジアンネットワークモデルを図 4.2 に示す．図 4.2 より目的変数は 11 月打刻回数と後期 $0.8 < D_I \leq 1.2$ の変数より有向グラフが引かれていることがわかる．よってこのモデルによる成績レベルの特徴は 11 月打刻回数と後期 $0.8 < D_I \leq 1.2$ の変数によって確率的に表現できる．

表 4.12: A,B 年度データのモデルによる成績レベル表現

| 11 月打刻回数 | 後期 $0.8 < D_I \leq 1.2$ | S | A | B | C | D |
|---|---|---|---|---|---|---|
| $\leq 59.5$ | $\leq 0.653846$ | 1.1 % | 3.2 % | 24.2 % | 43.2 % | 28.4 % |
| $\leq 59.5$ | $0.653846 <$ | 4.8 % | 42.9 % | 4.8 % | 33.3 % | 14.3 % |
| $59.5 <$ | $\leq 0.653846$ | 2.4 % | 17.1 % | 51.2 % | 26.8 % | 2.4 % |
| $59.5 <$ | $0.653846 <$ | 5.4 % | 42.5 % | 46.6 % | 5.4 % | 0.2 % |

表 4.12 より 11 月打刻回数が 59.5 回より低いと成績が悪くなりやすい傾向がみられ, また後期における各 $D_I$ が 0.8〜1.2 の間にある確率が約 0.65 より低いと成績結果が悪くなりやすい傾向を持つことが分かった．

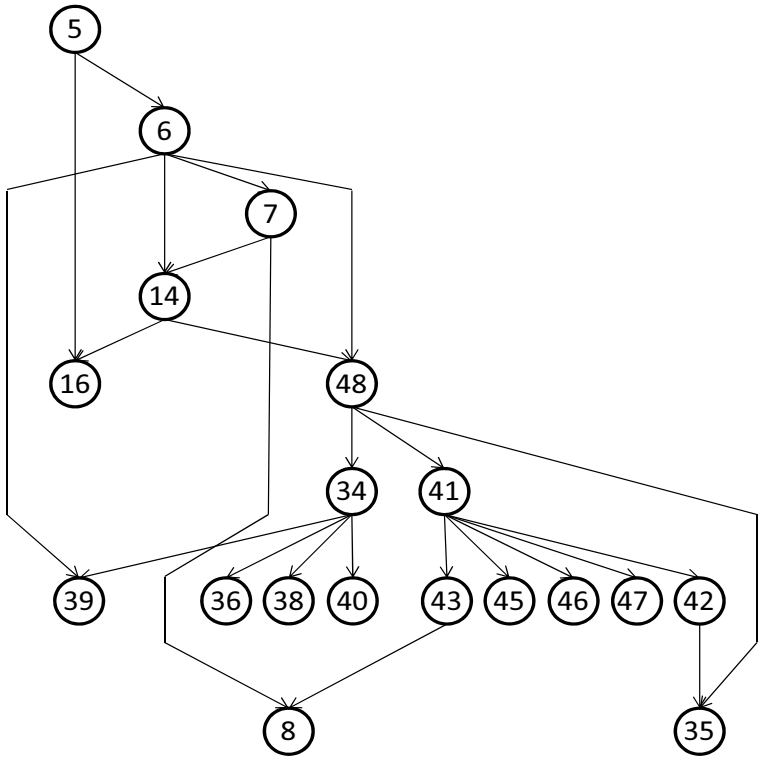


図 4.2: A,B 年度データによるベイジアンネットワークモデル

### 4.3.4　B,C 年度データ

#### B,C 年度データによる予測結果

B 年度 167 人,C 年度 166 人の計 333 人のデータを用いて予測を行う．情報利得により選択された変数は表 4.13 のようになった．

表 4.13: 情報利得により取捨選択された変数 (B,C 年度データ)

| 番号 | 変数名 |
|---|---|
| 3 | 1 年 6 月打刻回数 |
| 4 | 1 年 7 月打刻回数 |
| 5 | 1 年 10 月打刻回数 |
| 6 | 1 年 11 月打刻回数 |
| 7 | 1 年 12 月打刻回数 |
| 8 | 1 年 1 月打刻回数 |
| 9 | 前期 $D_I \leq 0.5$ の出現率 |
| 10 | 前期 $0.5 < D_I \leq 0.8$ の出現率 |
| 11 | 前期 $0.8 < D_I \leq 1.2$ の出現率 |

| | |
|---|---|
| 12 | 後期 $D_I \leq 0.5$ の出現率 |
| 13 | 後期 $0.5 < D_I \leq 0.8$ の出現率 |
| 14 | 後期 $0.8 < D_I \leq 1.2$ の出現率 |
| 15 | 前期各 $D_I$ 差 |
| 16 | 後期各 $D_I$ 差 |
| 32 | 専門 5（平均） |
| 34 | 1 年前期 |
| 35 | 1 年前期英語 |
| 36 | 1 年前期人文 |
| 38 | 1 年前期専門 |
| 39 | 1 年前期理科 |
| 40 | 1 年前期数学 |
| 41 | 1 年後期 |
| 42 | 1 年後期英語 |
| 43 | 1 年後期人文 |
| 44 | 1 年後期体育 |
| 45 | 1 年後期専門 |
| 46 | 1 年後期理科 |
| 47 | 1 年後期数学 |

B,C 年度のデータで予測を行うとき, 表 4.13 に書かれている変数が目的変数を予測するのに有効な変数であることが分かった．このときまず B,C 年度のデータで予測した際,leave one out 法による検証の結果を表 4.14 に示す.

表 4.14: ベイジアンネットワークによる予測結果 (leave one out 法)

| 実際＼予測 | S | A | B | C | D |
|---|---|---|---|---|---|
| S | 13 | 3 | 0 | 0 | 0 |
| A | 5 | 100 | 14 | 0 | 0 |
| B | 0 | 24 | 104 | 14 | 1 |
| C | 0 | 0 | 7 | 28 | 5 |
| D | 0 | 0 | 0 | 3 | 12 |

表 4.14 よりベイジアンネットワークによる予測的中精度は約 77.18 %となっており, 成績レベル S,D の予測的中精度が非常に高いことが確認できる．また B,C 年度を学習データとして A 年度を評価データとして用いた場合どうなるか検証した．このときの予測結果を表 4.15 に示す.

表 4.15: ベイジアンネットワークによる予測結果 (ホールドアウト法)

| 実際＼予測 | S | A | B | C | D |
|---|---|---|---|---|---|
| S | 4 | 1 | 0 | 0 | 0 |
| A | 4 | 48 | 9 | 0 | 0 |
| B | 0 | 6 | 63 | 4 | 1 |
| C | 0 | 0 | 5 | 19 | 2 |
| D | 0 | 0 | 0 | 0 | 5 |

表 4.15 よりベイジアンネットワークによる予測的中精度は約 81.29 %となった．表よりホールドアウト法による検証においても成績レベル S,D の学生の予測的中精度が高く,B,C 年度データによるモデルは成績優秀者と成績不振者を見つけるのに適していると考えられる．

次に leave one out 法において, 情報利得によって選択された変数から有効性の高い変数を順に選択していき予測的中精度が最も良くなったときの変数を表 4.16 に, 結果を表 4.17 に示す．

表 4.16: 予測的中精度が最良の時の変数 (B,C 年度データ)

| 番号 | 変数名 |
|---|---|
| 8 | 1 年 1 月打刻回数 |
| 32 | 専門 5（平均） |
| 34 | 1 年前期 |
| 41 | 1 年後期 |
| 45 | 1 年後期専門 |

表 4.17: ベイジアンネットワークによる予測結果 (leave one out 法, 情報利得)

| 実際＼予測 | S | A | B | C | D |
|---|---|---|---|---|---|
| S | 14 | 2 | 0 | 0 | 0 |
| A | 3 | 103 | 13 | 0 | 0 |
| B | 0 | 24 | 104 | 15 | 0 |
| C | 0 | 0 | 4 | 33 | 3 |
| D | 0 | 0 | 0 | 5 | 10 |

表 4.17 よりベイジアンネットワークによる予測的中精度は約 79.28 %となっていることが

わかり, 表 4.14 に比べて予測的中精度が向上していることがわかる．またこのモデルにおいて A 年度を評価データとしたときの予測結果を表 4.18 に示す．

表 4.18: ベイジアンネットワークによる予測結果 (ホールドアウト法, 情報利得)

| 実際＼予測 | S | A | B | C | D |
|---|---|---|---|---|---|
| S | 4 | 1 | 0 | 0 | 0 |
| A | 5 | 47 | 9 | 0 | 0 |
| B | 0 | 6 | 64 | 3 | 1 |
| C | 0 | 0 | 8 | 17 | 1 |
| D | 0 | 0 | 0 | 1 | 4 |

表 4.18 よりベイジアンネットワークによる予測的中精度は約 79.53 %となった．表 4.15 と比べて説明変数を減らして予測をした場合, このモデルにおいては予測的中精度が下がってしまうことが分かった．以上より leave one out 法による検証では予測的中精度は良くなるが, 他の年度で予測する場合, 表 4.15 のモデルのほうが予測的中精度が良くなるため,B,C 年度においては情報利得で選ばれた変数を絞らずに全体で用いたほうが良いことがわかった．しかし変数を絞ったとしても B,C 年度データによるモデルは成績レベル S,D の予測的中精度が高く,leave one out 法だけでなくホールドアウト法で予測しても同様のことが確認できた．

### B,C 年度データによる特徴分析

B,C 年度（2 学年度分）のデータで予測を行ったとき, 情報利得によって B,C 年度のデータで一番最適なモデルを構築してしまうと, 他年度のデータでは予測的中精度が下がってしまうことが分かった．よって B,C 年度においては変数を減らさずにモデルを構築したほうが最適なモデルとなる．このとき構築されたベイジアンネットワークモデルを図 4.3 に示す．

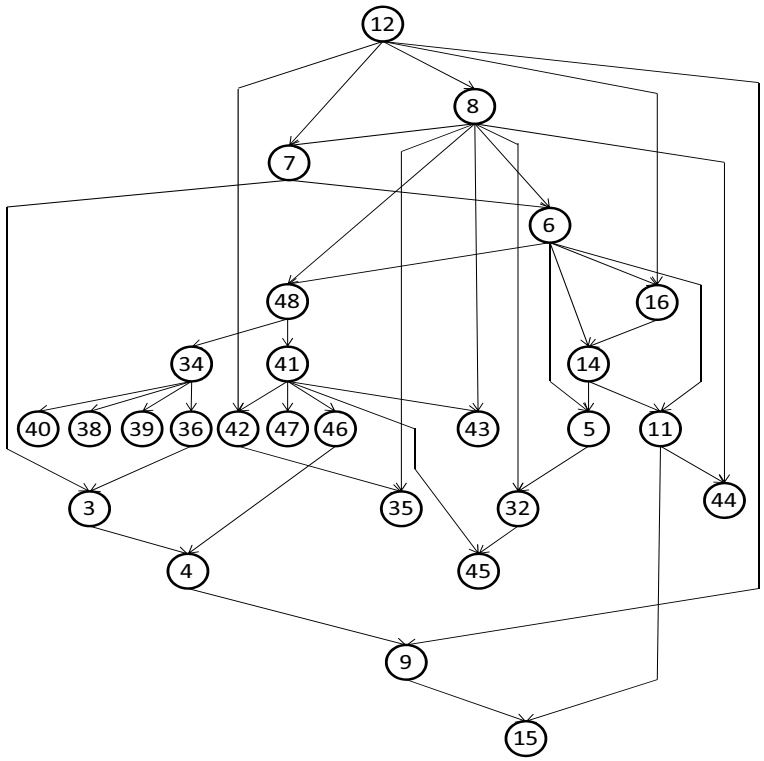


図 4.3: B,C 年度データによるベイジアンネットワークモデル

図 4.3 より目的変数は 11 月打刻回数と 1 月打刻回数の変数より有向グラフが引かれていることがわかる．よってこのモデルによる成績レベルの特徴は 11 月打刻回数と 1 月打刻回数の変数によって確率的に表現できる．このときの確率結果がどうなっているかを表 4.19 に示す．

表 4.19: B,C 年度データのモデルによる成績レベル表現

| 11 月打刻回数 | 1 月打刻回数 | S | A | B | C | D |
|---|---|---|---|---|---|---|
| $\leq$ 59.5 | $\leq$ 16 | 3.4 % | 3.4 % | 3.4 % | 24.1 % | 65.5 % |
| $\leq$ 59.5 | 16 - 39.5 | 2.6 % | 2.6 % | 28.2 % | 53.8 % | 12.8 % |
| $\leq$ 59.5 | 39.5 $<$ | 4 % | 28 % | 20 % | 20 % | 28 % |
| 59.5 - 78.5 | 16 - 39.5 | 3.4 % | 3.4 % | 44.8 % | 37.9 % | 10.3 % |
| 59.5 - 78.5 | 39.5 $<$ | 2.4 % | 32.4 % | 52.9 % | 11.9 % | 0.3 % |
| 78.5 $<$ | 39.5 $<$ | 9.6 % | 49.5 % | 38.1 % | 2.5 % | 0.4 % |

表 4.19 より 11 月打刻回数が 59.5 より低いと 1 月打刻回数にもよるが全体的に成績レベルが悪くなりやすいことがわかる．また 11 月打刻回数が 78.5 より高いと成績レベルは良くな

りやすいことがわかる.

### 4.3.5　A,C 年度データ

#### A,C 年度データによる予測結果

A 年度 171 人,C 年度 166 人の計 337 人のデータを用いて予測を行う．情報利得により選択された変数は表 4.20 のようになった.

表 4.20: 情報利得により取捨選択された変数 (A,C 年度データ)

| 番号 | 変数名 |
|---|---|
| 3 | 1 年 6 月打刻回数 |
| 4 | 1 年 7 月打刻回数 |
| 5 | 1 年 10 月打刻回数 |
| 6 | 1 年 11 月打刻回数 |
| 7 | 1 年 12 月打刻回数 |
| 8 | 1 年 1 月打刻回数 |
| 9 | 前期 $D_I \leq 0.5$ の出現率 |
| 11 | 前期 $0.8 < D_I \leq 1.2$ の出現率 |
| 12 | 後期 $D_I \leq 0.5$ の出現率 |
| 13 | 後期 $0.5 < D_I \leq 0.8$ の出現率 |
| 14 | 後期 $0.8 < D_I \leq 1.2$ の出現率 |
| 15 | 前期各 $D_I$ 差 |
| 16 | 後期各 $D_I$ 差 |
| 26 | 教養 10（平均） |
| 34 | 1 年前期 |
| 35 | 1 年前期英語 |
| 36 | 1 年前期人文 |
| 37 | 1 年前期体育 |
| 38 | 1 年前期専門 |
| 39 | 1 年前期理科 |
| 40 | 1 年前期数学 |
| 41 | 1 年後期 |
| 42 | 1 年後期英語 |
| 43 | 1 年後期人文 |
| 44 | 1 年後期体育 |
| 45 | 1 年後期専門 |
| 46 | 1 年後期理科 |
| 47 | 1 年後期数学 |

A,C 年度のデータで予測を行うとき, 表 4.20 に書かれている変数が目的変数を予測するのに有効な変数であることが分かった．このときまず A,C 年度のデータで予測する際,leave one out 法による検証結果を表 4.21 に示す.

表 4.21: ベイジアンネットワークによる予測結果 (leave one out 法)

| 実際＼予測 | S | A | B | C | D |
|---|---|---|---|---|---|
| S | 7 | 5 | 0 | 0 | 0 |
| A | 1 | 97 | 21 | 0 | 0 |
| B | 0 | 25 | 108 | 12 | 0 |
| C | 0 | 1 | 6 | 38 | 5 |
| D | 0 | 0 | 0 | 4 | 7 |

表 4.21 よりベイジアンネットワークによる予測的中精度は約 76.26 %となり, 上記で述べてきたパターンによる予測的中精度の中で一番低い精度であることが確認できる．また A,C 年度を学習データとして B 年度を評価データとして用いた場合どうなるか検証した．このときの予測結果を表 4.22 に示す.

表 4.22: ベイジアンネットワークによる予測結果 (ホールドアウト法)

| 実際＼予測 | S | A | B | C | D |
|---|---|---|---|---|---|
| S | 3 | 6 | 0 | 0 | 0 |
| A | 0 | 55 | 6 | 0 | 0 |
| B | 0 | 14 | 49 | 9 | 0 |
| C | 0 | 0 | 3 | 12 | 1 |
| D | 0 | 0 | 0 | 5 | 4 |

表 4.22 よりベイジアンネットワークによる予測的中精度は約 73.65 %となり, A,C 年度データによるモデルは全体的に予測的中精度が他のパターンと比べて低く,A,B 年度データによるモデルと同様に成績レベル S と D の学生を予測するのにはあまり向いていないことが分かった.

次に leave one out 法において, 情報利得によって選択された変数から有効性の高い変数を順に選択していき予測的中精度が最も良くなったときの変数を表 4.23 に, 結果を表 4.24 に示す.

表 4.23: 予測的中精度が最良の時の変数 (A,C 年度データ)

| 番号 | 変数名 |
|---|---|
| 3 | 1 年 6 月打刻回数 |
| 4 | 1 年 7 月打刻回数 |
| 5 | 1 年 10 月打刻回数 |
| 6 | 1 年 11 月打刻回数 |
| 7 | 1 年 12 月打刻回数 |
| 8 | 1 年 1 月打刻回数 |
| 9 | 前期 $D_I \leq 0.5$ の出現率 |
| 11 | 前期 $0.8 < D_I \leq 1.2$ の出現率 |
| 12 | 後期 $D_I \leq 0.5$ の出現率 |
| 34 | 1 年前期 |
| 36 | 1 年前期人文 |
| 38 | 1 年前期専門 |
| 39 | 1 年前期理科 |
| 40 | 1 年前期数学 |
| 41 | 1 年後期 |
| 42 | 1 年後期英語 |
| 43 | 1 年後期人文 |
| 45 | 1 年後期専門 |
| 46 | 1 年後期理科 |
| 47 | 1 年後期数学 |

表 4.24: ベイジアンネットワークによる予測結果 (leave one out 法, 情報利得)

| 実際＼予測 | S | A | B | C | D |
|---|---|---|---|---|---|
| S | 7 | 5 | 0 | 0 | 0 |
| A | 1 | 98 | 20 | 0 | 0 |
| B | 0 | 25 | 108 | 12 | 0 |
| C | 0 | 1 | 6 | 38 | 5 |
| D | 0 | 0 | 0 | 4 | 7 |

表 4.24 よりベイジアンネットワークによる予測的中精度は約 76.56 %となった．表 4.21 に比べて予測的中精度が向上していることがわかる．またこのモデルにおいて B 年度を評価データとしたときの予測結果を表 4.25 に示す

表 4.25: ベイジアンネットワークによる予測結果 (ホールドアウト法, 情報利得)

| 実際＼予測 | S | A | B | C | D |
|---|---|---|---|---|---|
| S | 3 | 6 | 0 | 0 | 0 |
| A | 0 | 55 | 6 | 0 | 0 |
| B | 0 | 14 | 49 | 9 | 0 |
| C | 0 | 0 | 3 | 12 | 1 |
| D | 0 | 0 | 0 | 5 | 4 |

表 4.25 よりベイジアンネットワークによる予測的中精度は約 73.65 %となった．表 4.22 と比べて説明変数を減らして予測を行っても予測的中精度を保っており, 変数をへらして予測を行うほうが良いモデルであることがわかった．ただ A,C 年度データによるモデルは変数を減らしても, 全体的な予測的中精度が他のパターンと比べてやや低く, また成績レベルに関しても成績レベル A の予測的中精度は高いが成績レベル S や D はあまり予測できていない.

### A,C 年度データによる特徴分析

A,C 年度（2 学年度分）のデータで予測を行ったとき, 情報利得で選択された変数から有効性の高い変数を順に選択したときの結果が最も良かったので, そのときに構築されたベイジアンネットワークモデルを図 4.4 に示す.

図 4.4 より目的変数は 7 月打刻回数と 1 年後期 GPA の変数より有向グラフが引かれていることがわかる．よってこのモデルによる成績レベルの特徴は 7 月打刻回数と 1 年後期 GPA の変数によって確率的に表現できる．このときの確率結果がどうなっているかを表 4.26 に示す.

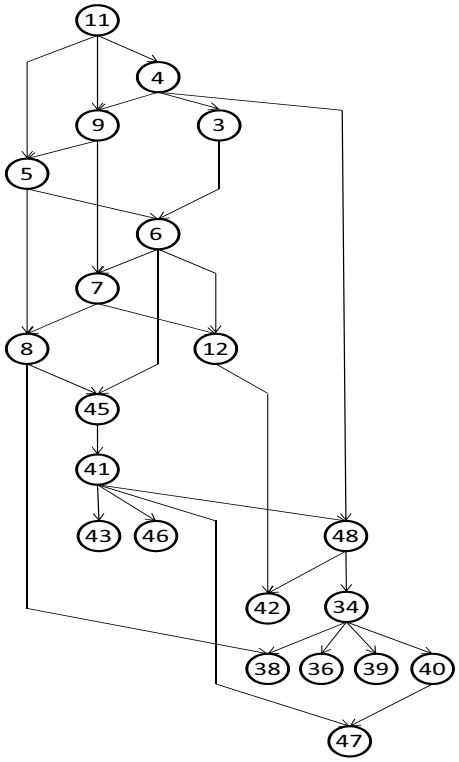


図 4.4: A,C 年度データによるベイジアンネットワークモデル

表 4.26: A,C 年度データのモデルによる成績レベル表現

| 1 年後期 GPA | 7 月打刻回数 | S | A | B | C | D |
|---|---|---|---|---|---|---|
| $\leq$ 0.982 | $\leq$ 39 | 4.8 % | 4.8 % | 4.8 % | 23.8 % | 61.9 % |
| $\leq$ 0.982 | 39 $<$ | 2.4 % | 2.4 % | 7.3 % | 61 % | 26.8 % |
| 0.982 - 1.625 | $\leq$ 39 | 9.1 % | 9.1 % | 9.1 % | 63.6 % | 9.1 % |
| 0.982 - 1.625 | 39 $<$ | 1 % | 1 % | 39.4 % | 57.6 % | 1 % |
| 1.625 - 2.339 | $\leq$ 39 | 11.1 % | 11.1 % | 11.1 % | 55.6 % | 11.1 % |
| 1.625 - 2.339 | 39 $<$ | 0.4 % | 10.6 % | 85.5 % | 3 % | 0.4 % |
| 2.339 - 2.673 | 39 $<$ | 0.9 % | 55.8 % | 41.6 % | 0.9 % | 0.9 % |
| 2.673 - 3.232 | $\leq$ 39 | 14.3 % | 42.9 % | 14.3 % | 14.3 % | 14.3 % |
| 2.673 - 3.232 | 39 $<$ | 2.1 % | 93.1 % | 3.4 % | 0.7 % | 0.7 % |
| 3.232 $<$ | 39 $<$ | 53.5 % | 39.5 % | 2.3 % | 2.3 % | 2.3 % |

表 4.26 より 1 年後期 GPA の値が低いほど成績レベルは悪くなりやすく, 高いほど成績レベルは良くなりやすい傾向を持つことが分かった．また 7 月打刻回数においても値が 39 より低いと成績レベルが悪くなりやすく, 値が 39 より高いと成績レベルは良くなりやすい傾向を持つことが分かった．

以上の結果より学生に対して指導を行うという点で考えると成績不振者の予測的中精度が高い B,C 年度データによるモデルで予測するのが効果的であると考えられる．また B,C 年度データによるモデルでは成績優秀者の予測的中精度も高いため成績が良くなる学生の傾向も事前に推測できる．また本章で述べた表 4.3 の変数番号 17〜33 の変数は予測する際に情報利得結果よりほとんどが変数として有効でないと判定されてしまうため, 本研究の結果からは成績レベルを予測するのに入室時間による特徴は見られないことが確認できた．

# 第5章　むすび

本章では, 本研究の結果, 今後の課題・展望について述べる.

## 5.1　本研究で得られた結果

本稿では, 名古屋工業大学を卒業した学生データを 3 学年度分用意し, ベイジアンネットワークによる予測技術を用いて予測を行った．また予測を行うにあたり, 更なる予測的中精度の向上や早期の指導を行うために変数の変更・追加や予測時期の変更を行った．A,B,C 年度データにおける予測,A,B 年度における予測,B,C 年度における予測,A,C 年度における予測の 4 パターンを試し, 構築されたモデルから打刻によるデータの有用性について特徴分析した.

ベイジアンネットワークによる予測の結果, どのパターンも約 8 割ほどの予測的中精度を持っていることが分かった．またその時の成績レベル予測表をみると B,C 年度のデータによる予測に関して成績レベル S,D の予測的中精度が leave one out 法, ホールドアウト法共に高く成績優秀者や成績不振者を予測したい場合に最も適していることが分かった．またベイジアンネットワークによって作成されたモデルから, 予測したい成績レベルは 11 月打刻回数と 1 月打刻回数から確率的に表現できることが分かり, 打刻データの有用性を見つけることができた.

## 5.2　今後の課題・展望

本研究では成績データ, 打刻データ, 出欠データを基データとして予測を行っている．また打刻データ, 出欠データに関しては出欠補正データという形で用いることで代用している. そのため予測を行うにあたってはデータをある程度信頼性のあるものとして採用しているが実際はどれだけ誤差が含まれているかはわからない．これはシステム上の問題によるところも大きいため今後も正確なデータを収集できるかどうかはわからないができるだけ正確なデータを得る為に, 補正方法の工夫やシステムの改善が必要であると考えている.

また本研究で用いているデータは名古屋工業大学におけるある学科のデータ 3 学年度分を使用しており, 本稿で述べた結果が他の学科あるいは他の大学でも利用できるかが分からない．同じ学科であったとしてもカリキュラムの違う学年度のデータに対応できるかはわからない．そのため今後は他学科のデータや他大学のデータを用いて本研究で行った操作を同様に検証することや, 新たなデータによる変数の追加を行いカリキュラムなどに縛られない変数による特徴を見つけることが今後の問題として言える．本研究で用いている打刻に関するデータは, 最近では出欠確認システムを導入している大学も多いためカリキュラムに縛られ

ない変数としては有用であると言える．今後の課題としては家から大学までの通学手段や通学にかかる時間を考えた打刻データによる特徴を見つけていきたいと考えている．

# 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，黒柳奨　准教授，伊藤宏隆　助教に心から感謝致します.

また, 本研究の実験のためのデータの提供元である, 出欠システム及びコースマネー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，情報基板センター教職員の皆様に心から感謝いたします.

最後に，本研究に多大な御協力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] 伊藤宏隆, 舟橋健司, 中野智文, 内匠逸, 松尾啓志, 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,4 巻,2 号, pp.15-21（2008）

[2] 岡田正，高橋参吉，藤原正敏，ICT 基礎教育研究会，“ネットワーク社会における情報の活用と技術”，実教出版（2010）

[3] 山中洋雄，中山実，清水康敬 “ICT 活用での形成的評価による学習成績・意欲に関する一考察”，電子情報数新学会技術研究報告．ET，教育工学 108 巻，247 号，pp.39-44（2008）

[4] 本村陽一，竹中毅，石垣司，“サービス工学の技術：ビッグデータの活用と実践”，東京電機大学出版局（2012）

[5] 文部科学省，“学習者等の視点に立った適切な e-Learning の在り方に関する調査研究報告書”（2007）

[6] 松尾啓志，“情報基板システムが支えるケータイ世代の学びの場とは?”, サイエンティフィック・システム研究会教育環境分科会, 第 1 回会合（2009）

[7] 堀江匠, “データマイニングによる学生の修学傾向分析とその修学指導への適用有効性の検証”, 平成 20 年度名古屋工業大学卒業研究論文（2008）

[8] 伊藤宏隆, 舟橋健司, 内匠逸, 松尾啓志, “IC カード出欠データと CMS 学習データを用いたデータマイニング”, 日本 e-Learning 学会誌，9 巻, pp.95-108（2009）

[9] 伊藤暁人, “ニューラルネットワークによる学生の成績予測とその学習指導への適用可能性の検討”, 平成 22 年度名古屋工業大学卒業研究論文（2010）

[10] K. Itoh, H. Itoh, K. Funahashi, ”Forecasting students’ grades using a Bayesian network model and an evaluation of its usefulness”, Proc. SNPD2012, pp.331-336 (2012)

[11] 平田大智，“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）

[12] 稲垣諒，“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）

[13] 本村陽一，“ベイジアンネットワーク:入門からヒューマンモデリングへの応用まで”，行動計量学会セミナー資料（2004）

[14] 鈴木譲，“ベイジアンネットワーク入門”，培風館（2009）

[15] 繁桝数男，上野真臣，本村陽一，“ベイジアンネットワーク概説”，培風館（2006）

[16] N. Friedman, D. Geiger, M. Goldszmidt, “Bayesian Network Classifers”, Machine learning, pp.131-163 (1997)

[17] 松尾啓志，“IC カード出欠システムを用いた不登校学生早期把握と災害時人情報把握への取り組み”，サイエンティフィック・システム研究会教育環境文科会，第 1 回会合（2012）

[18] Weka
http://www.cs.waikato.ac.nz/ml/weka/

# 付録　Wekaの操作方法について

本研究ではベイジアンネットワーク手法を用いるためにフリーソフトウェアである Weka を採用している．Weka はベイジアンネットワークだけでなく他のデータ解析や予測モデリングなどを行う機能を備えているが, ここでは本研究で用いたベイジアンネットワーク手法の実行方法について記述しておく．なお本研究で用いている Weka は Version3-6-7 である．

## 実験の流れ

まず予測を行う前にデータのセット方法について述べる．Weka では Excel で作成した「.csv」形式のデータを用いることができるため, あらかじめ説明変数と目的変数が記録されているデータを作成しておく．データのセット方法として「前処理」のページから「ファイルを開く」ボタンで使用するデータを選ぶことでセットできる．データをセットできたら情報利得による変数選択を行う.「属性選択」のページから属性検証より「GainRatioAttributeEval」を選び, 検索方法から「Ranker」を選ぶ．目的変数を選び開始ボタンを押すと出力画面に目的変数を予測するのに有用な変数を順に並べた結果が表示される．この結果から必要な変数部分のみ前処理ページの変数欄から選択し他は消去ボタンで変数を減らしておく．次にベイジアンネットワークの実行方法を述べる．分類のページより分類器で「bayes」の BayesNet を選択する．本研究ではベイジアンネットワーク予測における確率モデル評価指標として BIC, 学習方法として山登り法, モデルの構築方法として TAN を採用している．設定は分類器の部分で選んだ BayesNet の欄を押すと設定画面が出てくるのでその画面中の searchAlgorithm から「HillClimber」を選択することで山登り法の設定ができる．また HillCimber の欄を押すと山登り法の設定画面が出てくる．その設定画面の intAsNaiveBayes を「False」に, maxNrOfparents を「2」に,scoreType を「BAYES」にすることで BIC,TAN の設定ができる．あとは leave one out 法による検証で予測するのならテストオプション欄の交差検証をデータ数に設定し, ホールドアウト法による検証で予測するのなら供給テストセットで評価データとして用いるデータをセットすることでベイジアンネットワークによる各種設定方法の予測をすることができる．

# 発表論文リスト

1. 伊藤雄真，伊藤宏隆，舟橋健司，山本大介，齋藤彰一，松尾啓史，内匠逸，“学生データの変数の改善による将来の学生の成績レベル予測”，電気・電子・情報関係学会東海支部連合大会，O2-6（2014）

2. H. Itoh, Y. Itoh, K. Funahashi, D. Yamamoto, S. Saito, H. Matsuo, I. Takumi, “Forecasting Students’ Future Academic Records Using Bayesian Network”, SCIS&ISIS，TP4-3-3-(4)（2014）

3. 伊藤雄真，伊藤宏隆，舟橋健司，松尾啓史，内匠逸，“出欠状況を考慮した将来の学生の成績レベル予測及び特徴分析”，電気・電子・情報学会東海支部連合大会，A3-4（2015）

4. H. Itoh, Y. Itoh, K. Funahashi, “Forecasting Future Students’ Academic Level and Analyzing Students’ Feature Using Schooling Logs”, IEEE GCCE, pp. 288-291（2015）