平成26年度 卒業論文

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

指導教員
舟橋　健司　准教授
伊藤　宏隆　助教

名古屋工業大学　工学部　情報工学科
平成22年度入学　22115121番

平田　大智

目 次

第1章　はじめに

近年，IT 技術の発展と共に世の中には様々な情報が増え続けている．そのため，蓄積された大量のデータの中から有用な情報を得るデータマイニングが注目されており，実際に商業や医療の分野で有効利用されている．有名な事例で,『紙おむつを買う人はビールも買うことが多い』ことが発見されたことで，両者を近くに置くことで売り上げが上昇した．という話や，身近な所では，インターネットを利用した買い物では関連商品が表示されたり，購入履歴からおすすめの商品が表示される例がある．

また，教育における分野でも様々な電子化が進んでいる．名古屋工業大学では，早期の修学指導を目的としたコースマネージメントシステムと IC カードによる出欠システムを連携した双方向型教育支援システムが 2007 年より導入されている [1]．コースマネージメントシステムは，情報技術やインターネットを使った e-Learning を支援するシステムであり，教材の作成支援，課題の提出管理，小テストの実施，学生の受講管理を行う機能を有している．これにより，個々の学生のデータが逐次蓄積される．また，IC カードによる出欠システムは，IC カードにより学生の出欠状況を把握し，学生の修学指導に役立てようとするものであり，IC カード化された学生の身分証を各教室に設置された IC カードリーダにかざすことで，時刻情報を記録する．記録された時刻情報を教員が Web 上で確認することにより，学生の出欠状況を把握することができる．これらの蓄積された学生のデータを参照することで，総合的な成績評価が可能になる．このように蓄積されたデータに対してデータマイニングを行うことにより，成績評価にとどまらない有用な情報を見つけ出すことができると考えられる．

過去の関連研究として，学生の早期学習指導を目的とし，パターン認識に強力なニューラルネットワークを用いて成績予測を行う研究 [2] や，成績データと打刻データから将来の成績レベルを予測する研究 [3]，教員の修学指導の負担を減らす事を目的とし，今後指導を与えるべき学生を，未来事象の予測に活用されているベイジアンネットワークを利用して予測する研究 [4] がある．

本研究では，今後指導を与えるべき学生，関連研究でも取り上げられたいわゆる『要注意学生』を半期ごとのデータを用いて予測を行い，その精度の向上をはかる．半期ごとに予測を行う事により，学業不振に陥る学生の急な学力低下を，通年データを利用した場合に比べ，前後の時期のデータに影響を受けないため，より確実に拾うことができ，修学指導が必要な学生を広く拾うことができる可能性がある．また，半期ごとに広く要注意学生を予測することにより，新規の要注意学生を発見することができ，通年データで予測をするよりも累計の要注意発見数は多くなることが考えられる．また，もう一つ着目したのが,『要注意学生』の定義の見直しである．従来研究における定義は,『1 年前期と後期の GPA がともに 1.0 以上である，留年もしくは退学した学生』であったが，(前半の条件は，1 年次 GPA が 1.0 未満の学生はほぼ 100 %の割合で留年もしくは退学しており，予測が容易なためである．) 退学の理由として，学業不振という事が考えられ，そのような学生が本研究の予測の対象となるが，それ以外にも学業に問題はなくとも経済的理由によるものや，他大学受験という理由で退学するものも存在する．文部科学省の調査では，平成 24 年度のデータで最も中途退学者に多い理由はその他を除き，経済的理由であり平成 19 年度と比較し，割合が増えてい

る．[5] というものがある．そのため，本研究の目的は，学業不振となりうる学生の予測であり,『要注意学生』としてすべての退学者を一括りにしてしまうのは問題があると考え，その定義を見直した．

予測に用いるベイジアンネットワークは，因果的な特徴を有向グラフ構造により表し，個々の変数の関係を条件付き確率で表す確率推論のモデルであり，データマイニングにおいて未来事象の予測に利用されている手法である．また，構築されたモデルの精度検証は，leave one out 法を用い，正解率，再現率，適合率，F 値により評価を行った．

本論文の構成を説明する．第 2 章において本研究で用いるデータマイニングの手法や，予測手法の理論を述べ，第 3 章において本研究で用いる学生データの形式や拡張内容について述べる．また，第 4 章において『要注意学生』の定義を行い，第 3 章で述べたデータを用いて『要注意学生』を半期ごとに予測するモデルの提案及び検証する．最後に，第 5 章において本研究の結論と今後の課題を述べ，むすびとする．

ちなみに，本研究で用いられている学生のデータに関して，個人を特定できる情報（氏名，学籍番号）は一切含まれておらず，仮の番号を用いて管理しているため，本研究により個人情報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

本研究の『要注意学生』の予測手法として，未来事象の予測に活用されるベイジアンネットワークを採用した．本章ではベイジアンネットワークの概要と共に，予測精度向上に利用したデータマイニングの手法である属性選択とクラスリングに関して説明する．

2.1　ベイジアンネットワーク

ベイジアンネットワークとは，複数の確率変数の間の依存関係をグラフ構造によって表し，個々の変数の関係を条件付き確率で表した確率モデルである [6]．確率モデルとして，確率変数，その間の関係を表すグラフ構造，条件付き確率の集合によって定義される．これを用いた確率計算により，不確実性を含む事象の予測が可能となり，知的情報システムの適用例として，障害診断が挙げられる．ベイジアンネットワークの一例を図 2.1 に示す．

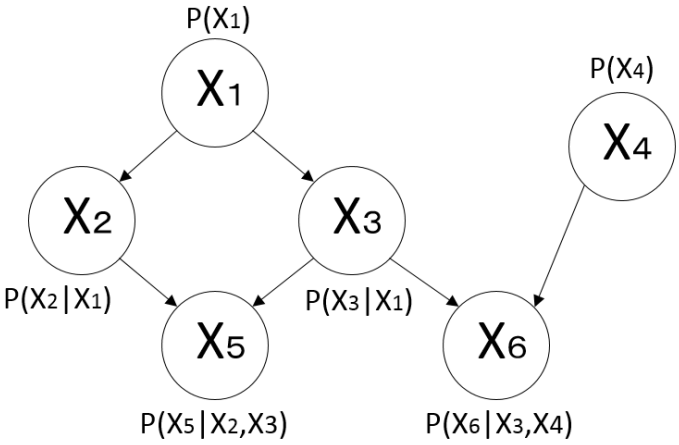

図 2.1: ベイジアンネットワークの例

この例は，確率変数 X_1, X_2, X_3, X_4, X_5, X_6と条件付き確率及び事前確率 $P(X_1)$, $P(X_4)$, $P(X_2|X_1)$, $P(X_3|X_1)$, $P(X_5|X_2, X_3)$, $P(X_6|X_3, X_4)$ が定義されており，それぞれの変数間が有効グラフにより結ばれている．これらの要素を決定することは，ベイジアンネットワークモデルを生成することと同義である．

2.1.1　ベイジアンネットワークのグラフ構造

予測に用いられるベイジアンネットワークは有向グラフにより構築されているため，そのグラフ構造により予測の結果は異なる．ここでは代表的な構造の説明を行う．

Naive bayes

図 2.2 に示すように，Naive Bayes はベイジアンネットワークの構造において最も簡単な構造であると言える．親ノードは一つしか存在せず，多くの場合は予測対象の目的変数が親ノードとなり，説明変数を子ノードとする．簡単な構造のため，条件付き確率の推定のみで構築することができるが，一概に子ノードを多くすれば精度が良くなるとは限らず，悪くなる可能性もあるため，適切な説明変数の選択が必要であると言える．有名な利用方法として，スパムメールの判別手法が挙げられる．

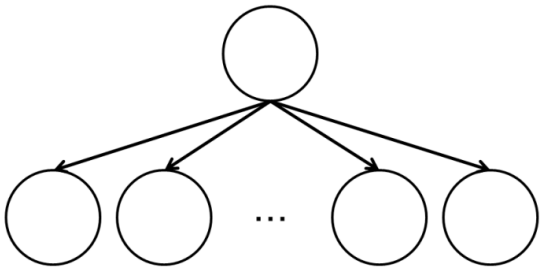

図 2.2: Naive Bayes の例

Tree Augmented Network

図 2.3 に示すのが Tree Augmented Network(TAN：以下 TAN と記述する) と呼ばれる構造である．Naive Bayes と似た構造をしているが，子ノードから他の子ノードにも 1 本のみ有向グラフが伸びており，子ノードは目的変数以外にも親ノードを持つ特徴がある．TAN 構造の決定指標には相互情報量が用いられる．

Free Network

Free Network は親ノードと子ノード数に制限が無いグラフ構造の総称である．はじめに挙げた図 2.1 も Free NetWork に分類される．しかし，Naive Bayes 同様，ノード数を増やせば精度が良くなるとは限らず，適切な変数の選択が必要である．また，親ノード数が増えるにつれ，必要となる条件付き確率が爆発的に増えてしまい，条件付き確率値に欠損が生まれる可能性もある．そのため，Free Network を用いる場合，親ノード数を制限し構造学習することが一般的である．

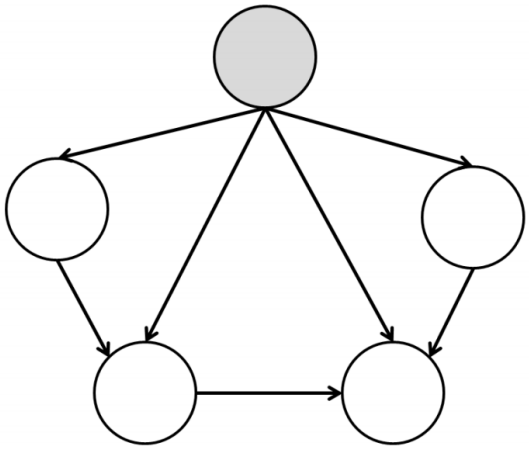

図 2.3: TAN の例

2.1.2　ベイジアンネットワークによる予測

ベイジアンネットワークを利用することで，一部の変数を観測した時，その他の変数の確率分布を求めたり，確率値が最も大きい状態をその変数の予測結果として得ることができる．これがベイジアンネットワークが未来予測の手法として用いられている理由である．確率計算に基づく予測は確率推論と呼ばれ，ベイジアンネットワークによる確率推論は以下の流れで行われる．

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．

という流れである．ここで，単純なモデル図 2.4 を用いて，計算の実行例を説明する．

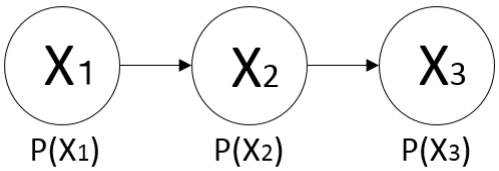

図 2.4: 単純なモデル例

変数間には図のような関係性があり，条件付き確率が与えられているとする．求めたい対象を X_2として，上流にある親ノードに与えられる観測情報を e^+，下流の子ノードに与えられる観測情報を e^-とする．求めたい事後確率 $P(X_2|e)$ は, e を e^+と e^-にわけ, X_2と e^-に注目してベイズの定理を使うと次のようになる．

$$\begin{aligned} P(X_2|e) &= P(X_2|e^+, e^-) \\ &= \frac{P(e^-|X_2, e^+)P(X_2|e^+)}{P(e^-|e^+)} \end{aligned}$$

また, e^+と e^-は X_2を固定した時は条件付き独立となり,$\alpha = \frac{1}{P(e^-|e^+)}$を X_2の値によらない正規化定数とす

れば，事後確率は次のようにできる．

$$P(X_2|e) = \alpha P(e^-|X_2)P(X_2|e^+) \tag{2.1}$$

このうち, e^+による X_2への寄与分，つまり親ノードから伝搬する確率を $P(X_2|e^+) = \pi(X_2)$ と書く．これは, $P(X_1|e^+)$ と X_2の条件付き確率を用い，次の式により求めることができる．

$$\pi(X_2) = \sum_{X_1} P(X_2|X_1)P(X_1|e^+) \tag{2.2}$$

$P(X_1|e^+) = \pi(X_1)$ は観測値が与えられているならば，その値は決定できる．観測値がなく，親ノードを持たない最上流のノードの場合，事前確率を与える．その上流に親ノードを保つ場合には式 (2.2) を再帰的に適用することでその値を求めることができる．

また, X_3から伝搬する確率を $P(e^-|X_2) = \lambda(X_2)$ とすると，定義されている条件付き確率 $P(X_3|X_2)$ を利用し次の式を用いればよい．

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

観測から得られる情報 e^-は X_2の値によらず独立であるため，次のように書き直せる．

$$\lambda(X_2) = \sum_{X_3} P(e^-|X_3)P(X_3|X_2) \tag{2.3}$$

ここで, $P(X_3|X_2)$ は事前に与えられており，親ノードからの伝搬と同様に $P(e^-|X_3) = \lambda(X_3)$ は観測情報が与えているならば値は決定できる．また，観測値がなくその下流に子ノードを持たない下端のノードの場合には，無情報であるため一様確率分布であるとして, X_3のすべての状態について等しい値とする．さらに下流に子ノードを保つ場合，式 (2.3) を再帰的に適用していけば値は定まるので,$\lambda(X)$ を計算することが可能である．

したがって，以上式 (2.2), (2.3) を式 (2.1) に代入することでノード X_2の事後確率を求める事ができる．同様に次の式により，任意のノードの事後確率も局所的に計算することができる．

$$P(X_j|e) = \alpha\lambda(X_j)\pi(X_j)$$

ベイジアンネットワーク内のすべてのパスがループを持たない場合，親ノードと子ノードが複数存在するような構造のネットワークでも，条件付き独立性の性質を用い，各ノードの上流，下流からの伝搬，上流，下流への伝搬の 4 種について計算することで任意のノードの事後確率を求める事ができる．

2.2　属性選択

属性選択は特徴選択，変数選択とも呼ばれ，すべての特徴集合のうち有用な部分集合だけを選択する手法のことである．不要で冗長なデータを除くことで，モデルの可読性が向上することや学習が高速化する長所を持つ．ここでは，属性選択として有名な主成分分析と，本研究に用いた情報利得について簡単に説明する．

2.2.1　主成分分析

主成分分析とは，多変量データを統合し，新たな総合指標を生み出す手法である．多くの変数に重みをつけることで少数の合成変数を作成するが，重みの付け方は，合成変数ができるだけ多く元の変数の情報量を含むようにつけられ，作成された合成変数は主成分と呼ばれる．また，主成分分析で得られる指標として，次のものが挙げられる．

固有値：主成分の分散に対応しており，その主成分がどの程度元のデータの情報を保持しているかを表す．

寄与率：ある主成分の固有値が表す情報が，すべての情報の中でどの程度の割合を占めるかを表す．

累積寄与率：各主成分の寄与率を大きい順に足したもので，そこまでの主成分でデータの持つ情報量が，どの程度説明されているかを表す．

また主成分数の選択は，一般的に累積寄与率が 70 %〜80 %あたりになる主成分が採用される．

2.2.2　情報利得

ベイジアンネットワークのモデル構築において，変数が多ければより精度が良くなるとは限らない．本研究では，多くの変数を利用しているため，中にはそれほど有用ではないデータが含まれており，予測のノイズとなっている可能性が考えられる．それらを除くために変数の取捨選択を行うが，その際に利用した指標が情報利得である．

情報利得は，カルバック・ライブラー情報量とも呼ばれ，2 つの確率分布の差異をはかる尺度であり，P, Q を離散確率分布とする時，次の式で定義される．

$$D(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

この情報利得を用いた変数選択指標として，本研究で利用した CFS(correlation based feature selection)[7] が挙げられる．ある変数と関連性の高い変数を選択する際に有効な手法である．CFS の値は以下の式で求めることができる．k は変数の個数, Z は目的変数を指す．この CFS 値を最大化するように変数 Y_iが選択される．

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

また, SU は情報量 H と情報利得 D を用いて次の式で求める事ができる．

$$SU(Y,Z) = 2 * \frac{D(Y||Z)}{H(Y) + H(Z)}$$

2.3　クラスタリング

ベイジアンネットワークを構築する際，説明変数は離散値である必要がある．本研究で用いる学生データは基本的に連続的な数値データであるため，離散化する必要ある．離散化の手法として用いたのが対象間の類似度に基づきグループ分けを行うクラスタリングであり，階層的クラスタリングの代表としてウォード法を，非階層的手法の代表として，K-means 法を説明する．ちなみに本研究においては，変数の離散化にウォード法を採用している．

2.3.1　ウォード法

2 つのクラスター P, Q を結合すると仮定したとき，それにより移動したクラスターの重心とクラスター内の各サンプルとの距離の 2 乗和 $L(P \cup Q)$ と，元々の 2 つのクラスター内での重心とそれぞれのサンプルとの距離の 2 乗和 $L(P)$, $L(Q)$ の差

$$\delta = L(P \cup Q) - L(P) - L(Q)$$

が最小となるようにクラスター同士を結合する手法．計算量は多いが，分類感度が良いため一般的によく用いられる．

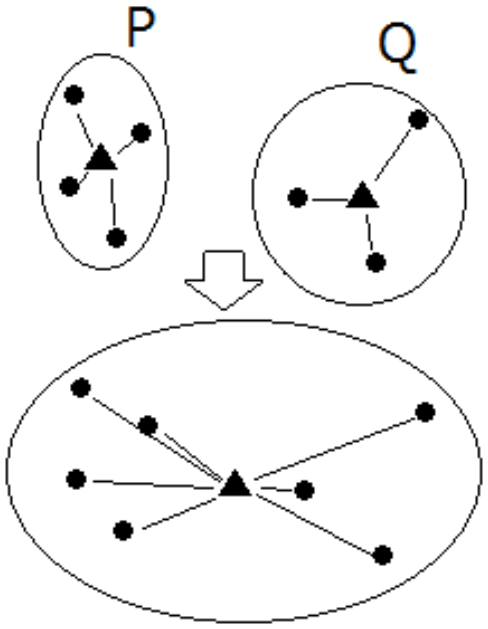

図 2.5: ウォード法のイメージ

2.3.2　K-means 法

クラスタの平均を用い，与えられたクラスタ数 K 個に分類するため，K-means 法や K-平均法と呼ばれる．アルゴリズムは単純であり，データ数を n，クラスタ数を K とした場合，次の流れで行われる．

1) 各データ $x_i(i = 1 \cdots n)$ に対してランダムにクラスタを割り振る．

2) 各クラスタのデータもとに中心 $V_j(j = 1 \cdots K)$ を計算する．基本的に計算は各要素の算術平均が使用される．

3) 各 x_iと各 V_jとの距離を求め痔 x_iを最も近い中心のクラスタに割り当て直す.

4) 上記の処理ですべての x_iのクラスタ割り当てが変化しなかった場合，あるいは変化量が事前に設定した一定の閾値を下回った場合，処理を終了する．そうでない場合，新たに割り振られたクラスタから V_jを再計算し，上記の処理を繰り返す.

単純なアルゴリズムで計算を行うため，実装が容易であり，実行も早い．そのため広く用いられているが，クラスタリングの結果は，初期クラスタのランダムな割り振りに大きく依存し，一度の処理で最良の結果が得られるとは限らない欠点も持ち合わせている．

第3章　本研究で用いるデータについて

ベイジアンネットワークによる予測モデルの構築において，その対象となるデータの質は，発見される新たな知識に直結しており，用いるデータの重要性はとても高い．本章では，本研究に用いるデータの概要と，予測に利用するために行ったデータの拡張について説明する．

3.1　用いる学生データの概要

本研究では，1 章で述べたコースマネージメントシステムや IC カード出欠システムより得られた名古屋工業大学を卒業した，338 名の学生データを用いている．338 名は 2 年度分に相当し，年度ごとに 171 名と 167 名に分けられる．主なデータの種類は 3 種であり，講義別成績データ，入退室時間に関するデータ (以下打刻データとする)，学生が卒業研究に着手した年次と卒業した年次が記載されたデータ (以下学生修学データとする) である．なお，データに記載されている番号は個人が特定できるような学籍番号ではなく，管理のためにつけられた仮の番号である．また，講義別成績データに記載されている講義名について，必須科目である英語や理系基礎科目など，全学生共通の講義名は変更されていないが，学生の学科が特定できるであろう科目は,「専門 1」や「演習 1」のように具体的な講義内容が分からないように変更されている．そのため，本研究により，個人情報が侵害されることはないことをあらためて記す．

3.2　データの拡張

講義別成績データは，学籍番号，講義名，GPA 数値，開講学期を 1 レコードとし，打刻データは，学籍番号，教室，打刻日，打刻時刻を 1 レコードとしたレコード形式で記録されている．全レコード数は 50 万にも及び，この形式のままではベイジアンネットワークによる予測モデル構築に利用しがたい．そのためにこれらのデータに行った拡張について説明する．

講義別成績データの拡張

個人の成績を表す指標として，広く一般に利用されている Grade Point Average（以下 GPA とする）に着目し，レコードデータから個人の GPA に変換し，科目ごとの GPA を算出した．また，GPA のみではわからない各評価（秀，優，良，可，不可，失格）の獲得数も，1 年次前期から 2 年次後期まで半期ごとに算出した．以下の表 3.1 が講義別成績データより拡張し，予測に利用した変数の一覧である．

表 3.1: 講義別成績データより拡張された変数一覧

番号	変数名	内容
1	1 年前期外国語 GPA	1 年次前期に受講した外国語に関する講義の GPA
2	1 年後期外国語 GPA	1 年次後期に受講した外国語に関する講義の GPA
3	1 年前期人文 GPA	1 年次前期に受講した人間文化に関する講義の GPA
4	1 年後期人文 GPA	1 年次後期に受講した人間文化に関する講義の GPA
5	1 年前期数学 GPA	1 年次前期に受講した数学系に関する講義の GPA
6	1 年後期数学 GPA	1 年次後期に受講した数学系に関する講義の GPA
7	1 年前期理科 GPA	1 年次前期に受講した理科系に関する講義の GPA
8	1 年後期理科 GPA	1 年次後期に受講した理科系に関する講義の GPA
9	1 年前期体育 GPA	1 年次前期に受講した体育科目に関する講義の GPA
10	1 年後期体育 GPA	1 年次後期に受講した体育科目に関する講義の GPA
11	1 年前期専門 GPA	1 年次前期に受講した専門科目に関する講義の GPA
12	1 年後期専門 GPA	1 年次後期に受講した専門科目に関する講義の GPA
13	1 年前期その他 GPA	1 年次前期に受講した上記に属さない講義の GPA
14	1 年後期その他 GPA	1 年次後期に受講した上記に属さない講義の GPA
15	2 年前期外国語 GPA	2 年次前期に受講した外国語に関する講義の GPA
16	2 年後期外国語 GPA	2 年次後期に受講した外国語に関する講義の GPA
17	2 年前期人文 GPA	2 年次前期に受講した人間文化に関する講義の GPA
18	2 年後期人文 GPA	2 年次後期に受講した人間文化に関する講義の GPA
19	2 年前期数学 GPA	2 年次前期に受講した数学系に関する講義の GPA
20	2 年後期数学 GPA	2 年次後期に受講した数学系に関する講義の GPA
21	2 年前期理科 GPA	2 年次前期に受講した理科系に関する講義の GPA
22	2 年後期理科 GPA	2 年次後期に受講した理科系に関する講義の GPA
23	2 年前期体育 GPA	2 年次前期に受講した体育科目に関する講義の GPA
24	2 年後期体育 GPA	2 年次後期に受講した体育科目に関する講義の GPA
25	2 年前期専門 GPA	2 年次前期に受講した専門科目に関する講義の GPA
26	2 年後期専門 GPA	2 年次後期に受講した専門科目に関する講義の GPA
27	2 年前期その他 GPA	2 年次前期に受講した上記に属さない講義の GPA
28	2 年後期その他 GPA	2 年次後期に受講した上記に属さない講義の GPA

番号	変数名	内容
29	1 年前期秀	1 年次前期に獲得した成績評価秀の数
30	1 年後期秀	1 年次後期に獲得した成績評価秀の数
31	1 年前期優	1 年次前期に獲得した成績評価優の数
32	1 年後期優	1 年次後期に獲得した成績評価優の数
33	1 年前期良	1 年次前期に獲得した成績評価良の数
34	1 年後期良	1 年次後期に獲得した成績評価良の数
35	1 年前期可	1 年次前期に獲得した成績評価可の数
36	1 年後期可	1 年次後期に獲得した成績評価可の数
37	1 年前期不可	1 年次前期に獲得した成績評価不可の数
38	1 年後期不可	1 年次後期に獲得した成績評価不可の数
39	1 年前期失格	1 年次前期に獲得した成績評価失格の数
40	1 年後期失格	1 年次後期に獲得した成績評価失格の数
41	2 年前期秀	2 年次前期に獲得した成績評価秀の数
42	2 年後期秀	2 年次後期に獲得した成績評価秀の数
43	2 年前期優	2 年次前期に獲得した成績評価優の数
44	2 年後期優	2 年次後期に獲得した成績評価優の数
45	2 年前期良	2 年次前期に獲得した成績評価良の数
46	2 年後期良	2 年次後期に獲得した成績評価良の数
47	2 年前期可	2 年次前期に獲得した成績評価可の数
48	2 年後期可	2 年次後期に獲得した成績評価可の数
49	2 年前期不可	2 年次前期に獲得した成績評価不可の数
50	2 年後期不可	2 年次後期に獲得した成績評価不可の数
51	2 年前期失格	2 年次前期に獲得した成績評価失格の数
52	2 年後期失格	2 年次後期に獲得した成績評価失格の数

打刻データの拡張

レコード形式として記録されている打刻データは，学籍番号，教室，打刻日，打刻時刻である．勤勉な学生は講義ごとに教室へ入室する時と退出する際に 2 度打刻を行うが，欠席した学生は打刻されない．そのため，学習姿勢をはかる指標として，学生個人の打刻回数に着目し，打刻日のデータからひと月ごと打刻回数へと拡張を行った．以下の表 3.2 が打刻データより拡張し，予測に利用した変数の一覧である．

表 3.2: 打刻データより拡張された変数一覧

番号	変数名	内容
1	1 年 4 月打刻数	1 年次 4 月に行った打刻の回数
2	1 年 5 月打刻数	1 年次 5 月に行った打刻の回数
3	1 年 6 月打刻数	1 年次 6 月に行った打刻の回数
4	1 年 7 月打刻数	1 年次 7 月に行った打刻の回数
5	1 年 8 月打刻数	1 年次 8 月に行った打刻の回数
6	1 年 9 月打刻数	1 年次 9 月に行った打刻の回数
7	1 年 10 月打刻数	1 年次 10 月に行った打刻の回数
8	1 年 11 月打刻数	1 年次 11 月に行った打刻の回数
9	1 年 12 月打刻数	1 年次 12 月に行った打刻の回数
10	1 年 1 月打刻数	1 年次 1 月に行った打刻の回数
12	2 年 4 月打刻数	2 年次 4 月に行った打刻の回数
13	2 年 5 月打刻数	2 年次 5 月に行った打刻の回数
14	2 年 6 月打刻数	2 年次 6 月に行った打刻の回数
15	2 年 7 月打刻数	2 年次 7 月に行った打刻の回数
16	2 年 8 月打刻数	2 年次 8 月に行った打刻の回数
17	2 年 9 月打刻数	2 年次 9 月に行った打刻の回数
18	2 年 10 月打刻数	2 年次 10 月に行った打刻の回数
19	2 年 11 月打刻数	2 年次 11 月に行った打刻の回数
20	2 年 12 月打刻数	2 年次 12 月に行った打刻の回数
21	2 年 1 月打刻数	2 年次 1 月に行った打刻の回数

要注意学生の予測は，これらの講義別成績データと打刻データから拡張し得られた変数群を利用した．これらの変数群は数値化されており，連続値である．ベイジアンネットワークに用いる確率変数は離散化されている必要があるため，これらの変数に離散化を行った．手法として，ウォード法によるクラスタリングを利用し変数ごとの属性数は４とし，モデル構築を行った．

第4章　要注意学生発見モデルの構築

本章では前章で説明した変数を用い，ベイジアンネットワークを利用した要注意学生発見モデル構築，及び検証について述べる．

4.1　発見の概要

本研究の『発見』とは，得られた学生データを用い，将来要注意学生となるか否かの『未来予測』に相当する．予測を行うことにより，将来的に要注意学生になるであろう学生に早期の修学指導を行うことが可能になり，修学環境の改善が期待できる．また，予測の精度を向上させることは，より多くの要注意学生を発見できることと同義であり，研究の目的となる．本節では，発見の対象者とした『要注意学生』の具体的な定義と，構築されたモデルの評価方法について以下に説明する．

4.1.1　発見対象者の定義

一般的に名古屋工業大学では，4 年次に卒業研究が開始されるが，卒業研究着手条件とされる単位数を取得できなかった学生は，卒業研究が開始できず，事実上留年となる．3 章で述べた学生の修学データには，卒業研究に着手した年次と卒業した年次が記載されている．そのデータをまとめた表が以下の表 4.1，表 4.2 である．ここで,『未着手』は記録上卒業研究に着手できてないことを表し,『退学』は卒業研究着手または卒業までに退学届が受理された学生数を,『在学中』は卒業しておらず，籍だけ置かれている学生を表している．

表 4.1: 各年度の卒業研究着手に要した年数

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

表 4.2: 各年度の卒業に要した年数

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

表から，データ対象の学生全 338 名のうち，283 名が 3 年で順調に卒業研究に着手しているが，反対に 55 名が 4 年次に卒業研究に着手できておらず，割合にすると全体の約 15 %が学業になんらかの問題を抱えていることがわかる．また，次の表 4.3 が 4 年で卒業できず，学業になんらかの問題を抱えたであろう学生 70 名に関する 1 年次の GPA データである．

表 4.3: 1 年次の GPA 値域別退学者及び留年者の割合

	1 年前期 GPA			1 年後期 GPA		
値域	全人数	退学，留年	割合	全人数	退学，留年	割合
0.0 以上 0.5 未満	5	5	100 %	11	11	100 %
0.5 以上 1.0 未満	8	6	75 %	12	12	100 %
1.0 以上 1.5 未満	11	7	64 %	31	14	45 %
1.5 以上 2.0 未満	48	23	48 %	67	15	22 %
2.0 以上 2.5 未満	105	16	15 %	96	10	10 %
2.5 以上 3.0 未満	100	10	10 %	70	2	3 %
3.0 以上 3.5 未満	53	3	6 %	42	4	10 %
3.5 以上 4.0 未満	8	0	0 %	7	0	0 %
合計	338	70		338	70	

上記の表から，GPA が高いほど退学，留年する学生の割合が低くなっていることがわかるが，GPA の高い値域の中にも少なからず退学，留年している学生がいることが確認できた．また，GPA が 1.0 を下回る学生に注目すると，前期，後期合わせて全人数 36 名中 34 名，割合にして約 94 %の学生が退学，留年していることがわかる．すなわち，1 年次の段階で GPA が 1.0 を前期または後期の段階で下回る場合，ほぼ確実に退学，留年するとも言える．これらから，1 年次の GPA が 1.0 を下回る場合，予測するまでもなく修学指導の対象とし，真に予測すべき対象となる学生は,『1 年次の GPA は 1.0 を上回るが，将来修学傾向が悪化し，退学または留年してしまう学生』であると言える．関連研究 [4] では，上記を予測すべき要注意学生の定義としていたが，本研究ではこの定義に関して，さらなる見直しを行う．

まず着目したのが，3 年で卒業研究に着手したが，卒業までに 5 年以上かかった学生である．これらの学生に考えられることは，卒業研究において 1 年で成果を残せず，指導教員の合格が得られなかった場合と就職活動に失敗し，戦略的に留年を選択した場合等が考えられる．これらの場合，3 年次までに学業不振で留年してしまう学生と性質が異なり，留年者としてまとめて要注意学生とするべきではないとした．また，次の図 4.1 は，文部科学省が行った調査 [5] で，平成 24 年度の中途退学者の状況である．

「その他」を除いて，最も高い割合を占めたのは「経済的理由」である．また，平成 19 年度より最も割合が増加しており，年々増加傾向にあると考えられる．本研究の予測の対象となるべき学生は「学業不振」を理由に退学してしまう学生の予測であり，早期の修学指導を行うことが目的である．そのため，経済的理由が最も多い退学理由である中，すべての退学者を一括りに『要注意学生』として予測の対象としてしまうには問題があると考える．以上から，本研究の予測対象となる要注意学生の定義に次のような条件を加えた．1)3 年で卒業研究に着手した場合，要注意学生とはしない．2) 入学から 3 年以内に退学した場合，データから除外する．1 つ目の条件により，戦略的に留年を選択した学生が除かれ，2 つ目の条件により，経済的な理由で退学した学生や，他大学受験の為退学した学生などといった学業不振ではない学生を予測対象

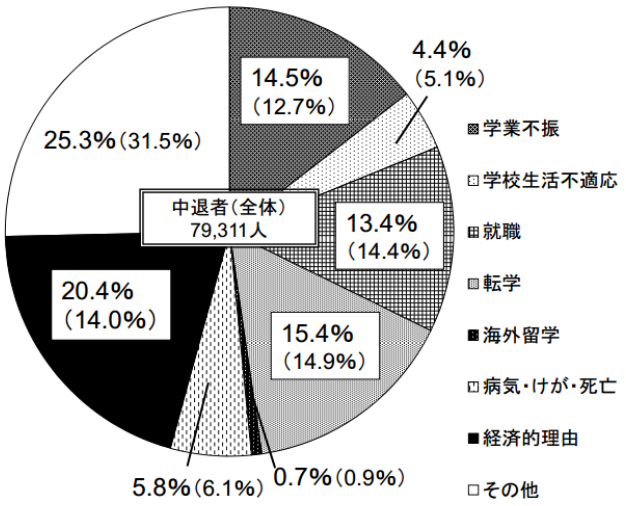

図 4.1: 平成 24 年度の中途退学者の状況（括弧内は平成 19 年度の値）

から除くことができる．この条件により，予測の対象となる学生は 302 名となり，その中で発見すべき要注意学生数は 41 名から 25 名となった．

4.1.2　構築された発見モデルの評価

要注意学生を予測するモデルを構築した際，予測の精度は発見される要注意学生数に直結し，重要であることは明白である．そのため，なんらかの指標で構築されたモデルを評価し，比較することが予測精度の向上に不可欠である．本研究では評価法に，leave one out 法を利用しモデル精度の評価を行った．また構築されたモデルを比較する指標として，以下に説明する，正解率 (Accuracy)，再現率 (Recall)，適合率 (Precision)，F 値 (F-measure) を利用した．

事実として，要注意学生である学生とそうでない学生が存在する学生集団に対して，一人ずつ要注意学生であるか否かの予測を行う．この時，要注意学生であることを Positive な事象であるとしたとき，実際に要注意学生に対し，要注意学生であると予測した場合を True Positive(以下 TP とする) と表す．この時それぞれの表記は次の表 4.4 のように表される．

表 4.4: 予測結果の表記一覧

	実際に要注意学生である	実際に要注意学生でない
要注意学生であると予測	True Positive(TP)	False Positive(FP)
要注意学生でないと予測	False Negative(FN)	True Negative(TN)

正解率 (Accuracy)：実際と予測に対する的中率を表す．

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

再現率 (Recall)：実際の要注意学生のうち，どれほど予測できたかを表す．

$$Recall = \frac{TP}{TP + FN}$$

適合率 (Precision)：要注意学生と予測した学生のうち，どれほど実際の要注意学生であったかを表す．

$$Precision = \frac{TP}{TP + FP}$$

F 値 (F-measure)：一般的に予測精度の評価指標とされる．適合率と再現率の調和平均である．

$$F - measure = \frac{2Recall * Precision}{Recall + Precision}$$

これらの指標を用いることで構築されたモデルの評価及び比較を行い，最も優れた精度のモデルを決定する．

4.2　ベイジアンネットワークによる要注意学生発見モデル

本節では，未来予測の手法としてベイジアンネットワークを利用し，3 章で述べたデータから，実際に要注意学生を予測するモデルの構築について説明する．

4.2.1　手法の概要

2 章で説明した通り，ベイジアンネットワークは確率変数，その間の関係を表すグラフ構造，条件付き確率の集合によって定義される．そのため，目的変数と説明変数の決定，有向グラフの学習，条件付き確率の推定が必要となる．本研究は修学に問題を抱えるであろう学生の予測であるため，目的変数は，Yes または No で表せる『要注意学生であるか否か』である．構築されたモデルにおいて，Yes が出力されたならば，その学生は要注意学生であること示し，修学指導が必要な学生であると考えられる．モデルの精度は説明変数によって異なり，その取捨選択が重要である．例えば説明変数を 1 年次のデータのみにすれば，1 年次の段階で目的変数である『要注意学生であるか否か』の判定が行えるため，早期の予測が可能となるが，2 年次までのデータを利用したものと比べれば，データ量が少なく，予測の精度は劣ることが考えられる．説明変数の取捨選択に関して，2 章でふれた属性選択の手法として用いられる CFS を利用した．さらに，ベイジアンネットワークに用いる確率変数は離散化されている必要がある．3 章で述べた GPA データや打刻データは数値化されており，連続値であるため離散化しなければならない．本研究では，離散化の方法として，ウォード法によるクラスタリングを利用し，属性数を 4 にすることでモデル構築を行った．有向グラフの学習と条件付き確率の推定に関しては，様々なデータ解析や予測モデリングのアルゴリズムを利用できるフリーのデータマイニングソフト『Weka』[8] を利用して行った．有効グラフ構造はすべて Naive Bayes 構造を採用している．

また，ベイジアンネットワークは出力形式がある事象の事後確率で出力される特徴を持つ．本研究の場合，目的変数は『要注意学生であるか否か』であり，Yes か No の二値的予測となる．一般的に閾値は 50 %とされ，事後確率が 50 %を超えた場合，予測モデルはある学生を要注意学生であると予測し，下回れば要注意学生ではないと予測される．そこで，閾値を任意に設定することで，より柔軟で正確な予測を行うことが可能となる．本研究では，事後確率の閾値を，50 %，30 %，事前確率の 3 通り設定し，それぞれのモデルで精度の検証を行った．また，事前確率は全体に対する実際の予測対象となる要注意学生の割合である．1 年次の GPA が 1.0 以上の学生数は 302 名，その内予測する要注意学生は 25 名であるため，事前確率は $25 \div 302 = 8.3$ % となる．

4.2.2　予測の時期とデータの範囲

本研究では予測に利用するデータの範囲を図 4.2 のような予測時期までのすべてのデータを利用し予測を行うモデルと，図 4.3 のような半期ごとのみのデータにより予測を行うモデル，半期ごとのみのデータに前回の予測結果を変数に加えたモデルの 3 つのパターンで半期ごとに要注意学生の予測モデルの構築を行った．

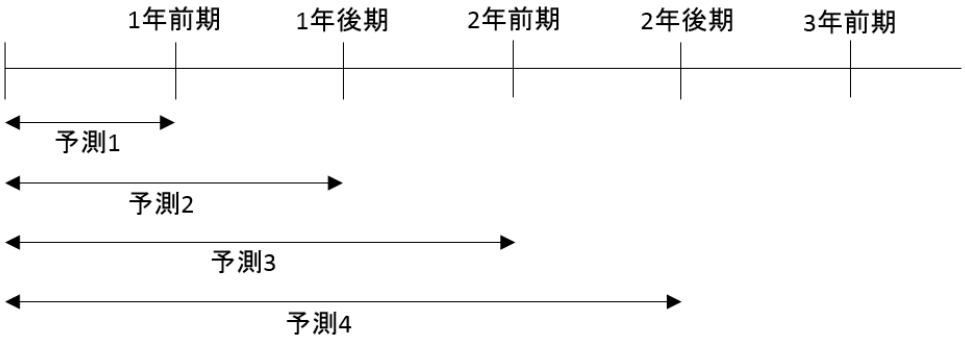

図 4.2: 予測時期までのすべてのデータを利用するモデルイメージ

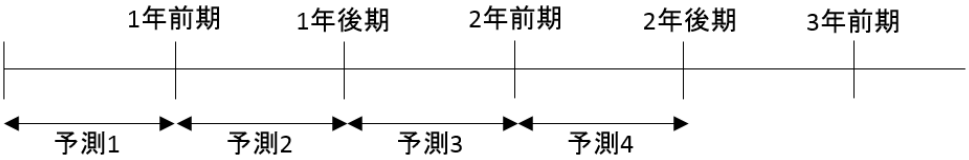

図 4.3: 半期ごとのデータのみを利用するモデルイメージ

関連研究に関して，予測の時期に関して研究ごとに違いは見られるが，利用するデータはその予測の時期までのすべてのデータを利用している．本研究で，新たに半期ごとのデータのみで予測を行うモデルを構築した理由は，要注意学生の成績の特徴である,『急な成績低下』を予測により反映させるためである．例えば，ある要注意学生は 1 年後期までの GPA は問題無かったが，専門科目が多くなる 2 年前期に急に成績が悪化したとする．この時，2 年次前期までのすべてのデータを利用した場合，1 年前期，後期の GPA に関して問題はなかったため，要注意学生ではないと予測される可能性があるが，2 年前期のみのデータで予測を行った場合，低い GPA から要注意学生であると予測することができる．反対に多くのデータを利用したからこそ見られる特徴も考えられるため，両方のパターンでモデル構築を行い，精度検証を行う．また，利用するデータは 3 章で説明した，科目別 GPA データのみの場合と，科目別 GPA データ，獲得成績データ，打刻データの 3 種類すべてのデータを利用した場合でモデル構築を行った．

また，半期ごとのデータのみで予測モデルの構築を行う際，前回の予測結果を新たな変数として導入し，発見精度の検証を行った．例えば，図 4.3 について予測 2 を行う場合，モデル構築に利用される変数は，1 年後期の半期データと予測 1 の判定結果，すなわち 1 年前期データによる予測結果である．この新たな変数を利用することで，予測時期までの全てのデータを利用するモデルとは異なる形で予測時期前の情報を半期ごとのデータに加えることができ，予測時期までの全てのデータを利用した場合とも，半期のみのデータを利用した場合とも異なる予測結果を得ることができる．

4.3　科目別 GPA のみを利用した要注意学生の発見

説明変数を科目別 GPA のみとし，ベイジアンネットワークを利用してモデル構築を行った．連続値である科目別 GPA はウォード法によるクラスタリングを行い，４つの属性値に離散化を行っている．変数の数は半期につき 7 変数である (詳しい内容は第 3 章表 3.1 を参照)．予測時期は半期ごととし，1 年前期から 2 年後期までのデータでモデル構築を行った．

4.3.1　予測時期までのすべてのデータを利用するモデル

図 4.2 に示したように，予測時期までのすべてのデータを利用するモデルで，説明変数として科目別 GPA のみを利用したモデルである．すべてで 4 つのモデルが構築されその精度一覧が以下の表 4.5 から表 4.8 である．グラフ構造はすべて Naive bayes 構造である．

1 年前期までのモデル

表 4.5: 1 年前期までの科目別 GPA のみを用いたモデルの精度一覧

閾値	正解率			再現率			適合率			F 値
	対象	的中		対象	的中		対象	的中		
50 %	302	270	89 %	25	5	20 %	17	5	29 %	0.238
30 %	302	260	86 %	25	9	36 %	35	9	26 %	0.300
8.3 %	302	215	71 %	25	16	64 %	94	16	17 %	0.269

1 年前期までのデータで全てで 7 変数しかなく，予測時期までのすべてのデータを利用するモデルの場合，最も説明変数が少ない予測モデルである．そのため，精度に関してはあまり良いとは言えない結果となった．F 値から，最も F 値がよくなった閾値は 30 %の時で，予測すべき要注意学生 25 名中，9 名を予測でき，1 年前期の段階で科目別 GPA を用いることで，36 %の要注意学生を発見できていることがわかる．

図 4.4: 1 年前期までの科目別 GPA を用いて構築されたグラフ

1 年後期までのモデル

1 年前期のデータ+後期のデータで，全 14 変数の科目別 GPA 値のみによる予測モデルは 1 年後期のデータを加えたことにより，前期のみの予測モデルよりも精度が向上しており，最も F 値が高くなったのは閾値 50 %のときで，予測すべき要注意学生 25 名中，14 名を予測でき，1 年後期の段階で科目別 GPA を用いることで 56 %の要注意学生を発見できていることがわかる．

表 4.6: 1 年後期までの科目別 GPA のみを用いたモデルの精度一覧

閾値	正解率			再現率			適合率			F 値
	対象	的中		対象	的中		対象	的中		
50 %	302	256	85 %	25	14	56 %	49	14	29 %	0.378
30 %	302	245	81 %	25	15	60 %	62	15	24 %	0.345
8.3 %	302	229	76 %	25	19	76 %	86	19	22 %	0.342

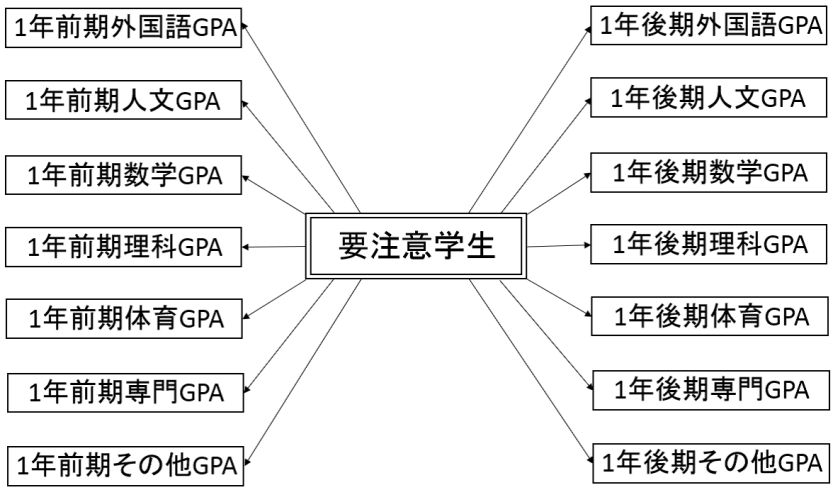

図 4.5: 1 年後期までの科目別 GPA を用いて構築されたグラフ

2 年前期までのモデル

表 4.7: 2 年前期までの科目別 GPA のみを用いたモデルの精度一覧

閾値	正解率			再現率			適合率			F 値
	対象	的中		対象	的中		対象	的中		
50 %	302	255	84 %	25	16	64 %	54	16	30 %	0.405
30 %	302	246	81 %	25	16	64 %	63	16	25 %	0.364
8.3 %	302	227	75 %	25	18	72 %	86	18	21 %	0.324

さらに 2 年前期の科目別 GPA データを加え，全 21 変数の科目別 GPA 値のみによる予測モデル結果である．最も F 値が高くなったのは，閾値 50 %の時で，予測すべき要注意学生 25 名中，16 名を予測でき，2 年前期までの科目別 GPA を用いることで 64 %の要注意学生を発見できていることがわかる．

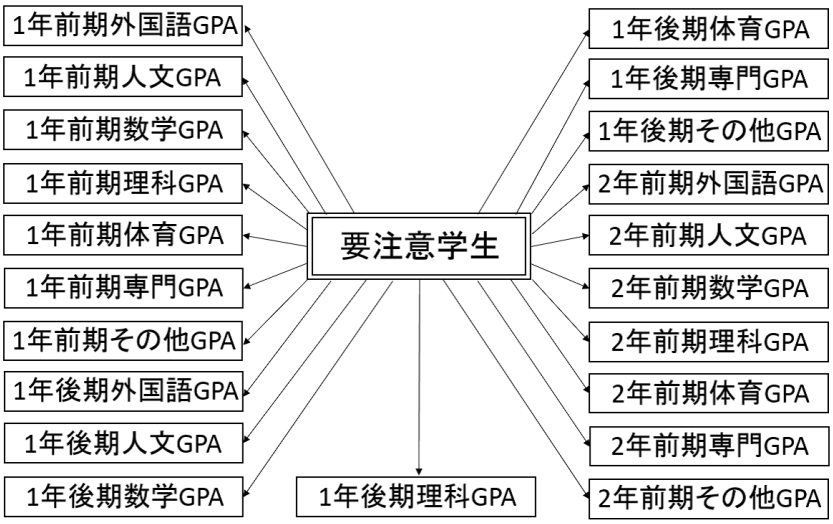

図 4.6: 2 年前期までの科目別 GPA を用いて構築されたグラフ

2 年後期までのモデル

表 4.8: 2 年後期までの科目別 GPA のみを用いたモデルの精度一覧

閾値	正解率			再現率			適合率			F 値
	対象	的中		対象	的中		対象	的中		
50 %	302	264	87 %	25	20	80 %	53	20	38 %	0.513
30 %	302	260	86 %	25	20	80 %	57	20	35 %	0.488
8.3 %	302	246	81 %	25	21	84 %	73	21	29 %	0.429

最後に 2 年後期の科目別 GPA データを加え，全 28 変数の科目別 GPA 値のみによる予測モデル結果である．最も F 値が高くなったのは，閾値 50 %の時で，予測すべき要注意学生 25 名中，20 名を予測でき，2 年後期までの科目別 GPA を用いることで 80 %の要注意学生を発見できていることがわかる．また，適合率の値から予測対象の 302 名中，53 名を指導対象者とすることで，予測すべき要注意学生の 80 %を発見できていることがわかる．

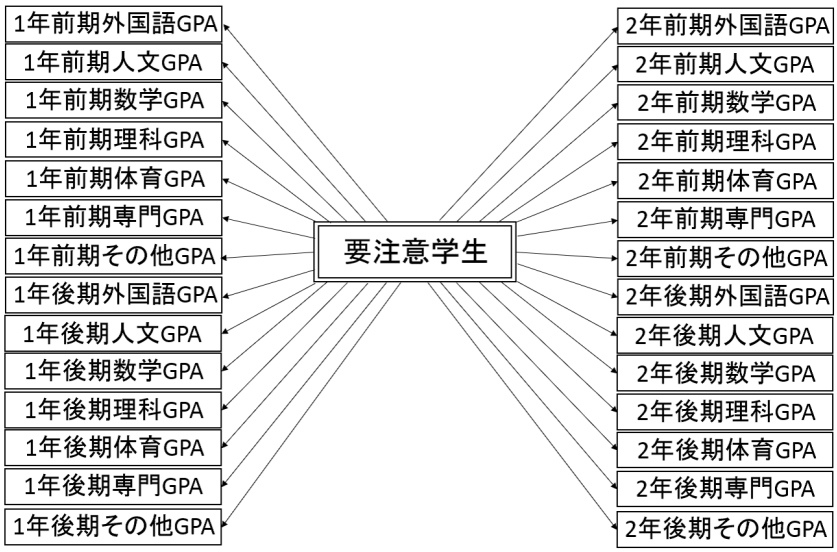

図 4.7: 2 年後期までの科目別 GPA を用いて構築されたグラフ

4.3.2　半期ごとのデータのみを利用するモデル

図 4.3 に示したように，半期ごとのデータのみを利用するモデルで，説明変数として科目別 GPA のみを利用したモデルである．すべてで 4 つのモデルが構築されるが，1 年前期の予測モデルは表 4.5 と一致するため省略している．構築されたモデルの精度一覧が以下の表 4.9 から表 4.11 である．グラフ構造はすべて Naive bayes 構造である．

1 年後期のみのモデル

表 4.9: 1 年後期のみの科目別 GPA のみを用いたモデルの精度一覧

閾値	正解率			再現率			適合率			F 値
	対象	的中		対象	的中		対象	的中		
50 %	302	265	88 %	25	10	40 %	32	10	31 %	0.351
30 %	302	258	85 %	25	16	64 %	51	16	31 %	0.421
8.3 %	302	221	73 %	25	17	68 %	90	17	19 %	0.296

1 年次後期のみの変数の為，予測に用いる科目別 GPA の説明変数の数は 7 変数のみである．最も F 値が高くなったのは閾値 30 %の時で，予測すべき要注意学生 25 名中，16 名を予測でき，1 年次後期のみの科目別 GPA により 64 %の学生を発見できていることがわかる．

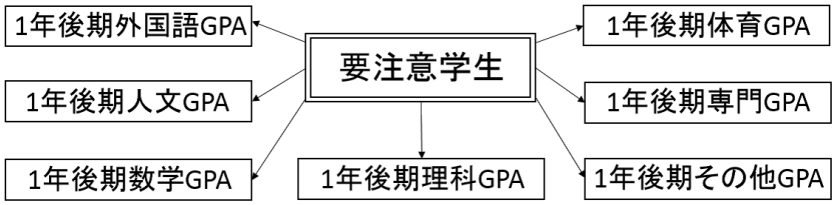

図 4.8: 1 年後期のみの科目別 GPA を用いて構築されたグラフ

2 年前期のみのモデル

表 4.10: 2 年前期のみの科目別 GPA のみを用いたモデルの精度一覧

閾値	正解率			再現率			適合率			F 値
	対象	的中		対象	的中		対象	的中		
50 %	302	271	90 %	25	14	56 %	34	14	41 %	0.475
30 %	302	260	86 %	25	14	56 %	45	14	31 %	0.400
8.3 %	302	228	75 %	25	17	68 %	83	17	20 %	0.315

同じく 2 年次前期のみの変数の為，予測に用いる科目別 GPA の説明変数の数は 7 変数のみである．最も F 値が高くなったのは閾値 50 %の時で，予測すべき要注意学生 25 名中，14 名を予測でき，2 年次前期のみの科目別 GPA により 56 %の学生を発見できていることがわかる．

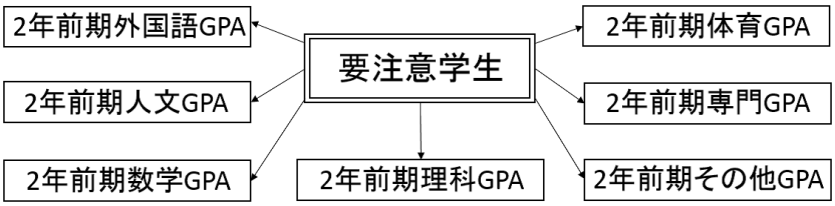

図 4.9: 2 前期のみの科目別 GPA を用いて構築されたグラフ

2 年後期のみのモデル

最後に 2 年次後期のみの科目別 GPA データを用いたモデルの精度結果である．説明変数の数は変わらず 7 変数のみである．最も F 値が高くなったのは閾値 50 %の時で，予測すべき学生 25 名中 17 名を予測でき，2 年後期のみの科目別 GPA により 68 %の学生を発見できていることがわかる．

表 4.11: 2 年後期のみの科目別 GPA のみを用いたモデルの精度一覧

閾値	正解率			再現率			適合率			F 値
	対象	的中		対象	的中		対象	的中		
50 %	302	280	93 %	25	17	68 %	31	17	55 %	0.607
30 %	302	275	91 %	25	19	76 %	40	19	48 %	0.585
8.3 %	302	250	83 %	25	21	84 %	69	21	30 %	0.447

図 4.10: 2 年後期のみの科目別 GPA を用いて構築されたグラフ

4.3.3　半期ごとのデータと前回の予測結果を利用するモデル

半期ごとのデータのみを利用するモデルに，新たな説明変数として，前回の予測結果を利用するモデルである．この変数により，半期のデータに，全てのデータを利用した場合とは異なる形でそれまでの時期の情報を加えることができる．すべてで 4 つのモデルが構築されるが，1 年前期の予測モデルは表 4.5 と一致するため省略している．

1 年後期のみと前回の予測結果を利用するモデル

表 4.12: 1 年後期のみの科目別 GPA と前回の予想結果を用いたモデルの精度一覧

閾値	正解率			再現率			適合率			F 値
	対象	的中		対象	的中		対象	的中		
50 %	302	263	87 %	25	11	44 %	36	11	31 %	0.361
30 %	302	256	85 %	25	14	56 %	49	14	29 %	0.378
8.3 %	302	222	74 %	25	18	72 %	91	18	20 %	0.310

最も F 値が高くなったのは，閾値 30 %の時で，F 値は 0.378 であった．しかし，1 年後期のみのモデルで最も良かった F 値は 0.421 であり，精度の低下が見られる．これは新たに追加した変数である，前回の予想結果がそもそもそれほど精度が良くなかったため，新たに変数として加えても予測精度の向上につながらなかったと考えられる．

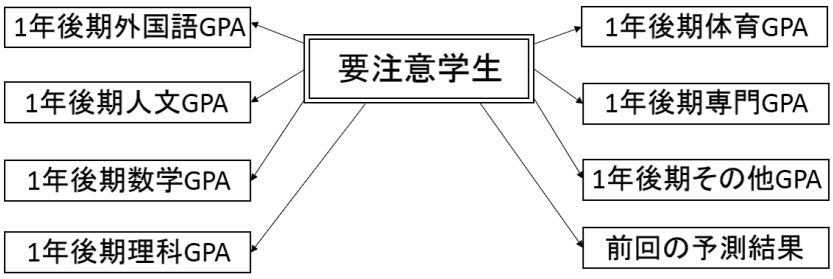

図 4.11: 1 年後期のみの科目別 GPA と前回の予測結果を用いて構築されたグラフ

2 年前期のみと前回の予測結果を利用するモデル

表 4.13: 2 年前期のみの科目別 GPA と前回の予想結果を用いたモデルの精度一覧

閾値	正解率			再現率			適合率			F 値
	対象	的中		対象	的中		対象	的中		
50 %	302	269	89 %	25	12	48 %	32	12	38 %	0.421
30 %	302	264	87 %	25	16	64 %	45	16	36 %	0.457
8.3 %	302	227	75 %	25	17	68 %	84	17	20 %	0.312

最も F 値が高くなったのは，閾値 30 %の時で，F 値は 0.457 であった．しかし，前期のみで新たに前回の予測結果を加える前のモデルで最もよい F 値は 0.475 であり，今回の場合も精度の低下が見られた．2 年前期においても，前回の予測精度がそれほど良くなかったため，変数として加えることで予測精度の向上につながることはなかった．

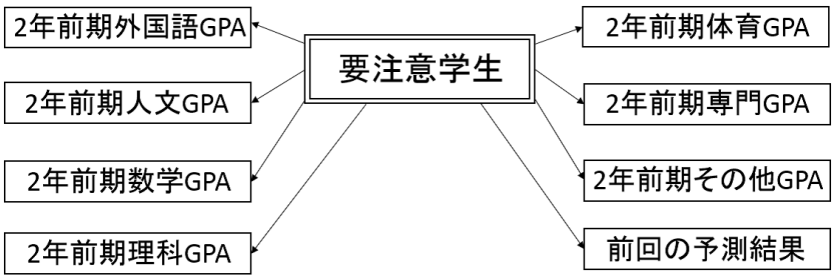

図 4.12: 2 年前期のみの科目別 GPA と前回の予測結果を用いて構築されたグラフ

2 年後期のみと前回の予測結果を利用するモデル

最も F 値が高くなったのは，閾値を 30 %の時で，その値は 0.537 であった．しかし，前回の予測結果を加える前のモデルで最もよい F 値は 0.607 であり，今回の場合も精度の低下が見られた．科目別 GPA のみで半期ごとのデータのみでモデルを構築する場合，前回の予測結果を利用することは精度の向上につながらないことが確認できた．

表 4.14: 2 年後期のみの科目別 GPA と前回の予想結果を用いたモデルの精度一覧

閾値	正解率			再現率			適合率			F 値
	対象	的中		対象	的中		対象	的中		
50 %	302	274	91 %	25	15	60 %	33	15	45 %	0.517
30 %	302	271	90 %	25	18	72 %	42	18	43 %	0.537
8.3 %	302	251	83 %	25	24	96 %	74	24	32 %	0.485

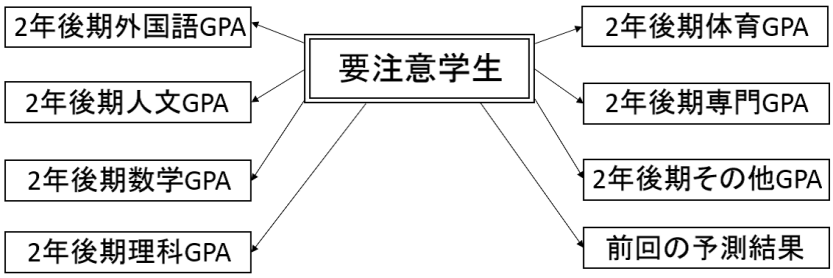

図 4.13: 2 年後期のみの科目別 GPA と前回の予測結果を用いて構築されたグラフ

4.3.4　科目別 GPA のみを利用したモデルの精度結果まとめ

科目別 GPA データのみを用いた予測モデルについて，半期ごとで最も高い F 値となったデータの利用範囲及び事後確率の閾値を表 4.15 にまとめる．

表 4.15: 科目別 GPA のみで各予測時期の最も F 値の高いモデルまとめ

予測時期	利用範囲	閾値	発見数	F 値
1 年前期	のみ	30 %	9 名	0.300
1 年後期	のみ	30 %	16 名	0.421
2 年前期	のみ	50 %	14 名	0.475
2 年後期	のみ	50 %	17 名	0.607

表 4.15 より，予測時期が遅ければ遅いほどより正確な予測ができていることがわかる．また，データの利用範囲について，どの予測時期を見ても，予測時期における半期のみのデータを利用したモデルのほうが高い精度であったことがわかる．このような結果になった考察として，変数の数の違いが考えられる．利用範囲を半期のみとした場合，どの時期においても 7 変数のみであったが，予測時期までのすべてのデータとした場合は，半期ごとに 7 変数ごと増えていき，変数の多さがモデル構築において，精度の高いベイジアンネットワークを構築する際の妨げになっている可能性があると考えられる．そのため変数が増えるタイミングである，1 年後期，2 年前期，2 年後期のそれぞれの予測時期において属性選択を行い，説明変数の削減を行った．属性選択には CFS を利用し，抽出されたそれぞれの時期の変数を表 4.16 に示す．

1 年後期までの 14 変数からは 8 変数が，2 年前期までの 21 変数からは 9 変数が，2 年後期までの 28 変数からは 11 変数が属性選択により抽出された．これらの抽出された変数を用いて構築したモデルの精度一覧を表 4.17 から表 4.19 に示す．

表 4.16: 各予測時期に属性選択により抽出された変数群一覧

1 年後期	2 年前期	2 年後期
1 年外国語前期	1 年外国語後期	1 年外国語後期
1 年外国語後期	1 年人文後期	1 年人文後期
1 年人文後期	1 年理科前期	2 年外国語後期
1 年体育後期	1 年専門後期	2 年人文前期
1 年理科前期	2 年外国語前期	2 年人文後期
1 年理科後期	2 年人文前期	2 年数学前期
1 年専門前期	2 年数学前期	2 年数学後期
1 年専門後期	2 年理科前期	2 年理科前期
	2 年専門前期	2 年理科後期
		2 年専門前期
		2 年専門後期
8 変数	9 変数	11 変数

1 年後期までの属性選択を行ったモデル

表 4.17: 1 年後期までの科目別 GPA で属性選択を行ったモデルの精度一覧

閾値	正解率			再現率			適合率			F 値
	対象	的中		対象	的中		対象	的中		
50 %	302	273	90 %	25	14	56 %	32	14	44 %	0.491
30 %	302	257	85 %	25	17	68 %	54	17	31 %	0.430
8.3 %	302	233	77 %	25	20	80 %	84	20	24 %	0.367

1 年後期までの，属性選択された科目別 GPA データ 8 変数を用いて構築されたモデルで最も F 値が高くなったのは，閾値 50 %ときで F 値 0.491 を示した．属性選択を行わずモデル構築をした場合の最も良い F 値は 0.378 であったため，属性選択による予測精度の向上がはっきりとわかる結果となった．

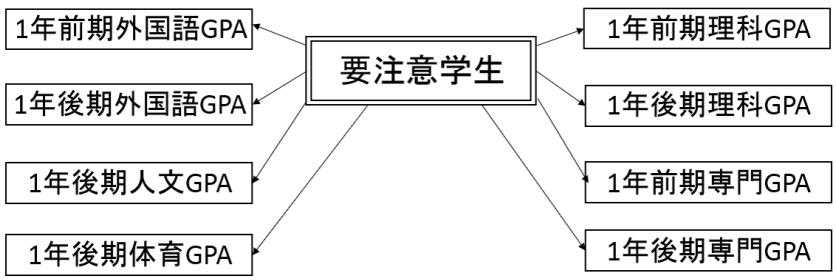

図 4.14: 1 年後期までの属性選択された科目別 GPA を用いて構築されたグラフ

2 年前期までの属性選択を行ったモデル

表 4.18: 2 年前期までの科目別 GPA で属性選択を行ったモデルの精度一覧

閾値	正解率			再現率			適合率			F 値
	対象	的中		対象	的中		対象	的中		
50 %	302	270	89 %	25	14	56 %	35	14	40 %	0.467
30 %	302	259	86 %	25	15	60 %	48	15	31 %	0.411
8.3 %	302	241	80 %	25	20	80 %	76	20	26 %	0.396

2 年前期までの，属性選択された科目別データ 9 変数を用いて構築されたモデルで最も F 値が高くなったのは，閾値 50 %の時で F 値は 0.467 を示した．属性選択を行わずモデル構築をした場合，最も良い F 値は 0.405 であったため，2 年前期までの場合においても属性選択の効果がはっきりと見られた．

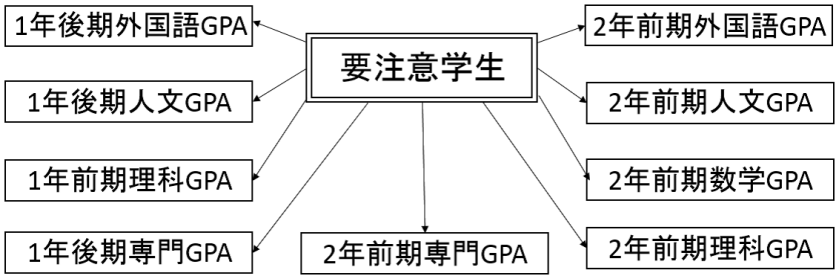

図 4.15: 2 年前期までの属性選択された科目別 GPA を用いて構築されたグラフ

2 年後期までの属性選択を行ったモデル

表 4.19: 2 年後期までの科目別 GPA で属性選択を行ったモデルの精度一覧

閾値	正解率			再現率			適合率			F 値
	対象	的中		対象	的中		対象	的中		
50 %	302	272	90 %	25	19	76 %	43	19	44 %	0.559
30 %	302	267	88 %	25	21	84 %	52	21	40 %	0.545
8.3 %	302	249	82 %	25	22	88 %	72	22	31 %	0.454

2 年後期までの，属性選択された科目別データ 11 変数を用いて構築されたモデルで最も F 値が高くなったのは，閾値 50 %の時で F 値は 0.559 を示した．属性選択を行わずモデル構築をした場合，最も良い F 値は 0.513 であったため，2 年後期までの場合においても属性選択の効果がはっきりと見られた．

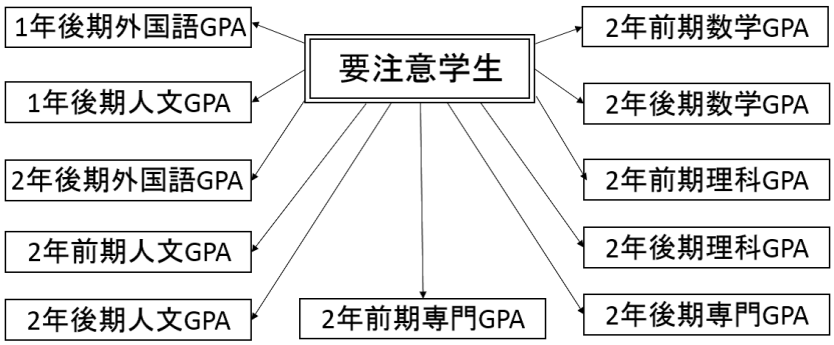

図 4.16: 2 年後期までの属性選択された科目別 GPA を用いて構築されたグラフ

属性選択によりどのような時期においても，変数削減の効果が予測精度向上につながる結果となった．最後に表 4.20 に属性選択を行った場合も含めた，各予測時期において最も高い F 値を示した手法についてまとめた．

表 4.20: 各予測時期における科目別 GPA のみで最も F 値の高くなった手法のまとめ

予測時期	利用範囲	閾値	発見数	F 値
1 年前期	のみ	30 %	9 名	0.300
1 年後期	までの属性選択	50 %	14 名	0.491
2 年前期	のみ	50 %	14 名	0.475
2 年後期	のみ	50 %	17 名	0.607

属性選択を行ったことにより，1 年後期の予測時期において，予測精度の向上が見られた．しかし，それ以外の時期に関しては半期のみのデータを利用した方が精度が良くなっていることが確認でき，必ずしも多くのデータから属性選択を行うほうが良くなるとは限らないことがわかる．学生の成績の急な変化を半期ごとのみのデータを利用するほうが捉えることができるため，このような結果になったと考えられる．

4.4　3 種類のデータを利用した要注意学生の発見

前節では，用いるデータを科目別 GPA のみとして予測モデルを構築した．本節では科目別 GPA データに加え，獲得成績数データと打刻データを加えた 3 種類のデータから要注意学生を半期ごとに予測するモデルを構築行った．利用するデータ範囲として，科目別 GPA データのみのときと同様に，予測時期までのすべてのデータを利用するモデルと，予測時期における半期のみのデータを利用するモデル，半期のみのデータに前回の予測結果の変数を加えたモデルの 3 パターンで予測モデルの構築を行った．

4.4.1　予測時期までのすべてのデータを利用するモデル

図 4.2 に示したように，予測時期までのすべてのデータを利用するモデルで，説明変数として科目別 GPA，獲得成績数，打刻データの 3 種類を利用したモデルである．1 年前期から 2 年後期までの半期毎に 4 つの予測モデルの構築を行った．利用した 3 種類の説明変数の詳細は第 3 章の表 3.1，表 3.2 に示している．構築する有向グラフ構造はすべて Naive bayes 構造である．また 3 種のデータを利用するため科目別 GPA データのみでモデル構築を行う場合に比べ説明変数の数が増えており，その数は半期で科目別 GPA データが 7 変数，獲得成績数データが 6 変数，打刻数データが 5 変数の全 18 変数である．予測時期まですべてのデータを利用する為，その数は半期ごとに増えていく．そのため，3 種類のデータを利用して予測モデルを構築する場合，どの予測時期においても属性選択により変数削減を行い，抽出された変数を用いてモデルを構築することで，予測精度の向上をはかる．

表 4.21 は予測時期までのすべてのデータを利用するとき，それぞれの予測時期で抽出された変数の一覧である．

表 4.21: 3 種類のデータから各予測時期に属性選択により抽出された変数群一覧

1 年前期	1 年後期	2 年前期	2 年後期
1 年理科前期	1 年外国語後期	1 年理科後期	1 年理科後期
1 年専門前期	1 年人文後期	2 年外国語前期	2 年外国語前期
1 年前期不可の数	1 年理科後期	2 年人文前期	2 年外国語後期
1 年 7 月打刻数	1 年専門後期	2 年専門前期	2 年人文前期
	1 年前期不可の数	1 年前期不可の数	2 年人文後期
	1 年後期不可の数	1 年後期失格の数	2 年専門後期
	1 年後期失格の数	2 年前期不可の数	1 年前期不可の数
	1 年 7 月打刻数	2 年前期失格の数	1 年後期失格の数
	1 年 11 月打刻数	1 年 11 月打刻数	2 年前期不可の数
	1 年 12 月打刻数	2 年 7 月打刻数	2 年前期失格の数
	1 年 1 月打刻数		2 年後期良の数
			2 年後期失格の数
			1 年 11 月打刻数
			2 年 11 月打刻数
			2 年 12 月打刻数
			2 年 1 月打刻数
4 変数	11 変数	10 変数	16 変数

1 年前期の予測時期では，全 18 変数から 4 変数が抽出され，1 年後期では全 36 変数から 11 変数が，2 年前期では全 54 変数から 10 変数が，2 年次後期では全 72 変数から 16 変数が抽出された．これらの抽出された変数を用いて構築されたモデルの予測精度の一覧を次の表 4.22 から表 4.25 に示す．

1 年前期までのモデル

表 4.22: 1 年前期までの属性選択された 3 種のデータを用いたモデルの精度一覧

閾値	正解率			再現率			適合率			F 値
	対象	的中		対象	的中		対象	的中		
50 %	302	272	90 %	25	8	32 %	21	8	38 %	0.348
30 %	302	256	85 %	25	10	40 %	41	10	24 %	0.303
8.3 %	302	229	76 %	25	15	60 %	78	15	19 %	0.291

最も高い F 値を示したのは閾値を 50 %にしたときで，F 値は 0.348 であった．科目別 GPA のみのモデルの最も高い F 値は 0.300 であったため，3 種のデータを利用したことによる精度の向上が確認できる．

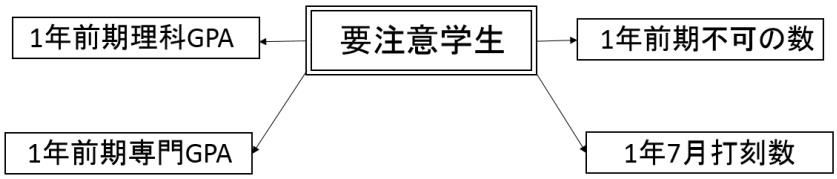

図 4.17: 1 年前期までの属性選択された 3 種のデータを用いて構築されたグラフ

1 年後期までのモデル

表 4.23: 1 年後期までの属性選択された 3 種のデータを用いたモデルの精度一覧

閾値	正解率			再現率			適合率			F 値
	対象	的中		対象	的中		対象	的中		
50 %	302	262	87 %	25	13	52 %	41	13	32 %	0.394
30 %	302	256	85 %	25	14	56 %	49	14	29 %	0.378
8.3 %	302	243	80 %	25	19	76 %	72	19	26 %	0.392

最も高い F 値を示したのは，閾値を 50 %にしたときで，1 年次後期までの科目別 GPA，獲得成績数，打刻データを用いることで予測すべき要注意学生 25 名中 13 名予測でき，52 %の要注意学生を発見することができる．

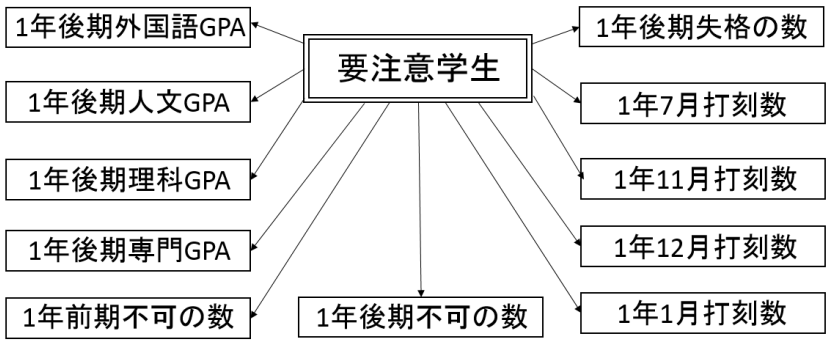

図 4.18: 1 年後期までの属性選択された 3 種のデータを用いて構築されたグラフ

2 年前期までのモデル

表 4.24: 2 年前期までの属性選択された 3 種のデータを用いたモデルの精度一覧

閾値	正解率			再現率			適合率			F 値
	対象	的中		対象	的中		対象	的中		
50 %	302	273	90 %	25	18	72 %	40	18	45 %	0.554
30 %	302	267	88 %	25	19	76 %	48	19	40 %	0.521
8.3 %	302	251	83 %	25	20	80 %	66	20	30 %	0.440

最も高い F 値を示したのは，閾値を 50 %にしたときで，2 年次前期までの科目別 GPA，獲得成績数，打刻データを用いることで予測すべき要注意学生 25 名中 18 名を予測でき，72 %の要注意学生を発見することができる．

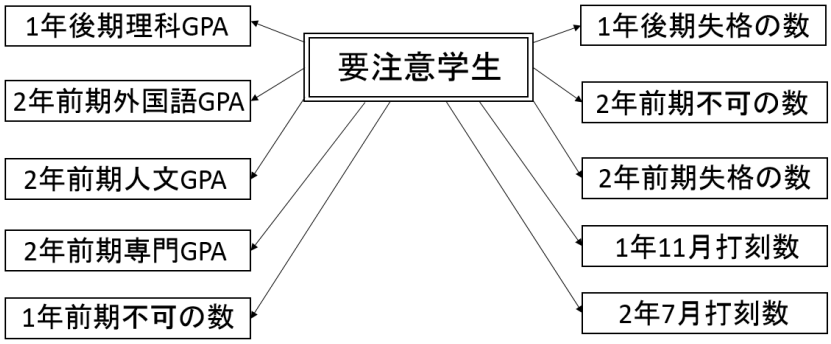

図 4.19: 2 年前期までの属性選択された 3 種のデータを用いて構築されたグラフ

2 年後期までのモデル

表 4.25: 2 年後期までの属性選択された 3 種のデータを用いたモデルの精度一覧

閾値	正解率			再現率			適合率			F 値
	対象	的中		対象	的中		対象	的中		
50 %	302	278	92 %	25	18	72 %	35	18	51 %	0.600
30 %	302	274	91 %	25	18	72 %	39	18	46 %	0.563
8.3 %	302	266	88 %	25	19	76 %	49	19	39 %	0.514

最も高い F 値を示したのは，閾値を 50 %にしたときで，2 年次後期までの科目別 GPA，獲得成績数，打刻データを用いることで予測すべき要注意学生 25 名中 18 名予測でき，72 %の要注意学生を発見することができる．

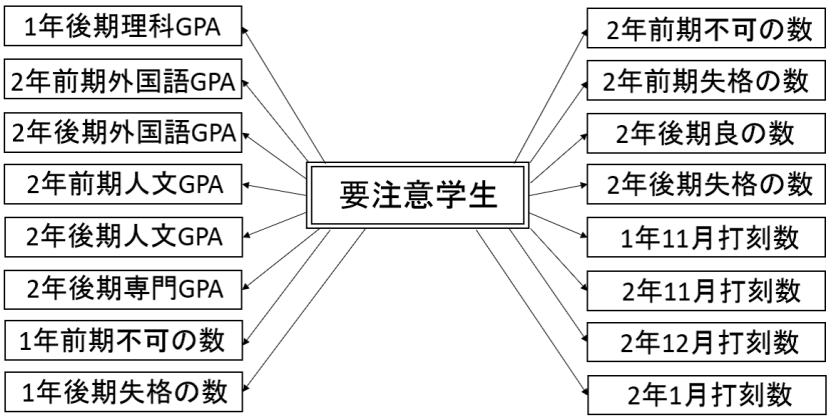

図 4.20: 2 年後期までの属性選択された 3 種のデータを用いて構築されたグラフ

4.4.2　半期ごとのデータのみを利用するモデル

図 4.3 に示したように，半期ごとのデータのみを利用するモデルで，説明変数として科目別 GPA に獲得成績数と打刻データを加えた 3 種類のデータを利用した構築するモデルである．1 年前期から半期ごとに 2 年後期までで 4 つのモデルが構築されるが，1 年前期の予測モデルは表 4.22 と一致するため省略する．また，それぞれの予測時期において，属性選択を行い変数の削減を行った．属性選択により抽出された変数を以下の表 4.26 に示す．なお，1 年前期の時期の属性選択は表 4.21 と一致するため，省略する．

1 年次後期のみの 18 変数からは 9 変数が，2 年次前期のみの 18 変数からは 6 変数が，2 年次後期のみの 18 変数からは 7 変数がそれぞれ抽出された．これらの抽出された変数を用いて構築されたモデルの予測精度の一覧を次の表 4.27 から表 4.29 に示す．

表 4.26: 3 種類のデータから各予測時期に属性選択により抽出された変数群一覧

1 年後期	2 年前期	2 年後期
1 年外国語後期	2 年外国語前期	2 年人文後期
1 年人文後期	2 年人文前期	2 年専門後期
1 年理科後期	2 年専門前期	2 年後期良の数
1 年専門後期	2 年前期不可の数	2 年後期失格の数
1 年後期不可の数	2 年前期失格の数	2 年 11 月打刻数
1 年後期失格の数	2 年 7 月打刻数	2 年 12 月打刻数
1 年 11 月打刻数		2 年 1 月打刻数
1 年 12 月打刻数		
1 年 1 月打刻数		
9 変数	6 変数	7 変数

1 年後期のみのモデル

表 4.27: 1 年後期のみの属性選択された 3 種のデータを用いたモデルの精度一覧

閾値	正解率			再現率			適合率			F 値
	対象	的中		対象	的中		対象	的中		
50 %	302	264	87 %	25	13	52 %	39	13	33 %	0.406
30 %	302	260	86 %	25	16	64 %	49	16	33 %	0.432
8.3 %	302	244	81 %	25	17	68 %	67	17	25 %	0.370

最も高い F 値を示したのは，閾値を 30 %にしたときで，1 年次後期のみの科目別 GPA，獲得成績数，打刻データを用いることで予測すべき要注意学生 25 名中 16 名予測でき，64 %の要注意学生を発見することができる．

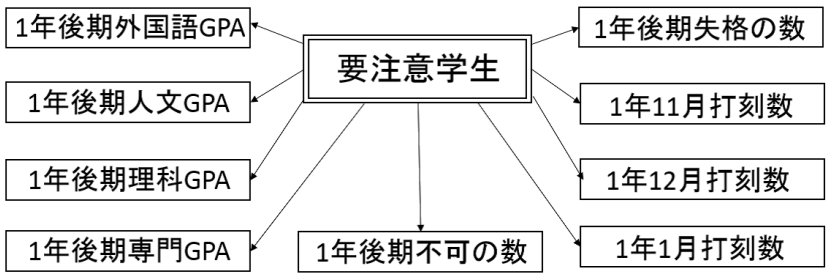

図 4.21: 1 年後期のみの属性選択された 3 種のデータを用いて構築されたグラフ

2 年前期のみのモデル

表 4.28: 2 年前期のみの属性選択された 3 種のデータを用いたモデルの精度一覧

閾値	正解率			再現率			適合率			F 値
	対象	的中		対象	的中		対象	的中		
50 %	302	272	90 %	25	14	56 %	33	14	42 %	0.483
30 %	302	265	88 %	25	17	68 %	46	17	37 %	0.479
8.3 %	302	255	84 %	25	19	76 %	60	19	32 %	0.447

最も高い F 値を示したのは，閾値を 50 %にしたときで，2 年前期のみの科目別 GPA，獲得成績数，打刻データを用いることで予測すべき要注意学生 25 名中 14 名予測でき，56 %の要注意学生を発見することができる．

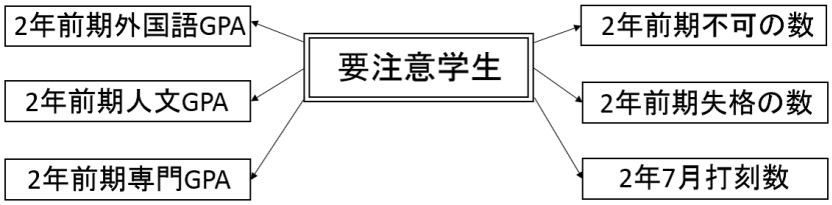

図 4.22: 2 年前期のみの属性選択された 3 種のデータを用いて構築されたグラフ

2 年後期のみのモデル

表 4.29: 2 年後期のみの属性選択された 3 種のデータを用いたモデルの精度一覧

閾値	正解率			再現率			適合率			F 値
	対象	的中		対象	的中		対象	的中		
50 %	302	277	92 %	25	15	60 %	30	15	50 %	0.545
30 %	302	277	92 %	25	16	64 %	32	16	50 %	0.561
8.3 %	302	259	86 %	25	19	76 %	56	19	34 %	0.469

最も高い F 値を示したのは，閾値を 30 %にしたときで，2 年次後期のみの科目別 GPA，獲得成績数，打刻データを用いることで予測すべき要注意学生 25 名中 16 名予測でき，64 %の要注意学生を発見することができる．

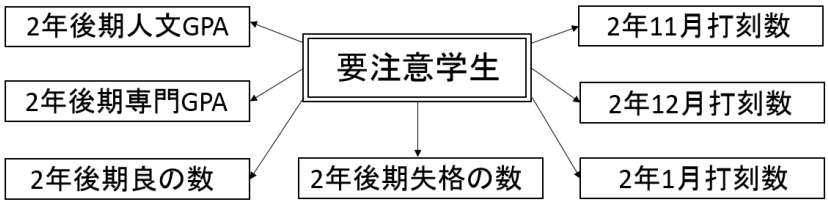

図 4.23: 2 年後期のみの属性選択された 3 種のデータを用いて構築されたグラフ

4.4.3　半期ごとのデータと前回の予測結果を利用するモデル

半期ごとのデータのみを利用するモデルに，新たな説明変数として，前回の予測結果を利用するモデルである．この変数により，半期のデータに，全てのデータを利用した場合とは異なる形でそれまでの時期の情報を加えることができる．すべてで 4 つのモデルが構築されるが，1 年前期の予測モデルは表 4.22 と一致するため省略している．

1 年後期のみと前回の予測結果を利用するモデル

表 4.30: 1 年後期のみの属性選択された 3 種のデータと前回の予測結果を用いたモデルの精度一覧

閾値	正解率			再現率			適合率			F 値
	対象	的中		対象	的中		対象	的中		
50 %	302	262	87 %	25	12	48 %	39	12	31 %	0.375
30 %	302	256	85 %	25	14	56 %	49	14	29 %	0.378
8.3 %	302	248	82 %	25	18	72 %	65	18	28 %	0.400

最も F 値が高くなったのは，閾値 8.3 %の時で F 値は 0.400 であった．前回の予測結果の変数を加えない 1 年後期のデータのみを利用したモデルで最も良かった F 値は 0.432 であったため，前回の予測結果は精度の向上につながらないことが確認できた．

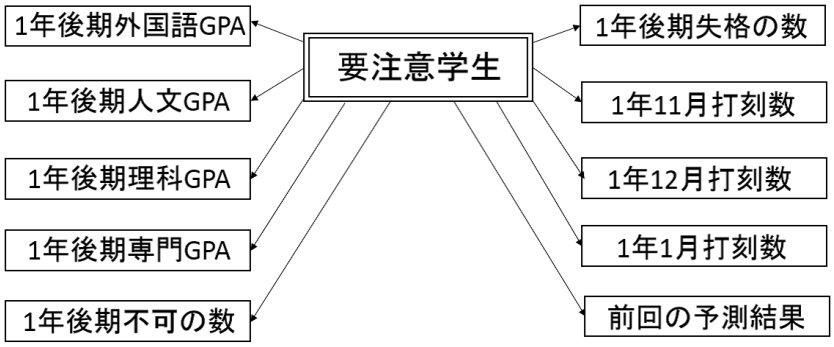

図 4.24: 1 年後期のみの属性選択された 3 種のデータと前回の予測結果を用いて構築されたグラフ

2 年前期のみと前回の予測結果を利用するモデル

表 4.31: 2 年前期のみの属性選択された 3 種のデータと前回の予測結果を用いたモデルの精度一覧

閾値	正解率			再現率			適合率			F 値
	対象	的中		対象	的中		対象	的中		
50 %	302	274	91 %	25	17	68 %	37	17	46 %	0.548
30 %	302	272	90 %	25	18	72 %	41	18	44 %	0.545
8.3 %	302	251	83 %	25	18	72 %	62	18	29 %	0.414

最も F 値が高くなったのは，閾値 50 %の時で，F 値 0.548 であった．前回の予測結果を変数として加えない 2 年前期のみのデータを利用したモデルで，最も良い F 値は 0.483 であり，前回の予測結果を変数に加える事で精度が向上することが確認できた．

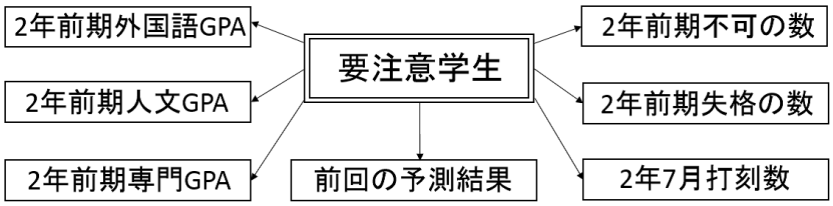

図 4.25: 2 年前期のみの属性選択された 3 種のデータと前回の予測結果を用いて構築されたグラフ

2 年後期のみと前回の予測結果を利用するモデル

表 4.32: 2 年後期のみの属性選択された 3 種のデータと前回の予測結果を用いたモデルの精度一覧

閾値	正解率			再現率			適合率			F 値
	対象	的中		対象	的中		対象	的中		
50 %	302	278	92 %	25	17	68 %	33	17	52 %	0.586
30 %	302	273	90 %	25	17	68 %	38	17	45 %	0.540
8.3 %	302	262	87 %	25	18	72 %	51	18	35 %	0.473

最も高い F 値となったのは，閾値 50 %の時で，F 値は 0.586 であった．前回の予測結果を変数として加える前のモデルで，最も良い F 値は 0.561 であり，前回の予測結果を変数として加えることで，予測精度の向上が確認できた．

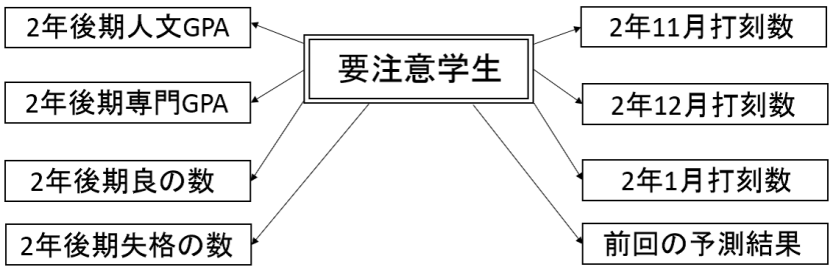

図 4.26: 2 年後期のみの属性選択された 3 種のデータと前回の予測結果を用いて構築されたグラフ

4.4.4　科目別 GPA，獲得成績数，打刻データを利用したモデルの精度結果まとめ

科目別 GPA，獲得成績数，打刻データの 3 種類のデータ説明変数として利用し，半期ごとに予測モデルを構築した場合，各予測時期において最も F 値が高くなる手法を表 4.33 にまとめる．

表 4.33: 科目別 GPA，獲得成績数，打刻データを利用した最も F 値の高いモデルまとめ

予測時期	利用範囲	閾値	発見数	F 値
1 年前期	のみの属性選択	50 %	8 名	0.348
1 年後期	のみの属性選択	30 %	16 名	0.432
2 年前期	までの属性選択	50 %	18 名	0.554
2 年後期	までの属性選択	50 %	18 名	0.600

データの利用範囲を見ると，予測時期が 1 年後期までは半期のみのデータを利用しほうが精度が高くなり，それ以降は全てのデータを利用したほうが精度が高くなることが確認できた．また，前回の予測結果は半期のみのデータを利用する場合には精度の向上が見られたが，全てのデータを利用する場合と比べると精度が劣ることが確認できた．

科目別 GPA のみを利用したモデルと科目別 GPA，獲得成績数，打刻データの 3 種のデータを利用したモデルの精度を比べると，F 値に関してはあまり差が見られなかったが，予測すべき学生の発見数は 3 種類のデータを用いたほうが多くなる結果であった．すなわち，要注意学生であるか否かを判断するにあたり，獲得成績数と打刻データも有効なデータであることがわかる．表 4.25 から，2 年後期の時点で科目別 GPA，獲得成績数，打刻データを利用し閾値 50 %で要注意学生であるか予測を行うことで，予測対象である 302 名中，35 名を修学指導対象とし，72 %の要注意学生を発見することができる．

4.5　要注意学生の発見時期の検証

本節では，予測対象 302 名の中の要注意学生 25 名が半期毎にどの時期で，本章で構築した予測モデルに要注意学生であると判断されるのか確認を行う．この確認により，前節の予測モデルの精度一覧では確認できなかった半期ごとの新規の要注意学生の発見数や，累計の要注意学生判定数を確認することができる．説明変数は科目別 GPA のみのモデルと科目別 GPA，獲得成績数，打刻データの 3 種類のデータを利用したモデル両方で確認を行う．また判定を行うモデルの閾値はどちらのモデルも 50 %とした．

4.5.1　科目別 GPA のみを利用するモデルの要注意判定確認

表 4.34，表 4.35 が半期ごとの科目別 GPA データのみを利用して構築された 4 つの予測モデルの要注意学生の判定詳細である．表左の数字は定義された予測すべき要注意学生 25 名の番号であり，表において○がついている部分が要注意学生であると予測された時期である．累計発見数が最も良かったモデルは，半期のみのデータで判定するモデルで，全てのデータを利用した判定よりも，半期のみの予測のほうが成績の変化の特徴を拾いやすく，広く要注意学生の発見ができていることが確認できた．しかし，2 年後期までの科目別 GPA データを用いても予測すべき要注意学生 25 名中，3 名が判定されていないこともわかる．

4.5.2　科目別 GPA，獲得成績数，打刻データを利用するモデルの要注意判定確認

表 4.36，表 4.37 が科目別 GPA，獲得成績数，打刻データを利用して構築された 3 つの予測モデルの要注意学生の判定詳細である．表左の数字は定義された予測すべき要注意学生 25 名の番号であり，表において○がついている部分が要注意学生であると予測された時期である．3 種のデータを利用した場合でも，半期のみのデータを用いて判定を行うほうが，累計の発見人数は多くなることが確認できた．しかし 3 種のデータを用いた場合でも，発見できた要注意学生の累積人数は 22 名で 3 名が発見できていないことが確認できた．

科目別 GPA データのみで発見できなかった学生番号は 4，7，17 であり，3 種のデータを用いて発見できなかった学生番号は 4，17，20 である．両方のモデルで判定を行えば 23 名までの要注意学生を発見することができるが，残り 2 名の学生はどちらのモデルでも発見することができない．この 2 名を発見できるような特徴が見られる変数を加える事で，さらなる精度向上につながることが考えられる．また，一つの予測モデルを利用し，要注意学生の予測を行った場合，出力は要注意学生と予測されるかされないかの二値的予測となるが，科目別 GPA のみで判定するモデルと 3 種類のデータで判定するモデルの両方で予測を行った場合，両方のモデルで要注意学生と判断されない場合，どちらか片方のモデルで要注意学生と判断される場合，両方のモデルで要注意学生であると判断される場合が考えられ，特に修学指導が必要な学生を予測することができる．

表 4.34: 科目別 GPA のみのモデルによる各時期の要注意判定詳細

	半期のみのデータで判定				半期のみ+前回の予測結果で判定		
	1 年前期	1 年後期	2 年前期	2 年後期	1 年後期	2 年前期	2 年後期
1				○			○
2			○	○		○	○
3			○	○			○
4							
5		○	○	○	○	○	○
6	○	○		○	○	○	○
7							
8			○	○			○
9		○	○		○	○	
10		○		○	○		○
11		○	○		○	○	
12			○				
13		○	○	○	○	○	○
14	○		○			○	
15				○			
16				○			○
17							
18				○			
19		○	○	○	○	○	○
20	○	○		○	○		○
21	○	○			○		
22	○		○	○	○	○	○
23			○	○		○	○
24			○	○		○	○
25		○	○	○	○	○	○
新規発見数	5	7	6	4	7	3	4
累計発見数	5	12	18	22	12	15	19
発見割合	20 %	48 %	72 %	88 %	48 %	60 %	76 %

表 4.35: 科目別 GPA のみのモデルによる各時期の要注意判定詳細

	全てのデータで判定				属性選択された全てのデータで判定		
	1 年前期	1 年後期	2 年前期	2 年後期	1 年後期	2 年前期	2 年後期
1				○			○
2			○	○			○
3			○	○		○	○
4							
5		○	○	○	○	○	○
6	○	○	○	○	○	○	○
7							
8				○		○	○
9		○	○	○	○	○	
10		○	○	○	○	○	○
11		○	○	○	○	○	○
12		○	○	○	○	○	○
13		○	○	○	○	○	○
14	○	○	○	○	○	○	○
15							○
16				○			
17				○			○
18							
19			○	○	○	○	○
20	○	○	○	○	○		○
21	○	○	○	○	○		
22	○	○	○	○	○	○	○
23							○
24		○	○	○	○	○	○
25		○	○	○	○	○	○
新規発見数	5	8	3	4	9	2	5
累計発見数	5	13	16	20	14	16	21
発見割合	20 %	52 %	64 %	80 %	56 %	64 %	84 %

表 4.36: 科目別 GPA，獲得成績数，打刻データのモデルによる各時期の要注意判定詳細

	半期のみのデータで判定				半期のみ+前回の予測結果で判定		
	1 年前期	1 年後期	2 年前期	2 年後期	1 年後期	2 年前期	2 年後期
1			○	○		○	○
2		○	○	○		○	○
3			○			○	○
4							
5		○	○	○	○	○	○
6	○	○	○	○	○	○	○
7		○			○		
8			○	○		○	○
9		○	○	○	○	○	○
10		○			○	○	
11	○	○			○	○	
12		○		○	○	○	○
13	○	○	○	○	○	○	○
14	○						
15			○	○		○	○
16				○			○
17							
18				○			○
19	○	○	○	○	○	○	○
20							
21	○	○			○		
22	○	○	○	○	○	○	○
23			○	○		○	○
24			○			○	○
25	○	○	○	○	○	○	○
新規発見数	8	6	6	2	5	7	2
累計発見数	8	14	20	22	13	20	22
発見割合	32 %	56 %	68 %	88 %	52 %	80 %	88 %

表 4.37: 科目別 GPA，獲得成績数，打刻データのモデルによる各時期の要注意判定詳細

	属性選択された全てのデータで判定			
	1 年前期	1 年後期	2 年前期	2 年後期
1				○
2		○	○	○
3			○	○
4				
5		○	○	○
6	○	○	○	○
7		○		
8			○	○
9		○	○	○
10		○	○	
11	○	○	○	
12		○	○	○
13	○	○	○	○
14	○		○	○
15			○	○
16				○
17				
18				
19	○	○	○	○
20				
21	○	○	○	○
22	○	○	○	○
23			○	○
24			○	○
25	○	○	○	○
新規発見数	8	6	5	2
累計発見数	8	14	19	21
発見割合	32 %	56 %	76 %	84 %

第5章　むすび

本研究では，名古屋工業大学に在学していた学生 338 名の講義別成績データと打刻データから，未来予測の手法としてベイジアンネットワークを利用し，半期毎に予測を行うことで，将来学業不振に陥るであろう学生の発見精度向上及び発見精度に関する検証を行った．2 章では本研究で利用したベイジアンネットワークや，属性選択などの理論を説明をした．3 章では予測に利用したデータに関して概要と拡張内容について説明をした．4 章では発見の対象者の定義を見直し，半期ごとに予測を行うことでの予測精度向上をはかった．

学業不振に陥る学生の特徴の一つに，急な成績低下が挙げられる．2 年次までのデータを利用し，さらに半期ごとのデータのみで予測を行うモデルを構築することにより，学生の成績のムラとも呼べる特徴を通年のデータを用いるよりも前後の時期の成績に関係しないため，確実に拾うことができ，より正確な予想が可能となった．具体的には，2 年次後期までのデータを用いれば，半期ごとに予測を行うことで，定義を見直した要注意学生の 9 割以上を発見できることが確認できた．また，科目別 GPA のみのモデルよりも獲得成績数と打刻データを加えたモデルのほうが発見される要注意学生が多く，学業不振に陥る学生の予測において，GPA データだけでなく，成績のムラを確認できる獲得成績数データや講義に対する姿勢が読み取れる打刻データに関しても有用性が確認できた．

獲得成績数データと打刻データを加えた場合に予測精度が向上したように，学生の特徴が見られる新たな変数が見つけられれば，さらなる予測精度の向上が可能になる．例えば，学生の受講態度を示す変数として打刻データを利用したが，出席数や遅刻数，欠席数をまとめたデータを利用することでより要注意学生の特徴を捉えることが可能になると考える．また，ベイジアンネットワークで予測モデルを構築する場合，利用する変数は予測精度に直結している．ベイジアンネットワークによるモデル構築において，連続値である変数を離散化する手法にウォード法によるクラスタリングを採用し，属性数を 4 に固定して離散化を行っていたが，利用する変数ごとに適切な属性数を設定し，属性数を固定ではなく可変にすることでさらなる精度の向上が期待できる．最後に，本研究では要注意学生の予測の精度に着目したが，より早い段階での要注意学生の特徴を発見し，予測の時期を早める事も修学指導を与えることを目的とした場合必要である．

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，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，情報基盤センター教職員の皆様に心から感謝致します.

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, 15-21 (2008)

[2] 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)

[3] 伊藤雄真,“ IC カード打刻データと修学データを用いた学生の将来の学習レベル予測と特徴分析 ”, 平成 24 年度名古屋工業大学卒業研究論文 (2012)

[4] 伊藤圭祐,“ 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)

[5] 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)

[6] 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)

[7] Duan, S. and Babu, S, “ Processing Forecasting Queries ”, *Proc. 2007 Intl, Conf. on Very Large Data Bases*, pp711-722 (2007)

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