平成25年度 修士論文

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

Study about Heuristics Anxious Student by Data Mining

指導教員
舟橋　健司　准教授
伊藤　宏隆　助教

名古屋工業大学大学院　工学研究科　情報工学専攻
平成24年度入学　24417505番

名前　伊藤　圭佑

目 次

第1章　はじめに

名古屋工業大学では2006年に「学びの場の構築」を目的としたプロジェクトを立ち上げ、学生に対する教育支援システムとしてICカード出欠管理システムと、Course Managemant Systemu（コースマネジメントシステム：以下CMSと記述する）を導入した[1]。ICカード出欠管理システムは、各教室に設置された端末に、各学生証に埋め込まれたICチップを認識させることで、学生の出席時間を学内に設置されたサーバに保持するシステムである。集積された時間情報は教員がWeb上で参照することができ、出欠の確定や学生の最終評価の指標などに活用されている。CMSは、情報技術やインターネットを使った教育指導を支援するシステムである。課題やレポートの提出管理、小テストの実施、学生の受講管理などをWeb上で管理することができる。これに伴い、学生の課題提出状況や小テストの結果及び評価が電子データで逐次集積されている。

現在では、学生に関する情報を電子データとして集積する事例は珍しいものではなくなっている。その背景には、Information and Communication Technology（情報通信技術：以下ICTと記述する）[3][4]の急速な発展が大きく関係している。ICTとは、コンピュータやネットワークに関連する技術・サービス・設備などの総称であり、先に述べたICカード出欠管理システムとCMSもこの一部である。ICTは様々な分野に進出・発展を遂げ、教育の分野では、e-Learningと呼称されるICTを活用した教育方法が多くの学校で採用されている。今日では8割以上の大学がe-Learningもといい ICTを取り入れている[2]。情報技術や通信技術が教育の場に活用されるようになってからは、副次的に大量のデータが産出されており、その一部には過去の各学生の成績、出席状況、学習履歴などの詳細な個人的情報も含まれている。情報の電子データ化とその大量集積は情報の保持性や参照スピードの向上に大きく寄与したが、近年ではそれだけでなく、データマイニングによって新たな知識や傾向を見つけ出し、修学環境を改善しようとする潮流が見られ始めている。データマイニングとは、大量のデータから機械的処理により要素間の関係性やパターンを見つけ出す行為及びその技術を意味し[5][6]、商業や医療などの分野では実用的に採用されている。

教育現場におけるデータマイニングの活用方法として、学生個人や学生全体に関する傾向を見つけ出し、その傾向をもとに今後の修学に関する指導を与えるという形態が提案されている。これに先立ち、修学傾向の分析に関係した多くの研究報告がなされている。例をいくつか挙げると、学生による授業アンケートをもとに学生個人の最終成績や学習状況の相互関係を調査したもの[7]や、講義の出席状況や課題提出状況からある学生の講義最終成績を予測したもの[8]、Web上で実施される学習の結果に応じた最適な教材の提供や学習ナビゲーションを目指したもの[9]などがあ

る。これらの研究では、目的や手法に差異があるが、既知の知識やデータから未知の傾向や事象を分析・予測するという共通点が存在する。例えば、「課題の提出が遅い学生は、成績が悪い可能性が高い」という確かな傾向が得られたとする。この場合、以後課題の提出が遅い学生に対し早期の修学指導を与えることで、学生の成績悪化の抑制が期待できる。このように、「〜ならば成績が悪くなる」と表現できる傾向やパターンを見つけ出せば、将来的に成績がかなり低迷してしまう学生や、学習の場から脱落してしまう学生（いわゆる”落ちこぼし”や消極的理由による退学者）を事前に発見できるため、学習環境を全体的に改善することができる。

しかし、先に指摘した研究報告では、修学環境や学生の分析までに留まったものや、実際どのように指導を与えるかに言及していないものが多い。せっかく獲得した知識や傾向も、修学指導にそのまま転用できるものであるとは言えない。どのように修学指導を与え、どのような学生を援助するかを具体的に考慮する必要がある。

修学指導の対象として挙げられるのが、近年の大学教育の場で目立ってきた消極的な理由による退学者の存在である。大学生が志半ばに退学する理由は、家庭の経済状況や学生自身の健康状況などの致し方ない場合や、転学などの積極的理由による場合が例として挙げられる。しかし中には大学生活に馴染めず退学してしまう学生や、能力的には難がないものの勤勉さに欠け最終的には学業からドロップアウトしてしまう学生などが多くの大学で指摘されている [10]。また、大学に籍を置きつつも学業にはほとんど参加していない学生や、就職や大学院入試の失敗による計画的な留年学生も少なからず存在している。これらの状況は名古屋工業大学も例に漏れず、具体的な数字は後に本稿で述べるが、無視できない人数の学生が学業から脱落していることが分かっている。その原因追究のため、本校でもデータマイニングの必要性が声高に唱えられているし、これは全国の大学でも要望のある事柄であるのは論を俟たない。

現在、多くの大学では、先述した学生の救援を目的として、教員と生徒による対話型修学指導を導入している [11]。その効果は既に指摘されており、学生と教員が面を向き合わせ学業や生活態度などについてアドバイスする指導方法が主流になりつつある。ただし、この指導方法には教員側の負担が大きくなるという問題が存在する。例えば、名古屋工業大学でも教員と学生の対話型修学指導を導入しているが、当大学では 1 人の教員に対し約 15 人の学生が在籍しており、1 人 1 人の学生の性格や環境を鑑みた指導を行うには時間的コストが多大になってしまう欠点がある。全学生に対する指導は指導者にとって相当な負担となる。また、何の判断材料もなしに指導を行うのは困難であり、場合によっては間違った指導を与えてしまう危険性も考えられる。修学環境や学生の分析だけでなく、修学指導を実用的にするための教員に対する支援も必要だと言える。

そこで我々は、『要注意学生』を定義し、その『要注意学生』を事前に発見する手法を提案している。『要注意学生』とは本研究にて我々が用いている造語であるが、これを言い換えると、『今後の修学において何かしらの懸念が予想され、事前の修学指導が必要であると考えられる学生』と表現できる。先述した、退学してしまう、あるいは、学業からドロップアウトしてしまう学生は

当然この『要注意学生』に該当する。つまり、修学環境の分析によって、今後指導を与える必要性がある学生（もしくは指導を与えることで修学状況の改善が見込める学生）の傾向を調査・分析する。これにより『要注意学生』になってしまう原因や兆候を知識として見つけ出す。さらに調査・分析結果を基にそのような学生を早期に発見する。そして、発見された『要注意学生』に対し然るべき修学指導を与える。これにより、指導対象となる学生を絞り込めるため、指導の時間的コストが削減できる。さらに選定された学生は分析によって定義された『要注意学生』であるから、指導の内容も決定しやすいというメリットもある。これらの問題点を解消することで、より実用的かつ効果的な修学指導が可能となる。

本研究では、まず、名古屋工業大学に在籍していた学生 338 名分の講義毎の成績（約 11 万レコード）、各教室の入退出時間（約 40 万レコード）や各学生の修学状況を元データとして、データ形式の拡張・変換・正規化・紐付けを行った。さらにそのデータを用いて統計的手法やデータマイニングによる修学環境の分析を行った。名古屋工業大学における学生の傾向を概観し、目新しい傾向や知識を発見するとともに、分析を進めていく過程で今後指導が必要だと考えられる『要注意学生』の定義のヒントを得た。そのヒントを基に『要注意学生』を定義し、指導対象となる学生を事前に発見する手法を検討した。本研究における要注意学生の発見は、「要注意学生であるかどうかの”予測”」とほぼ同じ意味があり、その予測手法にベイジアンネットワークによる手法を採用した。ベイジアンネットワークは、確率変数、有向グラフ構造、条件付き確率で定義される確率モデルであり、未来事象の予測に活用されている手法である。本研究では、ベイジアンネットワークによる『要注意学生』の発見の有用性を示すため、GPA のみを指標とした『要注意学生』の発見との比較を行った。これにより本研究の提案に相応しい予測手法及び予測モデルを提示した。

本稿の構成を説明する。2 章では、本研究で用いた統計的手法、データマイニング手法や予測手法の理論の概論を述べる。3 章では、本研究に用いるデータの形式やどのように拡張・変換・正規化・紐付けを行ったかを説明する。4 章では、3 章で述べたデータを用いて『要注意学生』を定義し、諸々の手法を用いて『要注意学生』を分析した。5 章では『要注意学生』を事前に見つけ出す手法の提案及びその検証を行った。そして 6 章では本研究のまとめを述べる。

ちなみに本研究では、学生のデータを扱うにおいて、個人を特定できる情報（指名や学籍番号）を一切排除した上で研究に着手しており、本文によって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　本研究に用いる手法の理論

本研究では、分析及び未来予測の手法を多く用いている。その多くはデータマイニングと呼ばれる知識発見の手法に分類されるものである。本研究では、修学状況の分析、『要注意学生』の発見において、各種手法を活用している。本節では属性選択、クラスタリング、ベイジアンネットワークについて説明する。

2.1　属性選択

属性選択 [12] は、複数ある特徴量のベクトルで記述されているデータを、全ての特徴量を利用せずに有用なものだけを取捨選択あるいは合成する行為を意味する。特徴選択、変数選択とも記される。機械学習によるパターン認識や、データマイニングによる分析や予測の際に多数の特徴量で形成されるデータを適用するが、無闇に全ての特徴量を用いても良い結果が得られない場合が多い。諸手法を用いて特徴量を限定あるいは合成することで、その結果を向上させることができる。

本節では、多数の変数で形成されたデータから新たな変数を作り出す主成分分析と、目的変数に大きく寄与している変数集合を抽出する指標である情報利得及び情報利得を用いた属性選択手法 Correlation based Feature Selection（CFS：以下 CFS と記述する）について説明する。

2.1.1　主成分分析

主成分分析（Principal Component Analysis：PCA とも表記される）は、多変数で表されるデータから、各変量によって新しい変数を作り出す手法である [13]。主成分分析を用いる目的は、情報の縮約であり、データを主要な変数に要約することで特徴の分析・把握を容易にする働きがある。「主要な変数」は主成分と呼称され、生成された各主成分に対し解釈を与えることで新たな意味を有した変数が生成される。以下より、主成分の数学的記述と、各主成分の計算方法、固有値の採択基準について述べる。

主成分の数学的記述

主成分分析は多変数で形成される複数の標本を持つデータを用いて行われる。ここで、以下の式 (2.1) で表される、変数が p 個、標本数 N、つまり各標本が $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \cdots\cdots, x_{ip})^\top$ と表さ

れるデータ X を仮定し、主成分を数学的に記述する。

$$X = \begin{pmatrix} x_{1,1} & \cdots & x_{1,N} \\ \vdots & \ddots & \vdots \\ x_{p,1} & \cdots & x_{p,N} \end{pmatrix} \tag{2.1}$$

主成分を z とすると、主成分 z は各標本 $\boldsymbol{x_i}$ を値とするベクトル変量 $\boldsymbol{x} = (x_1, x_2, \ldots\ldots, x_p)^\top$ を用いて表現される。実際に、以下の式 (2.2) を仮定する。

$$z = a_1 x_1 + a_2 x_2 + \cdots\cdots + a_p x_p \tag{2.2}$$

このとき、各係数をベクトルとした $\boldsymbol{a} = (a_1, a_2, \ldots\ldots, a_p)$ を $\sum_{i=1}^{p} a_i^2 = 1$ の条件下で、z の分散が最大となるように各値を変動させていく。最大の分散が得られたとき、この主成分を第 1 主成分とする。第 1 主成分が得られた際の主成分を z_1 というように記述すると、最大の分散が得られた際の係数ベクトル $\boldsymbol{a_1} = (a_{11}, a_{12}, \ldots\ldots, a_{1p})$ について、以下の式 (2.3) が成り立つ。

$$z_1 = a_{11} x_1 + a_{12} x_2 + \cdots\cdots + a_{1p} x_p \tag{2.3}$$

さらに、この z_1 とは無相関であり、z_1 の次に分散が最大となる係数ベクトル $\boldsymbol{a_2}$ を求める。そして求められた $\boldsymbol{a_2}$ によって作られる主成分 z_2 を第 2 主成分とする。同様に今までに得られた主成分とは無相関かつ分散が最大となる係数ベクトル $\boldsymbol{a_3}$ を探し出す。このような工程を繰り返すことで、主成分を多数生成する。一般的に、第 m 主成分 z_m は、係数ベクトル $\boldsymbol{a_m}$ を用いて以下の式 (2.4) のように表される。

$$z_m = a_{m1} x_1 + a_{m2} x_2 + \cdots\cdots + a_{mp} x_p = \boldsymbol{a_m x} \tag{2.4}$$

この m は変数の数 p 以下の自然数である。すなわち、主成分は元の変数の数 p 個まで作られる。しかし、大抵の場合は全ての主成分は用いられず、十分に全データを説明しうる分の主成分だけ採択され、不必要な主成分は無視される。これにより、主成分分析の主目的である情報の縮約が完了される。

主成分の計算方法

まずは 2 変数のデータを仮定した場合の主成分の計算方法を説明する。つまり各標本が $\boldsymbol{x_i} = (x_{1i}, x_{2i})^\top$ と表される場合を考える（i は標本番号）。このときのデータ行列 X は以下の式 (2.5) ように表される。

$$X = \begin{pmatrix} x_{1,1} & \cdots & x_{1,N} \\ x_{2,1} & \cdots & x_{2,N} \end{pmatrix} \tag{2.5}$$

まずは偏差関和行列 A を求め、さらに分散共分散行列 Σ を求める。各計算は以下の式 (2.6)、式 **??**eq:035) のようになる。

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \qquad \left(a_{ij} = \sum_{\lambda=1}^{N} (x_{i\lambda} - \bar{x_i})(x_{j\lambda} - \bar{x_j})\right) \tag{2.6}$$

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \qquad \left(\sigma_{ij} = \frac{1}{N-1} a_{ij}\right) \tag{2.7}$$

次に、第 1 主成分 z_1 を求める。z_1 は変数が 2 つであるから、係数ベクトル $\boldsymbol{a_1} = (a_{11}, a_{12})$、取る値が各標本 $\boldsymbol{x_i}$ となるベクトル変量 $\boldsymbol{x} = (x_1, x_2)^\top$ を用いて以下の式**??**eq:037) のように表される。

$$z_1 = a_{11}x_1 + a_{12}x_2 = \boldsymbol{a_1 x} \tag{2.8}$$

このとき、係数ベクトルは各値の 2 乗の和は 1 になるという制約があるため

$$\boldsymbol{a_1 a_1^\top} = 1 \tag{2.9}$$

が成り立つ。すなわち $\boldsymbol{a_1}$ は単位ベクトルになる。

式**??**eq:038) の条件の下で、z_1 の分散を最大とする $\boldsymbol{a_1}$ を決定する。ある変量 Z の分散を $V(Z)$ と表すとき、第 1 主成分 z_1 の分散は以下の式**??**eq:039) で求められる。

$$V(z_1) = V(\boldsymbol{a_1 x}) = \boldsymbol{a_1 \Sigma a_1^\top} \tag{2.10}$$

式**??**eq:038) の条件下で (2.10) の $V(z_1)$ を最大化する $\boldsymbol{a_1}$ を求めるには、ラグランジュの未定乗数 λ を用いた以下の式

$$\nu = \boldsymbol{a_1 \Sigma a_1^\top} - \lambda(\boldsymbol{a_1 a_1^\top} - 1) \tag{2.11}$$

の ν を最大化すればよい。式 (2.11) の両辺を $\boldsymbol{a_1}$ で微分して 0 とおくことにより

$$\frac{\delta \nu}{\delta \boldsymbol{a_1}} = 2\Sigma \boldsymbol{a_1^\top} - 2\lambda \boldsymbol{a_1^\top} = 0$$

$$\therefore (\Sigma - \lambda I)\boldsymbol{a_1^\top} = 0 \tag{2.12}$$

$$\therefore |\Sigma - \lambda I| = 0 \tag{2.13}$$

を得る。ただし I は 2 次の単位行列とする。この方程式を満たす $\boldsymbol{a_1^\top}$ を求めればよい。式 (2.13) は λ を固有値とした、行列 Σ の固有方程式であると見なすことができる。この固有値問題を解いて得られた σ に対応する固有ベクトルが係数ベクトル $\boldsymbol{a_i}$ に対応しているため、この固有値問題を解くことで、各主成分を決定することができる。

変数の個数が p になったとしても議論は同じものとなる。同様に式 (2.13) で表される固有値問題に帰着する。この際、λ の要素である $\lambda_1, \lambda_2, \ldots\ldots, \lambda_p$ を大きい順に並び換え、一番大きい固有値に対応する係数ベクトル $\boldsymbol{a_i}$ が第 1 主成分、その次に大きい固有値に対応する係数ベクトル $\boldsymbol{a_j}$ が第 2 主成分……といったように主成分の順番を決定する。

固有値の採択基準

主成分は元のデータの変数の数だけ生成されるが、全ての主成分を用いなくてもよい。ある程度に元のデータを説明できる分の主成分を採択すれば十分である。主成分分析において、用いる主成分の個数を決定する指標として、寄与率と累積寄与率が挙げられる。寄与率を直感的に説明すれば「ある主成分が全体のデータの何%を説明しているかを表すもの」だと言える。

固有値 λ_α の寄与率は、各変数の分散 σ_i を用いて以下の式 (2.14) で求められる。

$$C_\alpha = \frac{\lambda_\alpha}{\sum_{i=1}^{p} \sigma_i} \tag{2.14}$$

また、累積寄与率は寄与率の加算で計算される。例えば、第 M 主成分まで累積寄与率を計算するには、$C_1 + C_2 + \cdots\cdots + C_{M-1} + C_M$ を求めれば良い。全ての主成分の寄与率を加算すると、その値は 1 となることから、寄与率及び累積寄与率は%で表記される。

主成分を採択する基準として、「寄与率が何%以上であるものを採択する」「寄与率が大きい順に加算していって、累積寄与率が何%を上回る主成分の個数だけ採択する」といった基準が存在する。一般的には累積寄与率が 60 %〜80 %を上回った分だけを採択する。本研究では 70 %を基準として、主成分の採択を行っている。

2.1.2　情報利得と CFS

本研究では、元となるデータから多くの変数を定義している。分析や予測において、変数が多ければ多いほど良いというわけではなく、不必要な変数はノイズになることがある。これを避けるため、数ある変数の中から、必要な変数を取捨選択する必要がある。

属性選択をする際には、何かしらの指標を設けなければいけない。その指標の 1 つとして情報利得（Infomation Gain）が挙げられる。情報利得は、「2 つの確率分布の距離」とよく記述される (ただし距離の公理を満たしていないため、あくまで表現としての「距離」である)。

情報利得の定義は 2 つの確率分布 P と Q を用いて、以下の式 (2.15) で与えられる。

$$D(P||Q) = \sum_x P(x) log \frac{P(x)}{Q(x)} \tag{2.15}$$

$D(P||Q)$ が 2 つの確率分布 P と Q の情報利得である、と表すことができる。情報利得は分割前の平均情報量と分割後の平均情報量の差分でもあり、決定木の構築の際にも用いられている [14][15]。また、情報利得を用いた変数の選択は、属性数の多い変数が有利であるという特徴を持っている。ある変数の属性数を考慮して改良された情報利得比（Infomation Gain Ratio）も多く活用されている。情報利得比は情報利得を分割情報量で正規化されたものである。

また、情報利得を用いた変数選択指標として、CFS[16] が挙げられる。ある変数と関連性の高い変数を抜粋する際の指標として有効である。CFS は以下の式 (2.16) で与えられる。k は変数の個数、Z は目的変数を指す。この CFS を最大化するような変数 Y_i を抜粋する。ちなみに SU は情報量 H と情報利得 D で求めることができる。

$$CFS = \frac{\sum_{i=1}^{k} SU(Y_i, Z)}{\sqrt{k + \sum_{i=1}^{k} \sum_{j \neq i, j=1}^{k} SU(Y_i, Y_j)}} \tag{2.16}$$

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

2.2　クラスタリング

クラスタリング [17] は、対象間の類似度に基づき対象をグループ分けすることによって、グループの共通の性質が何であるかを検討しやすくするために使われる手法である。データマイニングに内包されている概念であり、元のデータ集合を特徴のある集団に分割するため、分析手法として用いられている。クラス（ラベル）が不要であるため、教師なし学習に分類される。

クラスタリングは主に階層的、非階層的に大別できる。階層的クラスタリングは似たもの同士を併合していくつかのグループにまとめていく手法であり、非階層は似た固体が同じグループになるように集合を分割していく手法である。本節では階層的の例として Ward's Method（ウォード法：以下よりウォード法と記す）、非階層的の例として K-means 法を解説する。

2.2.1　ウォード法

ウォード法は、各事例とクラスタの重心との距離の 2 乗値の和を用いた手法である。あるクラスタ G について、G に含まれる事例と重心との距離の 2 乗の和 $E(G)$ は、クラスタ G の重心 $M(G)$ を用いて、以下の式 (2.17) のように定義される。

$$E(G) = \sum_{x_i \in G} ||x_i - M(G)||^2 \tag{2.17}$$

そしてクラスタ G_i と G_j を併合したときの $E(G)$ の増分 $\delta E(G_i, G_j)$ と表記すると、以下の式 (2.18) のように表すことができる。

$$\delta E(G_i, G_j) = E(G_i \cup G_j) - E(G_i) - E(G_j) \tag{2.18}$$

この $\delta E(G_i, G_j)$ を基準（非類似度）として、2 つの集合を併合していく。以下より簡単なデータ列を用いて解説する。

10 ページ目の表 2.1 をデータ行列として、このデータのクラスタリングをウォード法により行う。最初はクラスタが存在せず全て単一の事例であるため、クラスタを事例の数だけ生成し $G_i = \{x_i\}$ としたとき、全てのクラスタにおいて $E(G_i) = 0$ である。つまり、各事例の距離がそのまま非類似度となる。このときの非類似度を表 2.2 に示す。この場合、x_1 と x_3 の非類似度が 1 であるため、この x_1 と x_3 を含んだ新たなクラスタ G_6 を生成する。

ここで新たに生成されたクラスタ G_6 の重心を計算する。$M(G_6) = (\frac{2+2}{2}, \frac{3+4}{2}) = (2, 3.5)$ である。次に G_6 と x_2, x_4, x_5 の各々との非類似度を算出する。例として、x_2 を要素とした G_2 と、G_6 の類似度 $\delta E(G_2, G_6)$ を求める。式は以下のようになる。

$$\delta E(G_2, G_6) = E(G_2 \cup G_6) - E(G_2) - E(G_6)$$

ここで各値、$E(G_2 \cup G_6), E(G_2), E(G_6)$ は以下のように求める。

$$\begin{aligned}
E(G_2 \cup G_6) &= \sum_{x_i \in G_2 \cup G_6} ||x_i - M(G_2 \cup G_6)||^2 \\
&= (2-3)^2 + (3-2.67)^2 + (5-3)^2 + (1-2.67)^2 + (2-3)^2 + (4-2.67)^2 \\
&= 10.7 \\
E(G_2) &= 0 \\
E(G_6) &= \sum_{x_i \in G_6} ||x_i - M(G_6)||^2 \\
&= (2-2)^2 + (3-3.5)^2 + (2-2)^2 + (4-3.5)^2 \\
&= 0.5
\end{aligned} \tag{2.19}$$

つまり、求める $\delta E(G_2, G_6)$ は $10.7 - 0 - 0.5 = 10.2$ となる。このようにして、クラスタと他事例及び他クラスタとの非類似度を算出する。次に完成する非類似度行列を表 2.3 に示す。ここで x_4 と x_5 の非類似度が表 2.3 にある中で最も値が小さいため、新たなクラスタ $G_7 = \{x_4, x_5\}$ を生成する。同様にクラスタ G_7 について、事例 x_2 とクラスタ G_6 の非類似度を算出する。以下の表 2.4 にその結果を示す。この結果より、新たなクラスタ $G_8 = G_6 \cup G_7$ を生成する。

このクラスタリングによって得られたデンドログラムと結果の可視化を図 2.1 と図 2.2 に示す。$\{x_1, x_3\}$ と $\{x_4, x_5\}$ と x_2 に分類されていることが分かる。

ウォード法は数あるクラスタリング手法の中でも汎用性が高い。分散最小法とも呼称され、数値型変数を有意に離散化する手法としても用いられている。

表 2.1: クラスタリング用データ列

事例	属性 1	属性 2
x_1	2	3
x_2	5	1
x_3	2	4
x_4	1	2
x_5	2	1

表 2.2: 1 回目の非類似度行列

	x_1	x_2	x_3	x_4	x_5
x_1	0.0	3.6	1.0	1.4	2.0
x_2	3.6	0.0	4.2	4.1	3.0
x_3	1.0	4.2	0.0	2.2	3.0
x_4	1.4	4.1	2.2	0.0	1.4
x_5	2.0	3.0	3.0	1.4	0.0

表 2.3: 2 回目の非類似度行列

	G_6	x_2	x_4	x_5
G_6	0.0	10.2	2.2	4.2
x_2	10.2	0.0	4.1	3.0
x_4	2.2	4.1	0.0	1.4
x_5	4.2	3.0	1.4	0.0

表 2.4: 3 回目の非類似度行列

	G_6	x_2	G_7
G_6	0.0	10.2	4.25
x_2	10.2	0.0	7.83
G_7	4.25	7.83	0.0

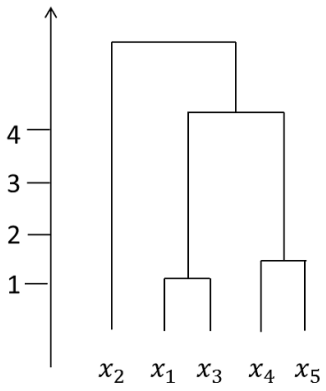

図 2.1: デンドログラム

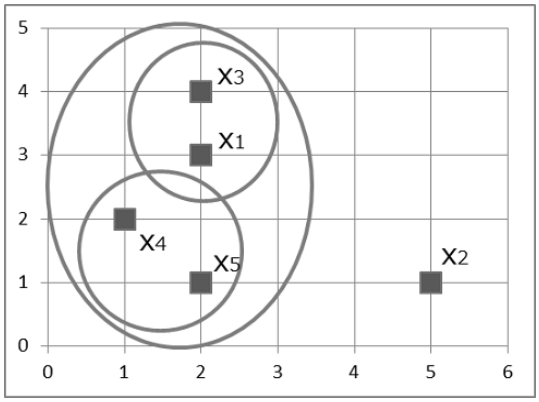

図 2.2: クラスタリング結果の可視化

2.2.2　K-means 法

K-means 法は、元のデータを似た事例同士が同じクラスタに属するように集合を分割する手法である。階層的クラスタリングと異なる点は、クラスタ数を予め定めておかなければならない点が挙げられる。

クラスタ数を k、データ（事例）の個数を n としたとき、K-means 法は次の手順で行われる。

1. k 個のシード $p^i(i = 1, 2, \ldots, k)$ を生成する。

2. 各データ $x^j(j = 1, 2, \ldots, n)$ に対し、最も近いシード p^i を求め、データ x^j をクラスタ C_i に入れる。

3. 各クラスタの重心を求め、その重心を新たなシード p^i とする。

4. 上記手順においてシード p^i が移動していた場合、手順 2 へ戻る。移動していなかった場合、クラスタリングを終了する

手順 2 において、ある事例と重心の距離を算出する。その距離の指標としてユークリッド距離が最も有名であり、本研究でも採用している。K-means 法の利点は、実装が容易かつ実行が早い点である。そのため、新規のサンプルを入力してもすぐに再計算できる。ゆえに多くの場面で活用されている。しかし、階層性はないため、クラスタリングの結果がクラスタ数や初期のシードに大きく影響を受けてしまうのが難点である。

2.3　ベイジアンネットワーク

ベイジアンネットワーク [18][19][20] とは、過去と現在の事象に確率的因果関係があると推定したグラフィカルモデルである。その特性を応用して、未来の事象の予測や分析の手法として多く用いられている。ベイジアンネットワークは、確率変数、変数間の有向グラフ、条件付き確率の 3 要素で定義される。その例として図 2.3 を示す。このモデルは確率変数 X, Y, Z, W、条件付き確率及び事前確率 $P(X), P(Z), P(Y|X, Z), P(W|Z)$、そして図 2.3 に示す有向グラフで定義されている。この 3 要素を決定することは、ベイジアンネットワークのモデルを生成することと同義である。

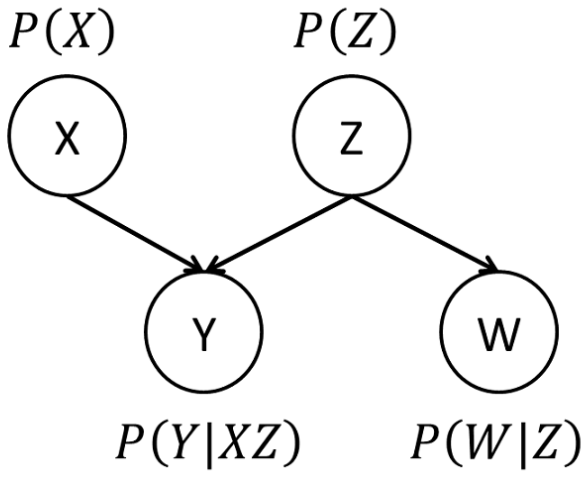

図 2.3: ベイジアンネットワークの一例

2.3.1　最適なベイジアンネットワークモデルの構築

ベイジアンネットワークは予測及び分析の手法として用いられるが、無闇にモデルを構築しても良好な結果を得ることはできない。上に述べた 3 要素を考慮し、最適なモデルを構築する必要がある。最適なモデルの構築は、最適な確率変数の選択、最適な有向グラフの獲得、最適な条件付き確率の獲得あるいは推論によって実現される。以下より、それぞれの手順や方法について述べる。

確率変数

ベイジアンネットワークに用いられる確率変数は、原則離散的でなければならない。つまり数値化変数を適用する場合は離散化を施す必要がある。ある属性が、ある企業に所属する社員の身長が記されたデータであったとする。身長は数値で記録されているため、そのままベイジアンネットワークの確率変数に用いることはできない。

離散化の手法としては、データの連続区間を等分割するか、クラスタリングによる分割を行う。前者は、事例が 100 個あったとしたら、33 個、33 個、34 個というように事例数が等しくなるように分割する。単純な方法ではあるものの、比較的良好な結果が得られることが多く報告されているため、この方法を採用する場合も少なくない。後者は、事例集合をクラスタリングによって分割することで、生成された集合 1 つ 1 つに有意性を付加することができる。特にウォード法がこの手法として活用されている。

有向グラフの構造

ベイジアンネットワークのグラフ構造はノードと変数の因果関係を表した矢印で形成される。一般に、矢印を指しているノードを親ノード、矢印に指されているノードを子ノードと呼称する。その構造によって、作られるモデルの名称とその特性が変わるため、モデル構築の際はグラフ構造を考慮する必要がある。ここでは代表的なグラフ構造について説明する。

(1) Naive Bayes　図 2.4 のように、ある 1 つの変数から、全ての他変数に矢印が伸びているグラフ構造を有したベイジアンネットワークを Naive Bayes と呼ぶ [21]。大抵の場合、予測対象の目的変数を親ノードとし、説明変数を子ノードとする。目的変数の事後確率はベイズの公式を用いて求められため、Naive Bayes の原理は他手法と比べ素朴だと言える。しかし、その素朴さに反して、得られる結果は良好である場合が多い。スパムメールの判別手法として有名な手法でもある。

(2) Tree Augmented Network　図 2.5 のように、Naive Bayes の子ノードから他の子ノードに 1 本だけ矢印を伸ばしたグラフ構造を、Tree Augmented Network（TAN：以下 TAN と記述す

る）と呼ぶ。ただし、ある部分グラフが有向閉路グラフとならないようにグラフ構造が決定される。グラフ構造の決定指標には、相互情報量が用いられる。

(3) Free Network　親ノード、子ノード数の制限がないグラフ構造を有したモデルを、Free Network と総称する。例えば、図 2.3 も Free Network と呼ばれるものの 1 つである。制限がないと雖も、無計画に複雑なグラフを構築したところで期待する精度が得られるとは限らない。また、あるノードに対する親ノードが増えるにつれ、必要となる条件付確率が爆発的に増え、条件付確率値に欠損が生まれる可能性もある。そのため、親ノードの個数などを制限した上で構造学習をする場合が多い。

本研究では、Naive Bayes 構造と Free Network 構造を採用している。

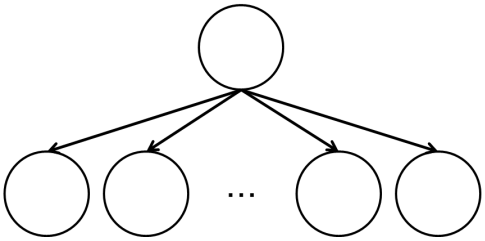

図 2.4: Naive Bayes 構造の一例

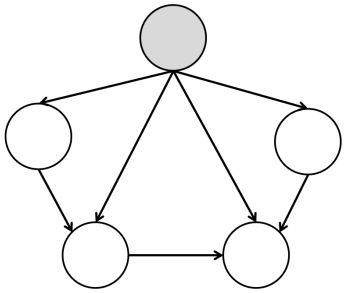

図 2.5: TAN 構造の一例

有向グラフの学習

有向グラフは、確率変数間の時間的因果関係を表している。図 2.3 の例では、確率変数 W は確率変数 Z の事象に関係性があると解釈できる。むろん、有向グラフを無闇に決定しても良いモデルは得られない。何かしらの手法で最適な有向グラフを見つけ出す必要がある。有向グラフ構造の決定には、いくつかのアルゴリズムが報告されている。本節では全探索アルゴリズムと K2 アルゴリズムについて概説する。

(1) 全探索アルゴリズム　全探索アルゴリズムは最も単純な方法である。全てのグラフ構造を想定し、最も良好な構造を採用する。当然最適な結果は得られるが、確率変数の数に応じて計算時間が指数関数的に増加していくため、現実的な方法ではない。確率変数が 3 つであれば、考えられる構造の数は 25 である。しかし、確率変数が 2 つ増えて 5 つとなっただけでも、考えられる構造の数は 29281 にも急増する。この問題は Non-deterministic Polynomial（NP）問題であり、探索方法に工夫が必要である。

(2) K2 アルゴリズム　K2 アルゴリズムは、上に述べた全探索アルゴリズムを計算時間面において改良したものである。ヒューリスティックを用いた手法であり、予め変数間の全順序関係を制約として設ける。これにより探索時間を大幅に減少することができる。

その他にも山登り法や、遺伝的ネットワークアルゴリズム、ニューラルネットを用いたグラフ構造の決定手法が存在する。本研究では主に K2 アルゴリズムを採用している。

条件付き確率の推定

各変数における事前確率及び条件付き確率を設定する必要がある。各変数間における条件付き確率値が格納された行列を Conditional Probability Table（CPT：以下 CPT と記す）と呼ぶ。条件付き確率は事前に明確である場合の方が少数であるため、大抵の場合は元のデータから推定・学習する。

データから条件付き確率を推定する方法を図 2.3 を用いて説明する。図 2.3 の確率変数 W の条件付き確率 $P(W|Z)$ について考える。ここで変数 Z と W の属性値をそれぞれ $\{1, -1\}$ の 2 値であると仮定する。このとき、各属性値の事例数が表 2.5 に示す通りであったして、条件付き確率 $P(W|Z)$ の推定を行う。

表 2.5: データの事例数表

		Z の属性値	
		1	-1
W の属性値	1	n_a	n_b
	-1	n_c	n_d

表 2.6: $P(W|Z)$ の CPT

		Z の属性値	
		1	-1
W の属性値	1	$\frac{n_a}{n_a+n_c}$	$\frac{n_b}{n_b+n_d}$
	-1	$\frac{n_c}{n_a+n_c}$	$\frac{n_d}{n_b+n_d}$

各事例数 $n_i(i = a, b, c, d)$ が十分に大きい場合は、最尤推定により $P(W = 1|Z = 1) = \frac{n_a}{n_a+n_c}$ といった具合で条件付き確率を定めることができる。この場合の CPT は図 2.6 の示す通りになるが、しかし、全事例数が十分に大きくない場合には、既存のデータの影響力が多大になり汎用性の欠いた条件付き確率が推定される問題が存在する。CPT のサイズが大きくなればなるほど、ある属性値を取る事例数が 0 になる可能性も大きくなる。この状態で最尤推定を行うと、条件付き確率が 0 となる場合が増加するため、汎化的なモデルとしては相応しくない。この解決法として、各事例数を無条件にいくつか加算する事が挙げられる。これにより、ある属性値を取る事例数がデータに存在しなくても、条件付き確率が 0 となることを回避することができる。また条件付き確率値の確率分布を考えることができ、観測値をパラメータとした Dirichlet 分布により条件付き確率を推定する場合もある。

また、データ自体に欠損を含む場合も考えられる。このデータは不完全データと呼称されるが、この場合は欠損を埋め合わせることで条件付き確率を推定する。その埋め合わせの方法として Bound and Collapse 法 [22] やニューラルネットを用いた手法などが挙げられる。

確率モデルの評価指標

最適なモデルを構築するには、確率モデルを評価する指標が重要となる。情報量基準と呼ばれる、統計モデルの良さを評価するための指標である。その例として 2 つの指標、赤池情報量基準（Akaike's Information Criterion：以下 AIC と記す）と最小記述長（Minimum Description Length：以下 MDL と記す）について以下より概説する。

(1) AIC　ベイジアンネットワークは、ほとんどの場合においてデータを基に構築される。このとき、構築に用いたデータにより適合した度合いは尤度と呼称される。尤度が大きければ大きいほど、より観測データに適していると評価できる。しかし、尤度の最大化を要望すると、同時にパラメータ数が増加してしまう傾向がある。パラメータ数が多すぎるとモデルの記述が複雑になってしまう。AIC はパラメータ数を一種のコストと見なしモデルを評価する指標である。あるモデルの尤度を L、パラメータ数を k とすると、指標値 AIC は以下の式 (2.20) のように計算できる。

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

AIC が小さいほど優れたモデルであると言える。尤度 L が大きいほど、もしくは、パラメータ数 k が小さいほど指標値 AIC は小さくなる。この指標を用いることで、モデル記述が簡潔であるが観測データに適合しているモデルを評価することができる。

(2) MDL　AIC は「より適合し、しかし短く」の思想から定められた指標であるが、データサイズが大きくなればなるほど複雑なモデルが有利になってしまう恐れを含んでいる。そこでデータサイズも一種のコストと見なした指標が MDL である。指標値 MDL はモデルの尤度を L、パラメータ数を k、データサイズ n を用いて以下の式 (2.21) のように表される。

$$MDL = -logL + \frac{k}{2}logn \tag{2.21}$$

AIC 同様に指標値が小さいほど優れたモデルであると言える。AIC と比較すると、データサイズが大きくなると記述の複雑なモデルは評価されなくなる特徴を持つ。

2.3.2　ベイジアンネットワークによる確率推論

ベイジアンネットワークのモデルは、グラフ構造のノードに相当する確率変数に入力（観測値）を与えることで、他の確率変数の事後確率を計算することができる。この一連の行為は確率推論と呼ばれ、事後確率値の変化を比較することで、事象の未来予測や分析を行うことができる。以下に簡単な例を交えて、確率推論の方法を説明する。

単純なモデル例を用いた確率推論

野球好きの 3 人家族を想定する。家族構成は父、母、そして息子である。父は野球チームであるジャイアンツの大ファンであり、対して母と息子はジャイアンツがあまり好きではない。父はジャイアンツが勝つと上機嫌になり、負けると機嫌が悪くなる。母はジャイアンツが負けると息子に電話してジャイアンツが敗北した旨を伝える。これらの事象をモデリングすると、図 2.6 のようになる。また条件付き確率を表 2.7、表 2.8、表 2.9 に示す。

まずはこのモデリングを用いて、事象の事前確率を求めることができる。例えば、前情報無しに父が不機嫌（$Father = Angry$）である確率を求めると、$P(F = Angry) = \sum_G P(G)P(F = Angry|G) = 0.41$ が得られる。つまり、10 日間のうち 4 日は不機嫌な日があると考えられる。こ

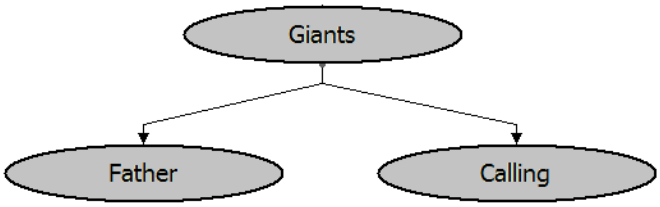

図 2.6: 野球好き一家の行動モデリング

表 2.7: Giants の事前確率 $P(G)$

G の属性値	Win	0.65
	Lost	0.35

表 2.8: Father の条件付き確率 $P(F|G)$

		G の属性値	
		Win	Lost
F の属性値	Happy	0.9	0.01
	Angry	0.1	0.99

表 2.9: Calling の条件付き確率 $P(C|G)$

		G の属性値	
		Win	Lost
C の属性値	Call	0.3	0.9
	No	0.7	0.1

のようにして、与えられた事前確率及び条件付き確率を用いてある事象の確率値を計算できる。

ここで、息子は野球チームジャイアンツの試合結果を知らない状況にあると仮定し、その状況下で母から電話がかかってきたとき、ジャイアンツの試合結果と父の機嫌はどうなっているか確率的に推論する。母から電話がかかってきた場合、息子は「属性値 $C = Call$」を観測したと言えるから、図 2.6 のノード $Calling\ (C)$ に観測値 $C = Call$ を入力する。さらに $P(G|C = Call)$ と $P(F|C = Call)$ を、確率の定義及びベイズの定理に則って計算する。それぞれは以下のように求めることができる。

$$\begin{aligned} P(G|C = Call) &= \frac{P(G, C = Call)}{P(C = Call)} \\ P(F|C = Call) &= \frac{P(F, C = Call)}{P(C = Call)} \end{aligned} \tag{2.22}$$

これらの式の分母にある $P(C = Call)$ は算出できないが、$\alpha = \frac{1}{P(C=Call)}$ として正規化定数扱いできる。つまり、分子にある $P(G \cap C = Call)$ 及び $P(F \cap C = Call)$ を計算し、後に全事象の確率値の和が 1 になるように正規化を行えばよい。また、$P(G \cap C = Call) = P(G) * P(C = Call|G)$ であるため、現時点で分かっている条件付き確率値と事前確率値を用いて計算することができる。例えば $P(G = Win|C = Call)$ は以下のように計算できる。

$$\begin{aligned} P(G = Win|C = Call) &= \frac{P(G = Win, C = Call)}{P(C = Call)} \\ &= \alpha P(G = Win, C = Call) \\ &= \alpha P(G = Win) * P(C = Call|G = Win) \\ &= \alpha * 0.65 * 0.3 = 0.195\alpha \end{aligned} \tag{2.23}$$

同様に $P(G = Lost|C = Call)$ を計算すると、0.315α が得られる。$P(G = Win|C = Call) + P(G = Lost|C = Call) = 1$ となるように正規化すると、$P(G = Win|C = Call) = 0.38$、$P(G = Lost|C = Call) = 0.62$ となる。何の情報も与えられなければ、表 2.7 の事前確率が表すようにジャイアンツの勝利が予測されるはずであったが、息子が「母からの電話」を観測したことでジャイアンツの敗北の確率が勝利の確率よりも高くなる結果となった。

これに続き、$P(F = Happy|C = Call)$ も求めたい。しかし、式 (2.22) にある $P(F \cap C = Call)$ を求める手立てが存在しない。そこで、モデル構造を考慮した結合確率 $P(F, G, C) = P(G)P(F|G)P(C|G)$

を用いて、より一般的に確率値の計算を行う。計算は以下のようになる。

$$\begin{aligned} P(F=Happy|C=Call) &= \frac{P(F=Happy, C=Call)}{P(C=Call)} \\ &= \frac{\sum_G P(G)P(F=Angry|G)P(C=Call|G)}{\sum_G \sum_F P(G)P(F|G)P(C=Call|G)} \\ &= \frac{\sum_G P(G)P(F=Angry|G)P(C=Call|G)}{\sum_G P(G)P(C=Call|G)\sum_F P(F|G)} \\ &= \frac{\sum_G P(G)P(F=Angry|G)P(C=Call|G)}{\sum_G P(G)P(C=Call|G)} \end{aligned}$$

式 (2.24) の 3 行目にある $\sum_F P(F|G)$ は 1 となることに留意したい。周辺化により厳密に確率値を計算することができる。この計算により、$P(F=Happy|C=Call)=0.35$、$P(F=Angry|C=Call)=0.65$ が得られる。当初の「父が不機嫌である確率」$P(F=Angry)$ は 0.41 であったが、母からの電話を観測すると、不機嫌である確率が増加しているのが分かる。一見して「母の電話」と「父の機嫌」は関係のない事象に思えるが、図 2.6 のようなモデルを構成することで、この 2 つの事象に確率的関係性があることが確認できる。

本例は単純なモデルであったが、より複雑なモデルを考えた場合、要素間の関係性を見つけ出すのは困難である。しかし、ベイジアンネットワークの確率推論を用いて、観測値の入力後の事後確率変動を検証することで、ある事象・要素間の関係性を発見することができる。このベイジアンネットワークの特性は、分析やある事象の未来予測などに活用されている。

確率推論の現状

上の節では原始的な確率推論とその効用を述べた。しかし、例で述べたような周辺化による確率値計算は、要素の増加に伴いその計算時間は爆発的に増加する。現象のモデリングはより複雑になることが考えられるため、その確率推論は現実的に不可能となってしまう。これに先立って、周辺化による計算ではなく、グラフ構造を考慮した確率伝播法による確率推論が推奨されている。確率伝播法は、あるノードに着目した時、親ノード群と子ノード群からの確率値変動だけを考慮することで計算時間を削減した手法である。この手法によりベイジアンネットワークの実用化が著しく進んでいる。

ただし確率伝播法にも問題点がある。図 2.3 のような、経路の方向性を無視した上で閉路が存在しないグラフ構造の場合は確率伝播法による確率推論は厳密であるが、そうでない場合、つまり閉路を含むグラフ構造を有したモデルにおいては厳密な確率推論ができない。この問題点を解消するために多くの研究報告がなされ、それらの中で最も有名なものが Junction Tree アルゴリズムである。Junction Tree アルゴリズムは閉路ありグラフを等価な閉路なしグラフに変換し確率推論を実行する手法である。要素の増加に伴い計算時間が増加してしまい現実的に運用できなくなってしまう課題もあるが、現時点では最も活用されている手法である。

第3章　本研究に用いるデータとその拡張及び変換

本章では、本研究に用いるデータの概要とその拡張及び変換について述べる。元のデータは各学生の講義別成績と入退室時間に関するものと、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ものがある。これらのデータを本研究の分析や要注意学生の発見に適用するために、データの拡張及び正規化を行った。

3.1　用いるデータの概要

本研究では、名古屋工業大学を卒業した 338 名の学生に関するデータを用いている。この 338 名は 2 年度分に相当し、171 名と 167 名に分けられる。データの種類は 3 つあり、講義別成績データ、入退室時間に関するデータ（以下打刻データと記述する）、そして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もの（以下学生修学データと記述する）である。

3.1.1　講義別成績データ

学生の講義別成績のデータは、レコード形式で保持されており、1 レコードが、学籍番号が暗号化された数字列、講義の成績、授業名、開講時期、これら 4 つの情報で構成されている。学籍番号は暗号化されているので個人を特定できないようになっている。

講義の成績は、教員が講義終了時に確定したもので、秀・優・良・可・不可・失格、これら 6 つの評価が存在する。秀が一番評価が高く、秀、優、良、可、不可になるにつれ評価が下がる。また、成績が秀・優・良・可であれば単位取得が認められ、不可・失格であれば認められない。不可と失格の差異は、評価可能かどうかにあり、各講義で実施される最終試験を受験しつつ単位取得条件を満たさなかった場合は成績が不可となり、最終試験を受けていなかった場合や講義への出席が所定の回数を満たしていない場合は、評価が不能であると捉えられ成績が失格となる。

記載されている授業名は実際のシラバスに載っている授業名ではなく、「専門 1」や「演習 1」のように講義を特定できないように改名されている。これは学生の学科を特定できないようにするための処置であり、基礎学習である英語、理系基礎科目、リベラルアーツに属する授業の名前は改名されていない。そのため、具体的な講義内容は分からなくとも、講義の科目は分かる状況にある。

3.1.2　打刻データ

打刻データは、1 章にて述べた IC カード出欠管理システムから産出されたデータである。学生の所持している IC チップが埋め込まれた学生証を各教室に設置された端末に反応させることで、教室の入室時刻と退出時刻が記録できる。本学では入退室時刻の記録を”打刻”と呼称しているため、このデータを”打刻データ”と呼んでいる。

講義別成績データ同様レコード形式で保持されており、1 レコードが、学籍番号が暗号化された数字列、打刻した日付（年/月/日）、打刻した時間、これら 3 つの情報で構成されている。暗号化された学籍番号は上で述べた講義別成績データのものと共通であるため、打刻状況と講義別成績との紐付けは容易である。また、打刻した日付から曜日やその日が休日であったかどうかも調べることができる。

3.1.3　学生修学データ

学生修学データは、本研究対象である 338 名の学生の卒業研究に着手した年次と卒業までに費やした年数が記録されているものである。本稿では便宜上”学生修学データ”と呼称している。名古屋工業大学では、1 年生から 3 年生までに各科目の講義を受講し、4 年生から研究室に所属し卒業研究に取り組む制度となっている。卒業研究に着手する条件は単位取得数によって取り決められており、既定の単位数に達しない場合は 4 年生になっても卒業研究に着手できない。また、通常は 4 年生修了時に卒業となるが、卒業条件も単位数によって条件付けされており、4 年次に卒業研究に着手できても、その年度に卒業できない学生が発生することもある。

3.2　データの拡張及び変換

講義別成績データと打刻データはレコード形式であり、総量は 50 万にも及ぶ。この形式のまま分析や本提案に適用するのは難しい。ゆえに、本研究ではこれらのデータの拡張及び変換を行った。その詳細を以下より述べる。

3.2.1　講義別成績データの拡張及び変換

本研究では学生個人の傾向により注目しており、元のデータの形式では適用が難しい。そこで、講義別成績データを学生個人の成績値と獲得成績数のデータに拡張した。

成績値の指標として、Grade Point Average（以下 GPA と記述する）[23] を採用した。この指標は、多くの大学教育機関で用いられているもので、当大学もその例外でない。GPA は各成績評価（秀・優・良・可・不可・失格）に割り振られた得点（4 点・3 点・2 点・1 点・0 点・0 点）と、

講義毎に決められている単位数を用いて式 3.1 により計算することができる。

$$GPA = \frac{\sum_{\text{受講した講義全て}} (\text{成績得点}) * (\text{講義の単位数})}{\sum_{\text{受講した講義全て}} (\text{講義の単位数})} \tag{3.1}$$

例えば、ある学生が全ての講義において秀の成績を獲得すれば、GPA は 4.0 となる。対して全ての講義において単位を取り逃せば GPA は 0.0 となる。この GPA の値を比較することで学生の修学態度や学習能力を比較評価することができる。

本研究では、全学生の各年次別の年間・前期・後期の GPA を算出した。さらに、講義をその科目や種類に応じて分類し、分類別の GPA も別途算出した。以下の表 3.1 にその分類の一覧を示す。

また、GPA とは別に、各成績評価の獲得数も学期・年次別に算出した。例えばある学生の GPA が 2.0 であっても、どの教科も良の評価を得ている場合と、秀と可の両極端な評価を得ている場合が考えられるからである。成績評価が秀・優・良・可・不可・失格の 6 種類存在するため、各学期において 6 つの変数が生成される。また、分類別の各評価の獲得数も算出しデータ化した。

表 3.1: 科目や種類に応じた分類の一覧

分類名	内容
外国語 GPA	外国語に関係する講義の GPA
人文 GPA	人間文化などのリベラルアーツに属する講義の GPA
数学 GPA	線形代数や微分積分などの数学教科に関係する講義の GPA
理科 GPA	物理や化学など理科に分類される講義の GPA
体育 GPA	体育科目の GPA
専門 GPA	専門科目の GPA
その他 GPA	レクリエーション的講義や学生交流型の講義など上記のものに属さない講義の GPA

3.2.2　打刻データの拡張及び変換

先述した通り、本研究では学生個人の傾向に注目しているため、打刻データに関しては各学生の打刻回数に着目した。打刻した日付の情報から月別打刻数、週間別打刻数、曜日別打刻数のデータを作成した。ただし、学生が受講している講義の数によって打刻数も増減するため、ある期間に受講している講義数も調査し、(打刻回数)/(受講数) の値を算出した。これにより一講義あたりの打刻数を概算することができる。ちなみに一般的には、教室への入室と退室に一回ずつ打刻を行うため、(打刻回数)/(受講数) の値が 2 であるのが望ましい。それよりも少ない場合は、講義に出席していない日が多い可能性が高いと考えられる。

第4章　『要注意学生』の概要とその傾向の分析

本章では3章で述べたデータを用いて、『要注意学生』の存在を指摘し、その『要注意学生』の修学状況の分析を行った。以下より、分析の環境の概説、分析の手順と方針、『要注意学生』の概要、データマイニングを用いた分析の結果について述べる。

4.1　分析環境

本研究では分析の際に活用するソフトウェアとして、MicroSoft Office Excel 2010と、フリーのデータマイニングツールであるWeka [24] を採用した。Wekaはニュージーランドのワイカト大学で開発されたソフトウェアで、JAVAで実装されている。主にデータ解析、予測モデリングや視覚化ツールとして多くのアルゴリズムが搭載されており、小規模のデータであれば難なく利用できる。具体的には、クラスタリング、統計分類、回帰分析、視覚化、特徴選択といった標準的データマイニングタスクが盛り込まれている。

Excelは統計検定やデータ整備を実行するツールが標準的に装備されており、グラフ化やその装飾も容易である。また、ExcelとWekaは共にCSVファイルを扱えるためデータ互換性が高く、相互に適用しやすい点が本研究にて採用された1つの要因となっている。

4.2　分析の手順と方針

分析を進めるには、データや手法を選択し、得られた結果に解釈を与える必要がある。しかし、データや手法の無闇な選択や、方針を持たない状態での分析結果の解釈は、最終的な良い分析から遠ざかる要因である。そのため、予め分析及びデータマイニングの手順や過程、方針を定めておく必要がある。

そこで、分析を進める前に、本研究の修学状況分析モデルを制定した。本研究の分析にはデータマイニングの手法を用いるため、データマイニングのプロセスを模範とした。データマイニングのプロセスは、以下のように大別できる。

1. データの準備（データの獲得、データクレンジング）

2. 諸手法によるパターンの発見

3. パターンの解釈

データの獲得については 3 章で述べており、既に完了していると言える。本研究の分析に必須な作業は、データクレンジング、パターンの発見、解釈である。ただし、データ形式や手法の無闇な選択は、結果の解釈を困難にする。目的を設定しなければ、得られた結果を正当に評価することができないからである。このことを踏まえた上で上記のプロセスを参考にして制定された修学状況分析モデルを図 4.1 に示す。

まずは分析の前に仮説を立てる。例えば、「出席率が日増しに悪くなる学生は、何かしらの科目の成績が悪い傾向にある」といったものや、「退学者は出席率が悪い」といった仮説を立てる。これにより、自ずと目的設定がなされる。前者の例ならば、出席率が悪い学生と良い学生の科目別成績値を比較することで、その結果を評価することができる。次に本研究では 3 章で述べたように異なる形式を持つ複数のデータを準備しているため、仮説に先立ったデータ形式を選択する。必要であればデータを適宜変換し、そのデータの特徴に応じた分析手法を選択する。得られた結果に対し、可視化や統計検定を用いて解釈を与える。この一連の手順により有用な知識や傾向を獲得する。

しかし、分析が必ずしも成功するとは限らない。データに不備があったり、分析手法の選択を誤ったりすれば、有意な分析結果は得られず知識の獲得には至らない。分析が失敗したときは、その原因を追究し、過程における選択を検討する必要がある。データマイニングのプロセスはしばしば”循環型”と形容され、成功・失敗に関わらず結果に応じて各ステップを試行錯誤していく必要があると提言されている。本分析もこれに倣い、各ステップの結果を吟味し、データの形式や手法を適宜検討し、有意な傾向や知識を得るまでプロセスを反復試行した。

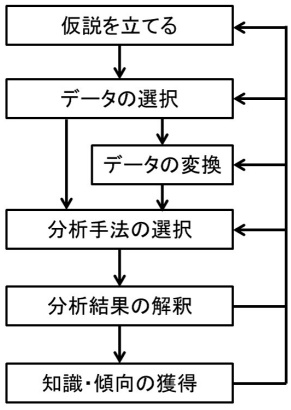

図 4.1: 修学状況分析モデル

4.3　学生の修学状況の一般的傾向調査による『要注意学生』の指摘

3 章で述べたデータを用いて、学生全体の修学状況を調査した。この調査を通して、どれだけの学生が、どのような修学態度でいるかを概観し、『要注意学生』に分類されるような学生の指摘を試みた。

用いたデータには名古屋工業大学卒業生 338 名の在籍時の修学状況が記録されており、その内約は 2 年度分（171 名と 167 名）である。これらを A 年度・B 年度と表記し、調査を進めた。まずは年度別の「卒業研究着手に要した年数」と「卒業に要した年数」の分布を 24 ページ目の表 4.1 と表 4.2 に示す。ちなみに、表 4.1 における「未着手」は記録上卒業研究に着手できていないことを意味する。また、表 4.2 における「在学中」は卒業しておらず籍だけは置かれている状況を意味し、両方において「退学」は卒業研究の着手及び卒業までに退学届が受理されたことを意味している。

名古屋工業大学では一般に 4 年次開始と同時に卒業研究に着手する。その際に指定の単位取得条件を満たしていない場合は卒業研究に着手できず、事実上の留年となる。表 4.1 に示す通り、調査対象の学生 338 名のうち、283 名が順調に卒業研究に着手しており、換言すれば、55 名が 4 年次開始時に卒業研究に着手できていない。割合にすると、全体の約 15 % が順調に学業に着手できていないことが分かった。また、卒業状況に関しても、338 名中 70 名の学生が 4 年で卒業できて

表 4.1: 各年度における「卒業研究着手に要した年数」の分布

	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: 各年度における「卒業に要した年数」の分布

	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 4.3: 4 年開始時に卒研着手した学生の「卒業に要した年数」の分布

	4 年	5 年	6 年	在学中	退学	合計
A 年度	134	10	0	0	1	145
B 年度	134	0	0	1	3	138
合計	268	10	0	1	4	283

いない状況にあることが分かった。割合にすると約 20 % が順調に卒業できていないことになる。

さらに、通常の場合であれば、卒業研究は 1 年間で行われるが、4 年次開始時に卒業研究着手が叶ったものの 4 年次終了時に卒業できなかった学生が 284 名中 15 名存在することが分かった（283 − 268 = 15）。そこで、4 年開始時に卒業研究に着手した学生の卒業年次を以下の表 4.3 に示す。A 年度において、4 年次開始時に卒業研究着手できたものの 1 年の留年してしまった学生（つまり 5 年次に卒業した学生）が 10 名いることが確認できた。また、卒業研究まで学業を進めているにも関わらず退学してしまった学生が 4 名存在した。この原因として、卒業研究あるいは卒業論文の執筆において成果を残せず指導教員の合格が得られなかった場合、卒業条件である取得単位を満足できなかった場合、就職活動の失敗による戦略的留年の場合などが挙げられる。

さらに、元のデータの 1 つである講義別成績データから、各学生の登録講義の合計単位数と取得できた単位数を調査した。当大学では学期の初めに、受講する講義を登録する手続き（履修登録）がある。登録された講義はデータとして保持され、最終成績も同データに記録されている。言い換えれば、ある学生が履修登録をしていなかった場合、講義別成績データに記録が付かないため、履修登録の有無を判断することができる。つまり、講義別成績データを参照することで、「履修登録をした（　単位取得の意志がある）」学生であるのか、「履修登録をしていなかった（　単位取得の意志が見られない）」学生であるのかを調査することができる。この調査を通して、学生の修学状況やその態度を分析した。

全学生 338 名のうち、卒業に 5 年以上費やした学生 70 名の登録単位数を調査した。表 4.4 に、各学期における履修登録を完了していない学生の人数を示す。1 年前期では全員が履修登録を完了していたが、1 年後期以降から履修登録していない学生が増えることを確認できた。学期の登録単位数が 0 となっていた場合、既に大学に来ていない可能性が高い。学期が進むに伴い、その人数が増加していることも分かった。また、履修登録をしていない学期が存在する学生は、70 名中 18 名であった。これは該当する 70 名の約 25 %に相当する。換言すれば、約 75 %の学生が単位取得の意志があるものの成績が低迷していたことが分かった。

次に、対象の 70 名の学期別 GPA を調査した。学期別 GPA の平均を表 4.5 に示す。各学期の GPA の平均より 1 年後期から GPA が低下していることが分かった。しかし個人単位では、1 年次から成績が低迷している学生もいれば、3 年次まで平均以上の成績を修めている学生もいることが確認できた。また、ある学期とその前学期の差が 1.0 以上の時期が存在する学生が 70 名中 36 名であったことから、入学以降に平均的な成績を残しつつもある時期を境に成績が急低下する学生が 70 名中の半数以上であることが分かった。よって、該当する 70 名の中には、1 年次から成績低迷の傾向を示す学生もいれば、ある時期まで成績が安定している学生もいることが分かった。

1 年次の成績値と今後の修学傾向の関係をさらに調査するため、全学生を対象に、1 年次前期と後期を GPA を等分割し、その区分に含まれる学生の人数と退学者及び留学者の人数を調査した。表 4.6 にその結果を示す。

表 4.4: 対象の 70 名において履修登録が確認できなかった学生の人数

学期	1 年前期	1 年後期	2 年前期	2 年後期	3 年前期	3 年後期
A 年度	0	1	2	2	4	4
B 年度	0	1	3	8	5	6
合計	0	2	5	10	9	10

表 4.5: 対象の 70 名と全体の学期別 GPA

学期	1 年前期	1 年後期	2 年前期	2 年後期	3 年前期	3 年後期
対象 70 名	1.80	1.36	1.24	0.97	1.11	0.92
全体	2.40	2.18	2.13	1.99	2.14	2.14

表 4.6: 1 年前期と 1 年後期の GPA 値域別の退学者・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 %
	338	70		338	70	

1 年前期と 1 年後期ともに、GPA が高い学生ほど退学・留年する割合が少なかった。しかし、GPA が 3.0 以上を記録している学生でも少数ながら退学・留年する場合があることも確認できた。ここで、1 年後期における GPA が 1.0 を下回る学生に着目すると、該当する 23 名中 23 名、つまり全員が退学・留年していることが分かった。これは、1 年前期も同様であり、GPA が 1.0 を下回る学生 13 名のうち、11 名が留年もしくは退学していることが確認できた。すなわち、1 年次の GPA が 1.0 を下回る学生は、今後留年もしくは退学する可能性がかなり高いと解釈できる。この調査結果より、本研究における真の『要注意学生』とは、「1 年次の GPA が 1.0 を上回っているが将来修学傾向が悪化する学生」であると考えられる。1 年次の成績がかなり悪い学生はその時点で修学指導の対象となり、そうでない学生の中でも、本研究による『要注意学生』の発見によって選定された学生も指導の対象になる。

つまり、本研究では、1 年各学期の GPA が 1.0 を上回るも以降に修学傾向が悪化する学生をどれだけ事前に発見できるかが肝要となることが分かった。この結果を踏まえ、以後の分析および

発見法の提案では、「1 年前期の GPA が 1.0 以上かつ 1 年後期の GPA が 1.0 以上であるが、将来的に『要注意学生』になる」学生に特に着目し、研究を進めた。

調査結果のまとめ

1. 調査対象である 338 名のうち、70 名が卒業に 5 年以上費やしている。これは全体の約 20 %に相当する。
2. 履修登録すらしていない時期があった学生は 18 名で 70 名の約 25 %である。換言すれば、70 名の約 75 %が、単位取得の意志がありながら成績が低迷し、最終的には留年する結果となっている。
3. 70 名のうち、約半数の学生の GPA が急激に低下している。
4. 1 年次後期の GPA が 1.0 を下回る学生は、100 %『要注意学生』となっている。また、前期でも 100 %近い割合を示している。つまり、1 年次の GPA が 1.0 を下回る学生は既に指導の対象として認められ、本研究では、GPA が 1.0 以上かつ今後『要注意学生』となる学生の指摘が肝要となる。

4.4　データマイニングを用いた『要注意学生』に関する分析

前節で述べた留年者及び退学者を例とした『要注意学生』に関する調査・分析を、データマイニングの手法により行った。用いた手法は、主成分分析とクラスタリングを応用した手法である。本節では用いた手法の適用の概要と分析結果について述べる。

4.4.1　手法の概要

本研究の分析にはデータマイニング手法の 1 つであるクラスタリングを活用している。これにより、同じ傾向・特性を持つ学生を 1 つのグループに集団化し、その集団の傾向を比較することで調査及び分析を行う。しかし本研究では、扱う属性数が多くクラスタリングの結果に解釈を与えるのが難しい。そこで、クラスタリングを行う前に、主成分分析により情報を縮約し、説明を与えやすい変数を作り出す。その後にクラスタリングを行うことで、結果に対する解釈をより平易化した。分析の手順は以下のようになる。

1. 対象の学生と属性変数の選択
2. 主成分分析による情報の縮約
3. 主成分の意味付け

4. 主成分スコアの値を用いたクラスタリング

5. クラスタ毎の傾向調査

6. 必要に応じて、別手法による調査

対象の学生　前節で述べた「1 年前期または 1 年後期の GPA が 1.0 未満である学生はほぼ全員留年あるいは退学をしていた」という知見をもとに、対象の学生は、「1 年前期と 1 年後期の GPA が共に 1.0 以上である」学生とした。これにより、1 年次の成績がそこまで悪くない学生であっても今後留年もしくは退学してしまう学生について調査を行う。対象となる学生は 307 名である。

属性変数　属性変数は 3 章で述べた 3 つのデータを主として、必要に応じて用いる属性変数の組み合わせ、あるいは縮小を行った。

主成分分析の概要　主成分分析における成分数は、累積寄与率が 70 %を超える成分数を採用した。

クラスタリングの概要　また、クラスタリングの手法は K-means 法を採用した。K-means 法におけるクラスタ数は事前に設定する必要がある。本分析ではクラスタ数を 10 に設定した結果について述べるものとする。用いる距離指標はユークリッド距離とした。

本分析では、科目別 GPA、獲得成績数、打刻回数について分析を行った。この分析により、1 年の成績がそこまで悪くなくとも『要注意学生』になる学生の傾向の調査を試みた。次節からその結果について述べる。

4.4.2　科目別 GPA に関する分析

4.4.1 節で述べた学生 307 名を対象として、科目別 GPA と未来の修学状況に関する分析を行った。属性変数は 1 年次の科目別 GPA データを用いた。まずは、主成分分析によりデータの縮約を試みた。主成分分析の結果を表 4.7 に示す。

第 6 主成分において累積寄与率が 70 %を超えたため、採用する主成分数を 6 とした。次に各主成分の固有ベクトルを表 4.8 に示す。そして、各主成分における固有ベクトルを解釈することで、主成分に意味付けを行った。その結果を表 4.9 に示す。

第 1 主成分では全ての固有値が負の値となっていた。このことから、第 1 主成分を「全体的な成績の悪さ」と意味付けした。第 2 主成分では、英語科目の値が大きくなっていた。このことから第 2 主成分を「言語能力の高さ」と意味付けした。第 3 主成分では、数学と理科の科目の値が低くなっていたため、「理系的素養の無さ」と名付けた。第 4 主成分では、第 3 主成分とは対照的に、前期よりも後期の値が大きかった。よって「やる気増加度」と名付けた。第 5 主成分でも、前

表 4.7: 各主成分の固有値、次成分とのその差分、寄与率、累積寄与率

	固有値	寄与率	累積寄与率
第 1 主成分	2.258	0.364	0.364
第 2 主成分	1.198	0.102	0.467
第 3 主成分	1.117	0.089	0.556
第 4 主成分	1.012	0.073	0.629
第 5 主成分	0.992	0.070	0.699
第 6 主成分	0.888	0.056	0.756
第 7 主成分	0.698	0.037	0.793
⋮	⋮	⋮	⋮
第 13 主成分	0.504	0.018	0.988
第 14 主成分	0.410	0.012	1.000

表 4.8: 各主成分の固有ベクトル

	学期	英語	人文	数学	体育	理科	専門	その他
第 1 主成分	前	-0.168	-0.226	-0.325	-0.120	-0.313	-0.342	-0.232
	後	-0.188	-0.206	-0.340	-0.183	-0.306	-0.359	-0.288
第 2 主成分	前	0.579	-0.150	-0.006	-0.106	0.046	-0.190	-0.260
	後	0.616	0.092	0.087	-0.034	0.114	-0.105	-0.322
第 3 主成分	前	0.334	0.221	-0.244	0.533	-0.306	0.204	0.282
	後	0.210	-0.056	-0.355	0.129	-0.237	-0.096	0.161
第 4 主成分	前	-0.271	0.099	-0.110	0.391	-0.126	-0.269	-0.568
	後	0.047	0.333	-0.066	0.358	0.261	0.103	0.130
第 5 主成分	前	0.004	-0.141	-0.323	-0.246	-0.256	0.046	-0.027
	後	0.149	0.356	-0.081	-0.529	0.027	0.350	0.434
第 6 主成分	前	0.078	0.625	0.133	0.000	0.101	-0.110	-0.001
	後	-0.184	0.538	-0.033	-0.329	-0.180	-0.195	-0.245

期よりも後期の値の方が高い傾向にあった。第 4 主成分と比較して、数学と理科科目の値が低く、後期専門科目の値が高かったため、「後期の専門科目に強い」と意味付けした。第 6 主成分では人文科目において高い値が確認できたため、「他分野に対する理解度」と意味付けした。

この主成分スコアを属性変数として、K-means 法によるクラスタリングを試行した。クラスタ数を 10 に設定した際のクラスタリング結果を表 4.10 に示す。各クラスタの人数、その中での留年者及び退学者の人数を記載する。

留年・退学者の割合が最も高かったクラスタはクラスタ 10 であった。このクラスタは全体的な成績がかなり悪く、しかし理系の素養は悪くないといった傾向が確認できた。留年者の人数が最も多かったクラスタはクラスタ 1 であった。このクラスタの傾向として、言語能力が低いという

表 4.9: 各主成分に与えた解釈

	解釈
第 1 主成分	全体的な成績の悪さ
第 2 主成分	言語能力の高さ
第 3 主成分	理系的素養の無さ
第 4 主成分	やる気増加度
第 5 主成分	専門科目に強い
第 6 主成分	他分野に対する理解度

表 4.10: 各クラスタにおける主成分スコアの平均値と留年者・退学者

	各主成分									
	成分 1	成分 2	成分 3	成分 4	成分 5	成分 6	人数	留年	退学	割合
1	1.32	-1.67	0.45	-0.07	0.15	-0.29	42	10	2	29 %
2	-1.09	-0.23	-0.51	0.40	-0.74	0.57	44	5	1	14 %
3	1.44	1.14	-0.27	-1.01	-0.62	-0.05	37	6	0	16 %
4	-0.42	-0.11	1.11	0.44	0.80	-0.04	36	1	0	3 %
5	3.37	0.55	0.60	-2.89	2.16	-0.22	4	1	0	25 %
6	-2.45	-0.78	-0.55	-0.39	0.00	-0.49	36	1	0	3 %
7	-2.41	0.98	-0.18	0.10	0.60	0.25	45	2	0	4 %
8	0.51	0.39	0.07	-0.03	-0.71	-1.44	26	1	3	15 %
9	2.28	-0.02	0.29	0.17	-0.13	1.15	35	7	3	29 %
10	4.47	1.47	-1.34	2.28	0.79	-0.51	10	2	2	40 %

傾向が確認できた。また、クラスタ 6 とクラスタ 7 はほとんど留年・退学者が含まれていなかった。この 2 つの共通点は、全体の成績が良好であることであり、「1 年次の成績が良い学生は留年・退学する確率が低くなる」という一般的に述べることのできる傾向が見出された。

さらに詳細な調査を進めるため、対象の学生と属性変数の削減を試みた。前述のクラスター分析では、全体の成績が悪い学生ほど留年・退学する確率が高くなることを示していた。さらに主成分分析により、体育科目とその他科目の GPA が未来の修学傾向に大きく寄与していないことも分かった。そこで、対象の学生を 1 年次後期の GPA1.0 以上 2.0 未満の学生に限定し、属性変数は体育科目とその他科目を省いたものとした。これにより、好成績を残した層を調査の対象から外し、順調に修学する学生と『要注意学生』となってしまう学生が入り混じっている中間層の傾向を調査した。表 4.6 に示す通り、対象の学生は 98 名、その内の 29 名（約 30 %）が留年及び退学した学生である。

同様の手順により調査を行った。まずは主成分分析により情報の縮約を試みた。各主成分の寄与率及び累積寄与率を表 4.11、各主成分の固有ベクトルを表 4.12 に示す。第 5 主成分にて累積寄

表 4.11: 各主成分の固有値、次成分とのその差分、寄与率、累積寄与率

	固有値	寄与率	累積寄与率
第 1 主成分	1.501	0.225	0.225
第 2 主成分	1.285	0.165	0.391
第 3 主成分	1.098	0.121	0.511
第 4 主成分	1.009	0.102	0.613
第 5 主成分	0.961	0.092	0.705
第 6 主成分	0.906	0.065	0.770
⋮	⋮	⋮	⋮
第 9 主成分	0.617	0.038	0.965
第 10 主成分	0.591	0.035	1.000

表 4.12: 各主成分の固有ベクトル

	学期	英語	人文	数学	理科	専門
第 1 主成分	前	0.092	0.239	0.544	0.454	0.310
	後	-0.028	0.134	0.413	0.365	0.119
第 2 主成分	前	-0.604	0.242	-0.127	0.076	0.126
	後	-0.640	0.141	-0.217	0.045	0.247
第 3 主成分	前	0.391	0.499	-0.070	-0.163	0.227
	後	0.206	0.411	-0.355	-0.171	0.385
第 4 主成分	前	0.149	0.002	0.080	-0.235	0.524
	後	-0.181	-0.690	0.083	-0.248	0.252
第 5 主成分	前	-0.070	0.200	-0.115	0.030	0.437
	後	0.166	-0.162	-0.432	0.403	-0.589

表 4.13: 各主成分に与えた解釈

	解釈
第 1 主成分	理系教科の強み
第 2 主成分	英語力の無さ
第 3 主成分	真面目であはるが理数系が苦手
第 4 主成分	他分野への興味の減退
第 5 主成分	講義レベルに追い付けていない

表 4.14: 各クラスタにおける主成分スコアの平均値と留年者・退学者

	各主成分					人数	留年	退学	割合
	成分 1	成分 2	成分 3	成分 4	成分 5				
1	-0.12	0.49	-0.79	-1.28	0.24	15	2	1	20.0 %
2	-0.77	1.07	-1.04	0.47	-0.34	14	5	0	35.7 %
3	0.67	0.74	0.56	0.99	-1.08	11	3	0	27.3 %
4	-1.54	-1.54	-0.27	1.04	1.53	5	3	1	80.0 %
5	-1.43	-1.08	0.79	-1.08	-1.32	6	0	1	16.7 %
6	1.26	-1.11	-1.60	-0.11	0.01	8	3	0	37.5 %
7	-0.19	1.12	0.67	0.46	0.44	13	4	1	38.5 %
8	1.38	-0.84	0.67	-0.12	0.37	12	1	1	16.7 %
9	-0.05	-1.71	0.15	0.98	-0.81	6	2	0	33.3 %
10	-0.29	-0.19	1.50	-0.84	0.98	8	0	1	12.5 %

与率が 70 %を超えたため、主成分数を 5 とした。また、固有ベクトルから、各主成分の意味付けを行った。その結果を表 4.13 に示す。

第 1 主成分では主に、数学と理科科目が大きな値を示していたため、「理系教科の強み」と意味付けした。第 2 主成分では英語科目に負の値を示していたため、「英語力の無さ」と意味付けした。第 3 主成分では、英語・人文・専門科目で正の値、数学・理科科目で負の値を示していた。1 年次に開講される英語・人文・専門科目は、出席率重視の講義が多く、対して数学・理科科目は試験結果重視の講義が多い。ゆえに第 3 主成分を「真面目ではあるが理数系が苦手」と意味付けした。第 4 主成分では、人文後期に負の値が見て取れたため、「他分野への興味の減退」と意味付けした。第 5 主成分では専門科目において前後期で大きな差が確認でき、数学科目でも後期の値がかなり低くなっていた。これより、「講義レベルに追い付けていない」と意味付けした。

さらにこの主成分スコアを変数として、K-means 法によるクラスタリングを行った。その結果を表 4.14 に示す。

最も留年・退学者の割合が高かったのはクラスタ 4 であった。このクラスタの傾向として、理系教科に弱く、英語力はあり、他分野への興味がなく、講義レベルに追い付けていないという点が指摘できた。同じように理系教科に弱いクラスタはクラスタ 5 であるが、留年・退学者の割合に差異があった。傾向の違いとして、クラスタ 5 の学生の方が真面目であり、他分野への興味があり、講義レベルにしっかりついてきているという点が挙げられた。クラスタ 5 と同様に割合が低いのはクラスタ 10 であり、共通点として、「真面目であるが理数系が苦手」「他分野に興味がある」という点が確認できた。同様の傾向示すをクラスタはないため、2 つのクラスタにおける固有の傾向であると言える。

4.4.3　科目別 GPA を用いた調査結果のまとめ

1. 307 名を対象に主成分分析を行ったところ、第 1 主成分として「全体的な成績の悪さ」が抽出された。
2. 体育科目とその他科目は今後の修学傾向に大きく寄与しない。
3. 1 年次成績中間層である 98 名を対象に主成分分析を行ったところ、第 1 主成分として「理系教科の強み」が抽出された。
4. 1 年次の成績中間層において、留年・退学者が多いクラスタは、理系教科に弱く、英語力はあり、他分野への興味がなく、講義レベルに追い付けていないという傾向を示していた。
5. 1 年次の成績中間層において、留年・退学者が少ないクラスタは、真面目であるが理数系が苦手、他分野に興味があるという傾向を示していた。

4.4.4　獲得成績数に関する分析

獲得成績数とは、講義の最終評価である「秀」や「不可」などの成績の獲得個数を意味する。3.2.1 節でも述べたように講義別成績から拡張して得られた獲得成績数データを用いて、成績評価の分布の差異における『要注意学生』の傾向分析を試みた。

対象の学生は 4.4.1 節でも述べた 307 名、属性変数は 1 年前期と後期の各成績の獲得数を示す 12 変数とした。まずは主成分分析により情報の縮約を行った。各主成分の寄与率及び累積寄与率を表 4.15、各主成分の固有ベクトルを図示化したものを図 4.2 に示す。

主成分分析した結果、第 5 主成分にて累積寄与率が 0.7 を超えたため、主成分数を 5 とした。また、固有ベクトルを検討することで各主成分の意味付けを行った。その結果を表 4.16 に示す。

図 4.2 から、第 1 主成分では、前後期共に秀と優の値が大きくなっていたため「優秀さ」と意味付けした。第 2 主成分では、前後期共に最も高評価である秀と、単位取得が認められない不可と失格の値が大きくなっていた。ゆえに、「成績の極端さ」と意味付けした。第 3 主成分では秀と良の値が大きく、優が小さい。また、不可と失格の値が増加していた。1 年次はチュートリアルも兼ねた講義も多いため秀が取りやすい背景がある。これらを踏まえ「優可の少なさ」と意味付けした。第 4 主成分では、前期と後期で値の分布に変化が見られた。前期では優の値が大きく、後期では良と失格の値が大きくなっていた。よって、「やる気喪失度」と意味付けした。第 5 主成分でも前期と後期において大きな変化が見られた。前期では可の値が大きくなっていたが、後期では平滑化されていた。このことから「学習レベル改善度」と意味付けした。

さらに主成分スコアを変数として K-means 法によるクラスタリングを行った。クラスタ数を 10 に設定した時の結果を表 4.17 に示す。

全体の傾向として、優秀でないクラスタほど、留年・退学者の割合が高くなる傾向が確認できた。また、成績が極端なほどその割合が高く、安定するほど割合が低くなることも確認できた。

クラスタ 7 は該当する 3 名中 3 名が留年していることが分かった。このクラスタの傾向として、優秀ではなく、成績が極端で、やる気喪失度が高く、学習レベルが追いついていない傾向が指摘できた。クラスタ 8 も割合が高かった。クラスタ 7 とクラスタ 8 の共通点として、優秀ではなく、成績が極端で、やる気の喪失が認められる点が確認できた。クラスタ 6 も割合が比較的高く、優秀ではなく、成績が極端である傾向があった。対してクラスタ 3 は成績が極端であるが、優秀さがあるため留年・退学者は少なかった。このことから、「成績が優秀ではなく獲得する成績が極端な学生は『要注意学生』になりやすい」ということが言える。

対して、割合が低いクラスタは、クラスタ 3、クラスタ 4、クラスタ 5 であった。クラスタ 3 とクラスタ 5 は優秀であるが、クラスタ 4 はそこまで優秀でなかった。しかし、成績が極端ではないという傾向を示していた。つまり、「優秀でなくても成績が極端でない学生は『要注意学生』になりにくい」ということが言える。

4.4.5　獲得成績数データを用いた調査結果のまとめ

1. 主成分分析の結果、「優秀さ」が第 1 主成分として抽出された。
2. 優秀ではなく、成績が極端な学生は『要注意人物』になる割合が高かった。
3. 優秀でなくても、成績が極端でなければ『要注意学生』になる確率が低くなる。

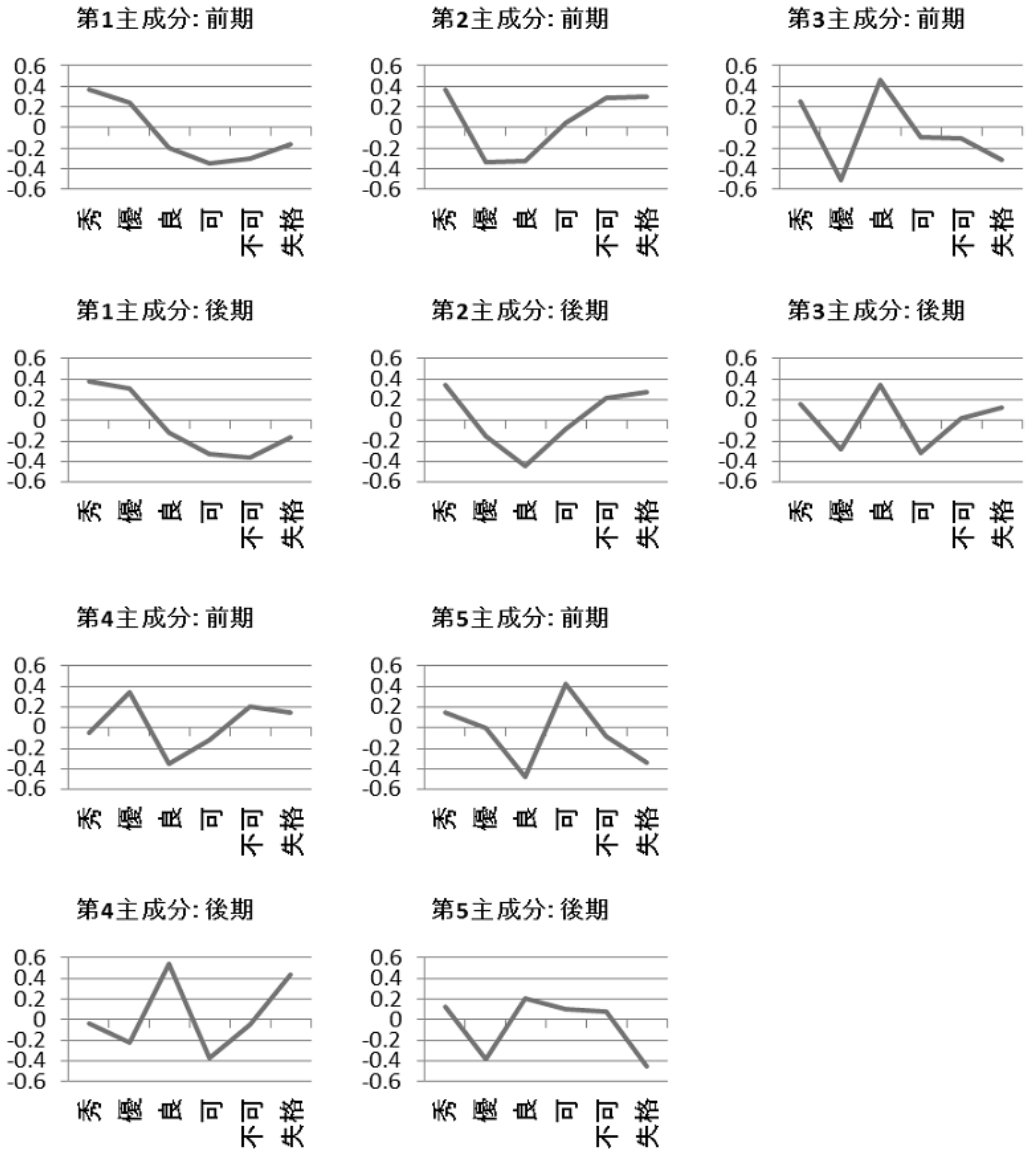

図 4.2: 各主成分の固有ベクトル

表 4.15: 各主成分の固有値、次成分とのその差分、寄与率、累積寄与率

	固有値	寄与率	累積寄与率
第 1 主成分	1.903	0.302	0.302
第 2 主成分	1.323	0.146	0.448
第 3 主成分	1.072	0.096	0.543
第 4 主成分	1.042	0.090	0.634
第 5 主成分	1.003	0.084	0.718
第 6 主成分	0.945	0.055	0.773
⋮	⋮	⋮	⋮
第 11 主成分	0.091	0.001	0.999
第 12 主成分	0.032	0.000	1.000

表 4.16: 各主成分に与えた解釈

	解釈
第 1 主成分	優秀さ
第 2 主成分	成績の極端さ
第 3 主成分	優可の少なさ
第 4 主成分	やる気喪失度
第 5 主成分	学習レベル追従度

表 4.17: 各クラスタにおける主成分スコアの平均値と留年者・退学者

	各主成分					人数	留年	退学	割合
	成分 1	成分 2	成分 3	成分 4	成分 5				
1	-0.59	-0.48	0.11	0.60	1.09	37	3	1	10.8 %
2	-1.84	-0.30	0.52	-1.18	0.06	39	4	1	12.8 %
3	2.44	1.58	1.01	0.01	0.47	29	0	1	3.4 %
4	0.32	-1.16	-0.90	0.08	-0.16	53	2	0	3.8 %
5	2.26	0.36	-0.84	0.53	0.11	32	1	0	3.1 %
6	-2.29	1.29	-1.59	-0.16	0.08	24	5	1	25.0 %
7	-4.10	3.82	0.80	3.81	-4.13	3	3	0	100.0 %
8	-2.50	1.01	0.51	0.79	-0.14	23	7	3	43.5 %
9	-0.14	-1.46	1.46	0.34	-0.44	27	1	2	11.1 %
10	1.57	0.19	0.14	-0.81	-0.63	40	6	0	15.0 %

4.4.6　打刻回数に関する分析

打刻データを用いて、出席率と打刻回数と『要注意学生』に関する調査を行った。3.1.2 節でも述べたが、”打刻”は学生が教室での入退室の際に学生証を端末に認識させることを意味する。打刻回数が出席状況の一端を表していると言える。

まずは、対象の学生を前述の 307 名、属性変数を月別打刻数（4 月から 8 月分と 10 月から 2 月分）の 10 変数として、調査を行った。主成分分析の結果を表 4.18 に示す。第 3 主成分にて累積寄与率が 0.7 を超えたため、成分数を 3 とした。さらに固有ベクトルを図 4.3 に示す。

第 1 主成分では全体的に値が大きくなっていた。よって「勤勉に出席したかどうか」と意味付けした。第 2 主成分では 4 月の値が大きく、8 月の値が小さくなっていた。さらに 4 月から 8 月、10 月から 1 月にかけて値が減少していることも確認できた。これより、第 2 主成分を「徐々に欠席率上昇」と意味付けした。第 3 主成分では前期（4 月から 8 月）の値が大きく、後期（10 月から 2 月）の値が小さくなっていた。そこで、「後期に出席状況悪化」と意味付けした。

さらに 3 つの主成分スコアを変数として、K-means 法によるクラスタリングを行った。表 4.20 にその結果を示す。クラスタ数は 10 とした。

クラスタ 8 の該当者は 1 名で、その 1 名は留年していた。このクラスタが示す特徴として、勤勉に出席しておらず、後期に出席しなくなっていることが確認できた。次に留年・退学の割合が高いクラスタはクラスタ 7 であった。このクラスタの傾向として、出席率がかなり悪いこと、出席率が学期開始時から低下していることが確認できた。対して、割合が低いクラスタ、クラスタ 1、クラスタ 4、クラスタ 6、クラスタ 9 に着目すると、総じて出席率が高いという傾向が確認できた。

このことから、1 年次の打刻回数が今後の修学状況を予測するのに重要な要因になりうることが分かった。

4.4.7　打刻情報を用いた調査結果のまとめ

1. 307 名を対象に主成分分析を行ったところ、第 1 主成分として「勤勉に出席したかどうか」が抽出された。

2. 打刻回数の少ないクラスタは留年・退学の割合が高く、回数の多いクラスタは割合が低い。

表 4.18: 各主成分の固有値、次成分とのその差分、寄与率、累積寄与率

	固有値	寄与率	累積寄与率
第 1 主成分	2.269	0.515	0.515
第 2 主成分	1.329	0.177	0.691
第 3 主成分	0.877	0.077	0.768
第 4 主成分	0.799	0.064	0.832
⋮	⋮	⋮	⋮
第 9 主成分	0.401	0.016	0.988
第 10 主成分	0.351	0.012	1.000

表 4.19: 各主成分に与えた解釈

	解釈
第 1 主成分	勤勉に出席したかどうか
第 2 主成分	徐々に欠席率上昇
第 3 主成分	後期に出席状況悪化

表 4.20: 各クラスタにおける主成分スコアの平均値と留年者・退学者

	各主成分			人数	留年	退学	割合
	成分 1	成分 2	成分 3				
1	1.03	-1.20	0.03	66	2	0	3.0 %
2	-2.15	-0.02	-0.04	44	9	3	27.3 %
3	-2.59	-2.07	1.62	10	1	0	10.0 %
4	1.77	1.82	0.77	47	3	1	8.5 %
5	-1.01	-3.17	2.42	7	1	0	14.3 %
6	0.87	0.73	-0.35	69	3	3	8.7 %
7	-5.05	1.47	0.15	16	7	1	50.0 %
8	-0.47	-1.93	3.44	1	1	0	100.0 %
9	3.35	-1.77	-0.96	11	1	0	9.1 %
10	-1.09	-0.43	-1.15	36	4	1	13.9 %

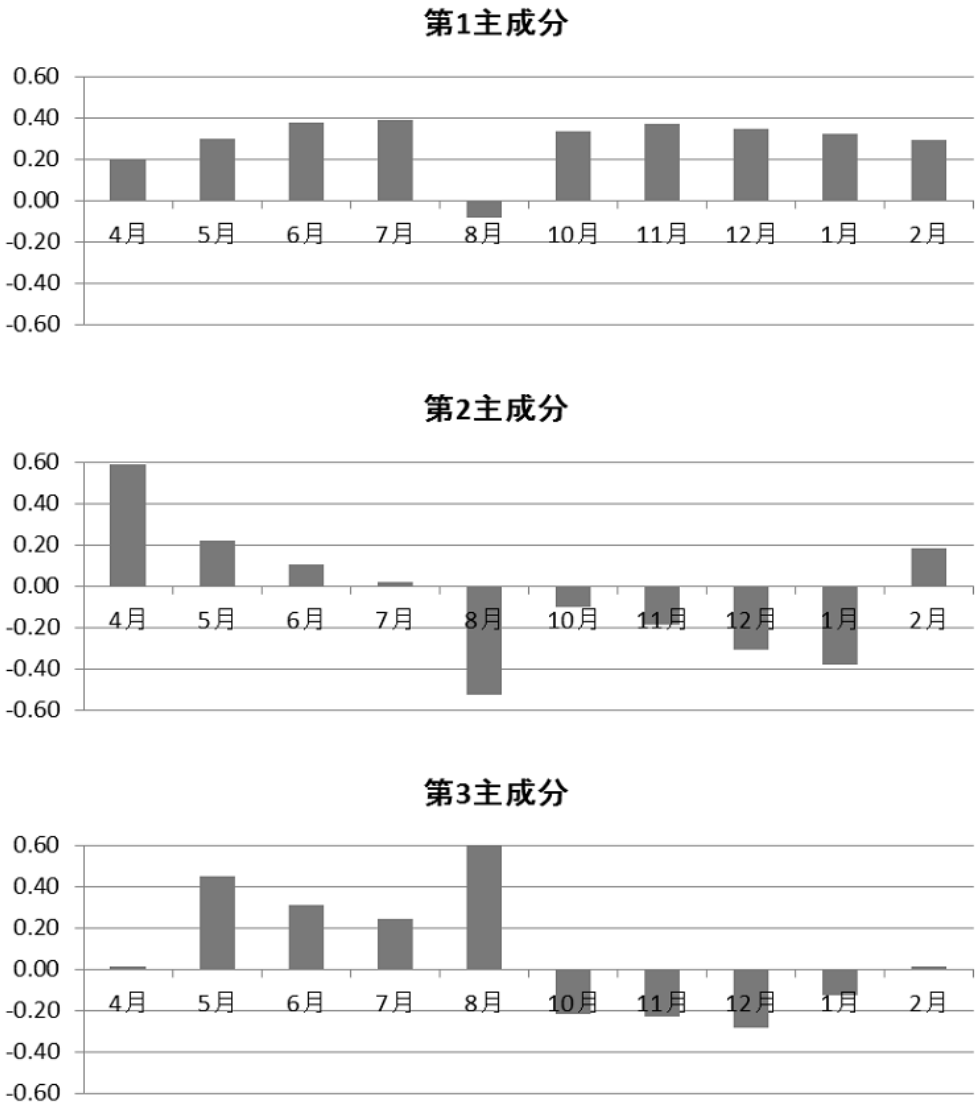

図 4.3: 各主成分の固有ベクトル

4.5　調査・分析の総括

本章では、『要注意学生』の概要と、各種データを用いて『要注意学生』の傾向の調査及び分析について述べた。本節ではそれらを総括する。

本研究では名古屋工業大学に所属していた学生 338 名を対象に、学生の修学状況と『要注意学生』について調査を行った。卒業研究の着手時期と卒業時期に着目して調査を進めた結果、338 名中、70 名が留年もしくは退学していることが分かった。これは全体の約 20 %に相当する。またその 70 名のうち、1 年前期と後期の GPA が 1.0 を下回る学生は、ほぼ 100 %の割合で留年・退学していることも分かった。

このことから、「1 年前期の GPA が 1.0 を下回るか、1 年後期の GPA が 1.0 を下回る学生」は既に修学指導の対象だとみなし、1 年各学期の GPA が 1.0 を上回りつつ、将来的に留年・退学してしまう学生を『要注意学生』であるとした。

さらに、前述の『要注意学生』の特別な傾向を見つけるべく、データマイニングの手法を応用した調査・分析を行った。用いたデータマイニングとして、主成分分析と K-means 法によるクラスタリングを活用した。本研究では、科目別 GPA、獲得成績数、打刻回数に関するデータを用いた。科目別 GPA に関する調査では、1 年次の成績中間層において、理系教科の成績が第 1 主成分として取り出せた。また特に留年・退学者の割合が高いクラスタ、低いクラスタの科目別 GPA における傾向を見出すことができた。獲得成績数に関する調査では、成績の優秀さ以外に、獲得する成績の極端さも主要因として取り出すことができた。打刻回数に関する調査では打刻回数の多さが今後の修学状況に寄与していることが分かった。

また、『要注意学生』の傾向の獲得とは別に、本研究で得られた成果として、用いた各種データが『要注意学生』に関する属性変数になりうることが分かった点が挙げられる。本研究では諸手法による『要注意学生』の発見を試みるが、データが『要注意学生』の是非に無関係であれば発見は上手くいかない。しかし、本分析によって各データと『要注意学生』に関係性が確認されたため、発見の際にも当データを登用する根拠を得ることができた。

次節では『要注意学生』の発見について述べる。本章で述べた分析・調査を基に、より有用な『要注意学生』の発見を試みる。

第5章　『要注意学生』の発見法の検討

前節では『要注意学生』の存在を指摘し、その傾向を分析した。本研究ではさらに、その結果をヒントとして、事前に『要注意学生』を発見する手法を提案する。提案手法が実用化されれば、『要注意学生』に対し予め警告を与えることができるため、全体の修学環境の改善が期待できる。本節ではまず、「発見」という語句に着目し、「発見」の概要を述べる。さらに、修学指導の対象となるべき学生、すなわち『要注意学生』の厳密な定義について述べる。

さらに、本研究では、そのような学生を早期に発見する手法として、ベイジアンネットワークの確率推論を用いた手法を提案している。本章では、GPA 値に閾値を設けた発見法と、ベイジアンネットワークによる手法を比較し、本提案の有用性を検証する。本章では、それらの手法の概要と発見精度について述べる。

5.1　発見の概要

本研究における「発見」は、換言すれば「未来予測」に相当する。既知（過去）の事象から未知（未来）の事象を予測する手法は多く存在し、その評価方法も既に確立されている。本節では「発見」及び未来予測の概要について述べる。具体的には、どのタイミングで予測をするのか、「発見」の対象者は誰になるのか、そして「発見」モデルの評価はどうするかについて以下より述べる。

5.1.1　発見の時期

発見の時期は、修学指導の時期とも言える。例えば、学生が入学して数週間経過した時点で『要注意学生』の発見を試みても既知の情報が乏しく有用な発見は実現できない可能性が高い。逆に、情報を求めるあまり修学指導の時期を遅らせすぎても、事前の指導の効果が薄れてしまう。情報量と指導の効果のバランスを考えながら、発見の時期を決定する必要がある。

本研究では発見の時期として、1 年次終了時を想定した。つまり、1 年次前期と後期に関する情報は既知であり、2 年次以降の情報や修学状況は分からない状況である。この既知の情報を頼りに、2 年次以降の修学状況の未来予測を試みた。

5.1.2　発見対象者および『要注意学生』の定義

発見・未来予測をするためには、その対象を予め定める必要がある。この対象は、本研究における「事前指導によって救い出したい学生」に相当する。本研究では望ましい発見を「既知の情報からは一見修学指導の対象になりうると判断できないが、実は将来的に修学状況が悪化する可能性が高い学生を拾い上げる」発見だと考えている。例えば、1 年次から GPA がほぼ 0.0 を示している学生を指導対象者であると判断するのは容易である。逆に 1 年次の成績が平均的であるが、将来修学状況が悪化してしまう学生を発見できれば、本研究の有用性を示すことができる。

4.3 節での調査では、「1 年各学期の GPA が 1.0 未満である学生は、ほぼ 100 %の割合で将来的に留年もしくは退学していた」という傾向が確認された。この知見をヒントに、「1 年前期の GPA が 1.0 未満か、または、1 年後期の GPA が 1.0 未満である学生」は無条件に修学指導の対象者とし、「1 年前期と 1 年後期の GPA がともに 1.0 以上である学生」の中に含まれる将来的に修学状況が悪化する学生を『要注意学生』もとい発見対象者とした。このような学生を発見することで、1 年次の成績が一見順調そうでありながらも 2 年次以降に状況が悪化する学生を救い出すことが期待できる。

5.1.3　発見モデルの評価

本研究の提案手法の実用を想定すると、「発見」モデルによって選別された学生は修学指導の対象として扱われ、指導者である教員に呼び出されたり、何かしらの警告が与えられたりすることが考えられる。本研究では指導のコストを削減することを目的としているため、なるべく指導の対象者は減らしたい。しかし、対象者の限定を考慮しすぎるあまり、実際の『要注意学生』を多数見逃してしまう恐れもある。理想の発見は、実際の『要注意学生』が全て発見されることが望ましいが、現実的に難しいと言える。これらのバランスを考慮した発見モデルを評価したい。

本研究では発見モデルの評価に、機械学習の評価法を用いている。ここで『要注意学生』であるかどうかの 2 値的予測を想定し、評価法の説明をする。

学生の中に事実として『要注意学生』である学生とそうでない学生が存在し、そして、各学生に対し『要注意学生』であるかどうかの予測を与える。ここで、『要注意学生』であることを（本研究目線において）Positive な事象であると捉えたとき、実際の『要注意学生』に対し「『要注意学生』である」という予測を与えられた場合を True Positive（以下 TP と記載する）と表記する。すなわち、Positive な事象が True（正解）であったことを意味している。この表記を基本としたとき、各状況に応じた表記は表 5.1 の通りになる。この表に示す事例数を用いて、正解率 $Accuracy$、再現率 $Recall$、適合率 $Precision$、そして予測精度の評価指標である $F-measure$ を算出する。これらの指標は以下の式 (5.1) のように求められる。

$$
\begin{aligned}
Accuracy &= \frac{TP+TN}{TP+FP+FN+TN} \\
Recall &= \frac{TP}{TP+FN} \\
Precision &= \frac{TP}{TP+FP} \\
F-measure &= \frac{2Recall*Precision}{Recall+Precision}
\end{aligned} \tag{5.1}
$$

正解率 $Accuracy$ は実際と予測の的中率を表している。再現率 $Recall$ は「実際の『要注意学生』のうち、何人の学生を発見・予測できたか」を表している。適合率 $Precision$ は「『要注意学生』と予測した学生のうち、何人の学生が実際の『要注意学生』であったか」を表している。$F-measure$ は適合率と再現率の調和平均である。

例として、20 名の学生を仮定し、この 20 名の中に『要注意学生』が 5 名いたとする。そして予測によって 8 名の学生を『要注意学生』と予測し、その予測された 8 名のうち 3 名が『要注意学生』であったとする。この場合において上で述べた 4 つの指標の値は、$Accuracy = \frac{3+10}{20} = 0.65$、$Recall = \frac{3}{5} = 0.6$、$Precision = \frac{3}{8} = 0.375$、$F-measure = \frac{2*0.375*0.6}{0.375+0.6} = 0.46$ となる。この結果を総評すると「全体としての正解率は 65 %である」「実際の『要注意学生』の 60 %を拾い上げた」「『要注意学生』と予測された学生のうち実際に『要注意学生』であったのは 37.5 %で、そうでない学生が 62.5 %である」と評価できる。

本研究で考えられるリスクとして、実際の『要注意学生』であるのに『要注意学生』と予測されない場合（= FN）と、実際の『要注意学生』でないのに『要注意学生』と予測される場合（= FP）が挙げられる。本研究では前者の FN をより回避したい場合として考えている。そのため、本研究では再現率 $Recall$ の評価を重要視した。

発見とその評価の例として図 5.1 を示す。全学生が 12 名であり、△マークが『要注意学生』、○マークが『要注意学生』でない学生であるとする。そして、枠で囲まれた学生が、発見・予測によって『要注意学生』であると認定された学生である。左の例では 3 名が『要注意学生』だと指摘され、右の例では 8 名が『要注意学生』だと指摘されたこととなる。このとき、左の発見例と右の発見例ではどちらが優れた発見であるかを説明する。

それぞれの正解率 $Accuracy$ は、左の例は 75 %、右の例では 67 %となる。さらに適合率 $Precision$ を計算すると、左の例は 67 %、右の例は 50 %となる。正解率 $Accuracy$ と適合率 $Precision$ を用いてモデル評価すると左の例の方が優れていると評価できる。しかし、本研究では「『要注意学生』であるのに『要注意学生』であると発見することができない場合」を最大のリスクと考えるため、再現率 $Recall$ の値をより重要視する。再現率 $Recall$ を計算すると、左の例では 50 %、右の例では 100 %となり、右の例の方が優れた発見であると言える。このように、全体の正解率よりも、いかに『要注意学生』をピックアップできるかに着目し、発見モデルの評価を行う。

表 5.1: 実際と予測結果に応じた表記

	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）

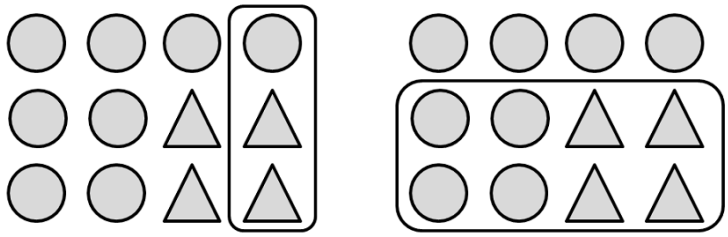

図 5.1: 発見の例。△マークが『要注意学生』

むろん、再現率 *Recall* に固執しすぎてもいけない。例えば、全学生に対し『要注意学生』であると予測を行えば、自ずと再現率 *Recall* は 100 %になる。この場合、本研究の目的であった、指導対象者の限定が達成されていないため、発見モデルとしては優れているとは言えない。再現率 *Recall* を重要視しつつ、他の評価指標の検証を加え、モデル評価の議論を行う必要がある。

5.2　GPA 値に閾値を設ける手法による『要注意学生』の発見

本研究ではベイジアンネットワークによる手法を『要注意学生』を早期発見する方法として提案しているが、その提案手法の有用性を示すため、GPA 値の比較による『要注意学生』の発見について検証した。これは、1 年次の成績から 2 年次以降の修学状況を予測するため、1 年次の GPA の良し悪しで『要注意学生』であるかどうかを判断する手法である。本節では手法の概要と発見精度について述べる。

5.2.1　手法の概要と発見精度

本手法は、「1 年次の修学状況が芳しくない学生は 2 年次以降も芳しくない」という素朴な考えに基づいたものである。例えば 1 年次の GPA が 1.0 である学生と 3.0 である学生がいた場合、その GPA 値だけを比較すれば、『要注意学生』である確率が高いと判断できるのは前者であるのは言うまでもない。このように、GPA 値を単純に比較することで、『要注意学生』の発見を試みた。

表 5.2: 1 年次通年 GPA に閾値を設けた場合の発見精度

閾値	Accuracy			Recall			Precision			F-measure
	対象	的中		対象	的中		対象	的中		
1.0	307	266	87 %	41	0	0 %	0	0	0 %	0.000
1.2	307	266	87 %	41	0	0 %	0	0	0 %	0.000
1.4	307	267	87 %	41	4	10 %	7	4	57 %	0.167
1.6	307	266	87 %	41	11	27 %	22	11	50 %	0.349
1.8	307	254	83 %	41	15	37 %	42	15	36 %	0.361
2.0	307	244	79 %	41	22	54 %	68	22	32 %	0.404
2.2	307	205	67 %	41	26	63 %	113	26	23 %	0.338
2.4	307	176	57 %	41	30	73 %	150	30	20 %	0.314
2.6	307	144	47 %	41	34	83 %	190	34	18 %	0.294
2.8	307	112	36 %	41	36	88 %	226	36	16 %	0.270
3.0	307	89	29 %	41	39	95 %	255	39	15 %	0.264
3.2	307	64	21 %	41	41	100 %	284	41	14 %	0.252
3.4	307	51	17 %	41	41	100 %	297	41	14 %	0.243
3.6	307	42	14 %	41	41	100 %	306	41	13 %	0.236
3.8	307	41	13 %	41	41	100 %	307	41	13 %	0.236
4.0	307	41	13 %	41	41	100 %	307	41	13 %	0.236

具体的には閾値を設けることで、その閾値を下回った学生を『要注意学生』であると認定する。このとき、どれだけの学生に対し正しい認定を行ったか（正解率 Accuracy）、実際の『要注意学生』に対しどれだけの人数を指摘できたか（再現率 Recall）、『要注意学生』と認定した学生の内どれだけの人数が本当に『要注意学生』であったか（適合率 Precision）、そして F-measure を閾値毎に算出した。

対象は 307 名とし、また GPA の比較には 1 年次通算の GPA を用いた。そして、閾値を 1.0 から 0.2 ずつ増加させ、前述した各指標を評価した。表 5.2 にその結果を示す。最も高い精度を示していたのは、閾値を 2.0 とした場合のものであった。F-measure は 0.404 を示し、1 年次成績中間層に潜む 54 %の『要注意学生』を発見することができていた。

ベイジアンネットワークによる手法では、GPA 閾値を 2.0 に設定した場合よりも高い精度を示す発見モデルの構築を試みた。

5.3　ベイジアンネットワークによる『要注意学生』の発見

本節ではベイジアンネットワークを用いた『要注意学生』の発見について述べる。ベイジアンネットワークの確率推論により、『要注意学生』である確率を算出し、指導対象者の発見を試みた。ベイジアンネットワークは確率変数、有向グラフ構造、条件付き確率で定義され、特に確率変数は離散化変数でなければいけない制約があるため、離散化手法にウォード法による手法を採用した。以下より、手法の概要と発見モデルの評価について述べる。

5.3.1　手法の概要

ベイジアンネットワークを構築するには、離散化された確率変数を用意し、アルゴリズムによって確率変数をノードとした有向グラフ構造を学習し、データから条件付き確率を推定する、といった手順を行う。以下にその全体フローを示す。

1. 対象の学生と属性変数の選択
2. 属性変数の取捨選択
3. 属性変数の離散化
4. 目的変数の決定
5. ベイジアンネットワークの有向グラフの学習
6. 条件付き確率の推定
7. 発見モデルの評価
8. 結果に応じて対象や変数を議論し、1. に戻る

対象の学生　全学生 338 名の内、「1 年前期と 1 年後期の GPA が共に 1.0 以上である学生」307 名を対象とした。また、場合に応じて発見の対象者を限定した。

属性変数の取捨選択　本研究では、GPA データ、獲得成績数データ、打刻回数データを用意している。これらのデータを全て活用しても良い発見モデルを構築することができない。データを組み合わせたり、一部削減したりすることで、発見モデルの精度向上を試みた。また、属性変数の取捨選択には、属性選択の手法として用いられる CFS を採用した。

属性変数の離散化　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ければならない。しかし、本研究で用意されているデータは数値化されているものが多い。数値化されたものを離散化する必要がある。本研究では、離散化の手法として、ウォード法によるクラスタリングを用いた離散化法を採用した。離散化する際の属性数は全変数において 3 もしくは 4 とし、それぞれについて発見モデルを構築した。

有向グラフ構造　本研究では、Naive Bayes 構造を仮定したものと、Free Network 構造のものを有向グラフの構造として採用した。Free Network 構造の場合は、最大の親ノード数を 3 とした。

有向グラフの学習と条件付き確率の推定　諸アルゴリズムによって、よりデータを説明し得る有向グラフを学習する。その学習アルゴリズムに、K2 アルゴリズムを採用した。またグラフ構造決定の過程において用いるモデル評価指標には AIC を採用した。

5.3.2　ベイジアンネットワークによる発見モデルの評価

本研究における発見モデルは予測モデルと換言できるが、ベイジアンネットワークが持つ他の予測モデルとは異なる特徴として、出力形式がある事象の事後確率である点が挙げられる。『要注意学生』であるかどうかの予測において、その出力結果は、「『要注意学生』である確率が 78.5 %、そうでない確率が 22.5 %」といった形式で表される。ベイジアンネットワークによる予測において、通常は事後確率が一番大きい事象が予測結果として示される。本研究の場合であれば、『要注意学生』であるかそうでないかの二値的予測であるため、『要注意学生』である確率が 50 %を越えれば予測モデルはある学生を「『要注意学生』である」とみなす。

しかし、閾値を設定することでより柔軟な予測・発見を行うことができる。例えば『要注意学生』である確率が 35 %である学生がいたとしよう。確率値が 50 %を超えていないため、通常では「『要注意学生』でない」とみなされるが、閾値を 30 %に設定することでこの学生も『要注意学生』として発見することができる。本研究では、事後確率の閾値を、50 %、30 %、事前確率に設定し、その発見精度を検証した。『要注意学生』の事前確率は、全体に対する該当学生の割合で求めることができる。対象の 307 名中、41 名が『要注意学生』であるため、事前確率は $41 \div 307 = 13.4$ % となる。

閾値を事前確率 13.4 %に設定することは、『要注意学生』であるかどうかの判別を事後確率が増加したか減少したかで判断することを意味している。ベイジアンネットワークに入力を与えていない状態では、『要注意学生』であるかどうかの確率値は 13.4 %である。ゆえに、既知情報を入力した際に、13.4 %の確率値が上昇すれば『要注意学生』であると見なされ、降下すれば『要注意学生』でないと見なされる。

また、各モデルにおける精度の評価方法は、leave one out 法とした。

5.3.3　属性変数を科目別 GPA とした『要注意学生』の発見

属性変数を科目別 GPA（前後期合わせて 14 変数）とし、ベイジアンネットワークによる発見モデルを構築した。属性変数の離散化は、GPA 値別の離散化、事例数が等しくなる等分割による離散化、ウォード法によるクラスタリングを採用した。それぞれの離散化結果は、付録 A にある表 A.1、表 A.2、表 A.3 に掲載しているので、適宜参照して頂きたい。GPA 値別の離散化のみに関して、属性数を 4 に限定した。生成されたデータ数は、GPA 値の離散化による 1 つ、ウォード法による 2 つの、計 3 つとなった。これらのデータを用いて、Naive Bayes 構造を仮定したモデルと、Free Network 構造によるモデルを構築した。

GPA 値別の離散化を施した場合のモデル

GPA 値別の離散化を施した場合の予測モデルについて述べる。Naive Bayes 構造を想定したモデルと、Free Network 構造によるモデルを構築した。Free Network 構造のモデルの構造を図 5.2 に示す。そして、閾値を 50 %、30 %、事前確率である 13.4 %にした場合の発見精度を、表 5.3 と表 5.4 に示す。

全体の正解率は、共に閾値が 50 %のものが最も高かった。しかし、Recall を比較すると、共に閾値が 13.4 %のものが最も高く、より多くの『要注意学生』を拾い上げる結果となった。Naive Bayes 構造でかつ閾値が 13.4 %である場合、307 名中 98 名を『要注意学生』と認定し、実際の『要注意学生』41 名のうち 27 名を発見することとなる。つまり、1 年次の成績中間層に潜む 66 %の『要注意学生』をするコストを、全学生の約 3 分の 1 を指導対象者とする程度に抑えられたと言える。

クラスタ数を 3 としたウォード法による離散化を施した場合のモデル

ウォード法による離散化を施した場合、各変数のクラスタ数を 3 とした予測モデルについて述べる。同様に Naive Bayes 構造を想定したモデルと、Free Network 構造によるモデルを構築した。Free Network 構造のモデルの構造を図 5.3 に示す。そして、閾値を 50 %、30 %、事前確率である 13.4 %にした場合の発見精度を、表 5.5 と表 5.6 に示す。

全体の正解率では、Free Network 構造・閾値 50 %のモデルが最も高かったが、『要注意学生』であると予測された学生は 1 人もいない結果となった。F-measure が最も高かったモデルは、Naive Bayes 構造の閾値 30 %のモデルであり、1 年次の成績中間層である 307 名のうち 81 名を指導対象とすることで、61 %の『要注意学生』を発見することができた。

クラスタ数を 4 としたウォード法による離散化を施した場合のモデル

ウォード法による離散化を施した場合、各変数のクラスタ数を 4 とした予測モデルについて述べる。Free Network 構造のモデルの構造を図 5.4 に示す。そして、閾値を 50 %、30 %、事前確率である 13.4 %にした場合の発見精度を、表 5.7 と表 5.8 に示す。

全体の正解率が最も高いのは、Free Network 構造・閾値 50 %のモデルであったが、F-measure は最も低いことが確認できた。対して最も F-measure が高かったモデルは Naive Bayes 構造・閾値 50 %のモデルであった。

また、Recall に着目すると、最もこの値が高かったモデルは、Free Network 構造・閾値 13.4 %のモデルであった。このモデルは、対象の 307 名中 112 名を指導対象者とすることで、1 年次成績中間層に潜む 68 %の『要注意学生』を発見できていた。

表 5.3: Naive Bayes 構造を想定したモデルの精度一覧（GPA 別の離散化）

閾値	Accuracy			Recall			Precision			F-measure
	対象	的中		対象	的中		対象	的中		
50 %	307	248	81 %	41	21	51 %	60	21	35 %	0.416
30 %	307	238	78 %	41	26	63 %	80	26	33 %	0.430
13.4 %	307	222	72 %	41	27	66 %	98	27	28 %	0.388

表 5.4: Free Network 構造を想定したモデルの精度一覧（GPA 別の離散化）

閾値	Accuracy			Recall			Precision			F-measure
	対象	的中		対象	的中		対象	的中		
50 %	307	263	86 %	41	4	10 %	11	4	36 %	0.154
30 %	307	253	82 %	41	13	32 %	39	13	33 %	0.325
13.4 %	307	218	71 %	41	25	61 %	98	25	26 %	0.360

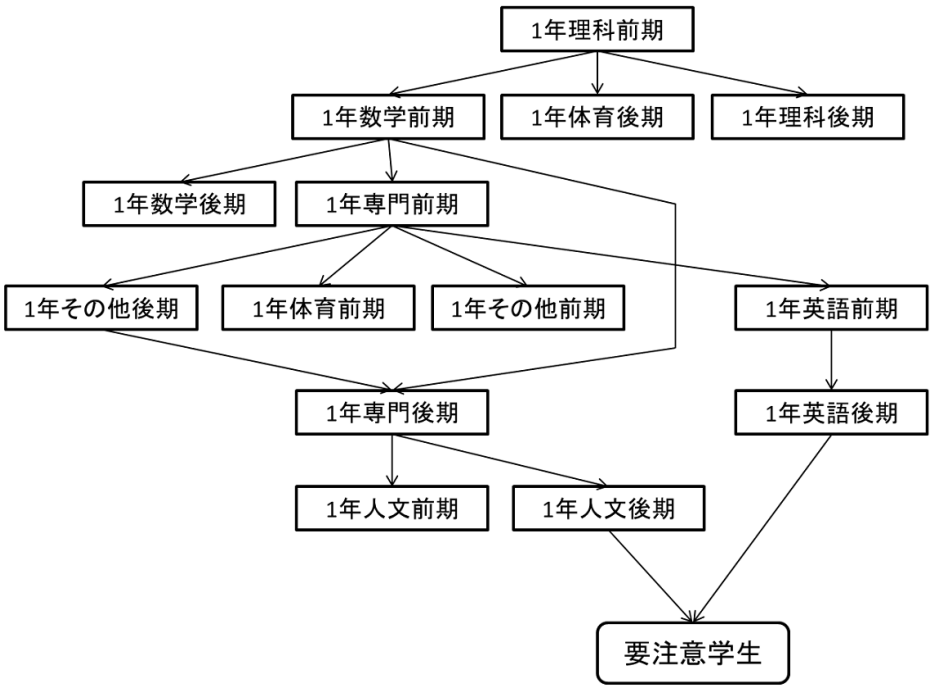

図 5.2: GPA 値別の離散化を施した場合のモデル構造

表 5.5: Naive Bayes 構造を想定したモデルの精度一覧（クラスタ数 3 のウォード法による離散化）

閾値	Accuracy			Recall			Precision			F-measure
	対象	的中		対象	的中		対象	的中		
50 %	307	246	80 %	41	21	51 %	62	21	34 %	0.408
30 %	307	235	77 %	41	25	61 %	81	25	31 %	0.410
13.4 %	307	215	70 %	41	27	66 %	105	27	26 %	0.370

表 5.6: Free Network 構造を想定したモデルの精度一覧（クラスタ数 3 のウォード法による離散化）

閾値	Accuracy			Recall			Precision			F-measure
	対象	的中		対象	的中		対象	的中		
50 %	307	258	84 %	41	0	0 %	8	0	0 %	0
30 %	307	244	79 %	41	10	24 %	42	10	24 %	0.241
13.4 %	307	206	67 %	41	22	54 %	104	22	21 %	0.303

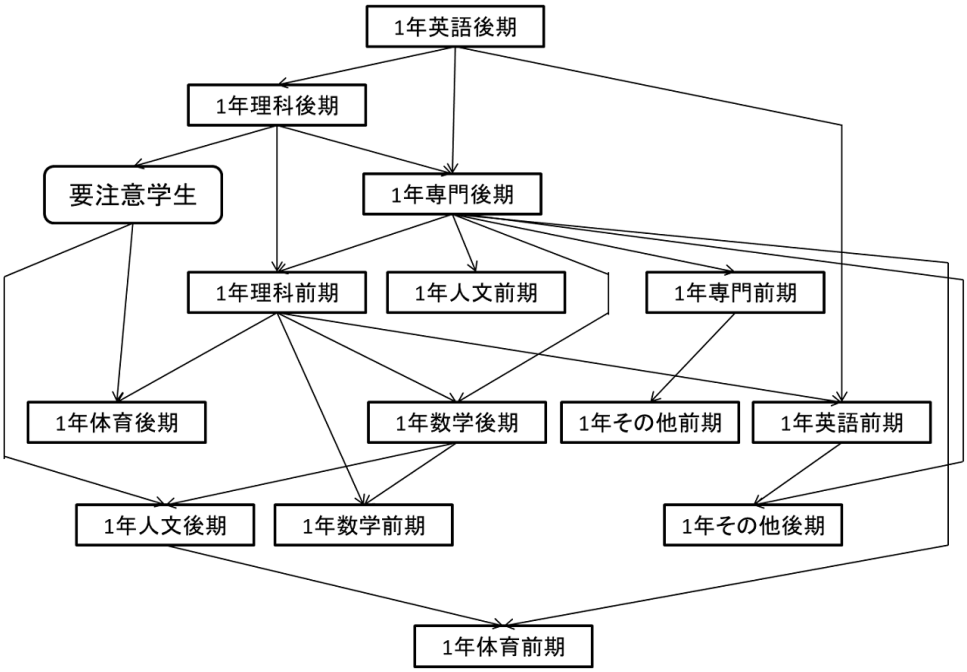

図 5.3: ウォード法による離散化を施した場合のモデル構造（クラスタ数：3）

表 5.7: Naive Bayes 構造を想定したモデルの精度一覧（クラスタ数 4 のウォード法による離散化）

閾値	Accuracy			Recall			Precision			F-measure
	対象	的中		対象	的中		対象	的中		
50 %	307	249	81 %	41	22	54 %	61	22	36 %	0.431
30 %	307	237	77 %	41	26	63 %	81	26	32 %	0.426
13.4 %	307	211	69 %	41	27	66 %	109	27	25 %	0.360

表 5.8: Free Network 構造を想定したモデルの精度一覧（クラスタ数 4 のウォード法による離散化）

閾値	Accuracy			Recall			Precision			F-measure
	対象	的中		対象	的中		対象	的中		
50 %	307	263	86 %	41	3	7 %	9	3	33 %	0.120
30 %	307	243	79 %	41	16	39 %	55	16	29 %	0.333
13.4 %	307	210	68 %	41	28	68 %	112	28	25 %	0.366

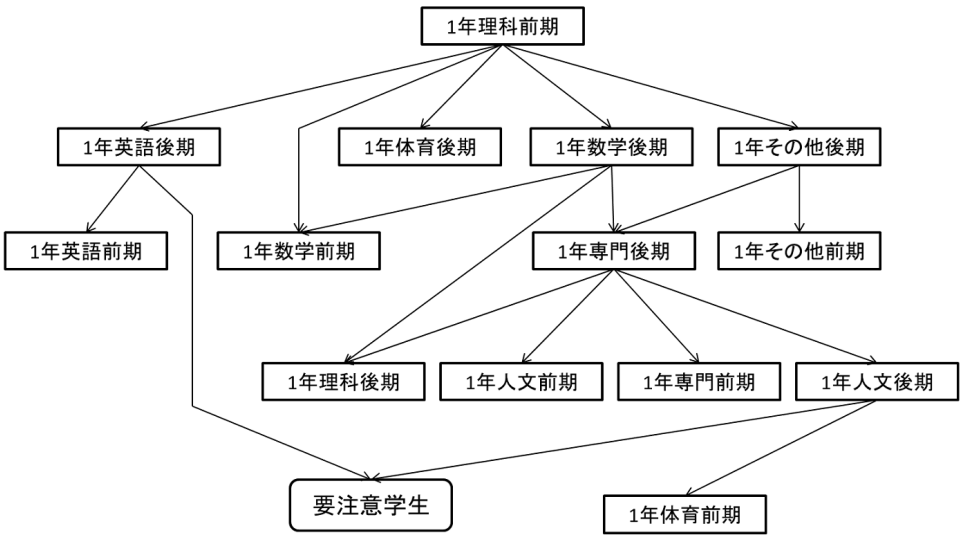

図 5.4: ウォード法による離散化を施した場合のモデル構造（クラスタ数：4）

5.3.4　全データから属性選択を行った場合の『要注意学生』の発見

前述の発見モデルは科目別 GPA のデータのみを用いてモデル構築を行った。次に、本研究における全データ（科目別 GPA、獲得成績、打刻回数データ。全 36 変数）を用いて CFS による属性選択を行い、より『要注意学生』の是非に寄与している変数を抽出しモデルの構築を試みた。

CFS による属性選択を行ったところ、14 変数が抽出された。表 5.9 に全変数を、表 5.10 にその変数郡を示す。この変数をウォード法によるクラスタリングで離散化した。離散化結果は、付録 A にある表 A.1、表 A.4、表 A.5 に掲載しているので、適宜参照して頂きたい。クラスタ数は 3 と 4 に設定し、グラフ構造は Naive Bayes 構造と Free Network 構造の 2 種類とした。ゆえに、計 4 種類の発見モデルが構築された。以下に構築されたモデルとその発見精度を示す。

表 5.9: 全変数群（36 変数）

	種類	変数
1	科目別 GPA	1 年英語前期
2	科目別 GPA	1 年英語後期
3	科目別 GPA	1 年人文前期
4	科目別 GPA	1 年人文後期
5	科目別 GPA	1 年数学前期
6	科目別 GPA	1 年数学後期
7	科目別 GPA	1 年体育前期
8	科目別 GPA	1 年体育後期
9	科目別 GPA	1 年理科前期
10	科目別 GPA	1 年理科後期
11	科目別 GPA	1 年専門前期
12	科目別 GPA	1 年専門後期
13	科目別 GPA	1 年その他前期
14	科目別 GPA	1 年その他後期
15	獲得成績数	1 年前期秀
16	獲得成績数	1 年前期優
17	獲得成績数	1 年前期良
18	獲得成績数	1 年前期可
19	獲得成績数	1 年前期不可
20	獲得成績数	1 年前期失格
21	獲得成績数	1 年後期秀
22	獲得成績数	1 年後期優
23	獲得成績数	1 年後期良
24	獲得成績数	1 年後期可
25	獲得成績数	1 年後期不可
26	獲得成績数	1 年後期失格

27	打刻回数	4 月打刻回数
28	打刻回数	5 月打刻回数
29	打刻回数	6 月打刻回数
30	打刻回数	7 月打刻回数
31	打刻回数	8 月打刻回数
32	打刻回数	9 月打刻回数
33	打刻回数	10 月打刻回数
34	打刻回数	11 月打刻回数
35	打刻回数	12 月打刻回数
36	打刻回数	1 月打刻回数

表 5.10: 指標 CFS によって抽出された変数郡

	種類	変数
1	科目別 GPA	1 年人文後期
2	科目別 GPA	1 年数学後期
3	科目別 GPA	1 年理科前期
4	科目別 GPA	1 年理科後期
5	科目別 GPA	1 年専門後期
6	獲得成績数	1 年前期不可
7	獲得成績数	1 年後期秀
8	獲得成績数	1 年後期不可
9	獲得成績数	1 年後期失格
10	打刻回数	5 月打刻回数
11	打刻回数	6 月打刻回数
12	打刻回数	7 月打刻回数
13	打刻回数	12 月打刻回数
14	打刻回数	1 月打刻回数

クラスタ数を 3 としたウォード法による離散化を施した場合のモデル

ウォード法による離散化を施した場合、各変数のクラスタ数を 3 とした予測モデルについて述べる。同様に Naive Bayes 構造を想定したモデルと、Free Network 構造のモデルを構築した。Free Network 構造のモデルの構造を 56 ページの図 5.5 に示す。そして、閾値を 50 %、30 %、事前確率である 13.4 %にした場合の発見精度を、56 ページの表 5.11 と表 5.12 に示す。

全体の正解率では、Free Network 構造・閾値 50 %のモデルが最も高かったが、Recall の結果が芳しくなかった。F-measure が最も高かったモデルは、Naive Bayes 構造の閾値 30 %のモデルであり、1 年次の成績中間層である 307 名のうち 74 名を指導対象とすることで、59 %の『要注意学生』を発見することができた。

また、科目別 GPA のみを用いた発見モデルと比較すると、各モデル構造共に、精度が改善されていることが確認できた。

クラスタ数を 4 としたウォード法による離散化を施した場合のモデル

ウォード法による離散化を施した場合、各変数のクラスタ数を 4 とした予測モデルについて述べる。Free Network 構造のモデルの構造を 52 ページの図 5.4 に示す。そして、閾値を 50 %、30 %、事前確率である 13.4 %にした場合の発見精度を、52 ページの表 5.7 と表 5.8 に示す。

全体の正解率が最も高いのは、Free Network 構造・閾値 50 %のモデルであったが、F-measure は最も低いことが確認できた。対して最も F-measure が高かったモデルは Naive Bayes 構造・閾値 50 %のモデルであった。

また、Recall に着目すると、最もこの値が高かったモデルは、Free Network 構造・閾値 13.4 %のモデルであった。このモデルは、対象の 307 名中 91 名を指導対象者とすることで、1 年次成績中間層に潜む 68 %の『要注意学生』を発見できていた。

また、科目別 GPA のみを用いた発見モデルと比較すると、クラスタ数が 3 のときと同様に Naive Bayes 構造は F-measure が改善された。しかし、Free Network 構造は各指標が悪化していることが確認できた。

表 5.11: Naive Bayes 構造を想定したモデルの精度一覧（クラスタ数 3 のウォード法による離散化）

閾値	Accuracy			Recall			Precision			F-measure
	対象	的中		対象	的中		対象	的中		
50 %	307	247	80 %	41	20	49 %	59	20	34 %	0.400
30 %	307	240	78 %	41	24	59 %	74	24	32 %	0.417
13.4 %	307	225	73 %	41	26	63 %	93	26	28 %	0.388

表 5.12: Free Network 構造を想定したモデルの精度一覧（クラスタ数 3 のウォード法による離散化）

閾値	Accuracy			Recall			Precision			F-measure
	対象	的中		対象	的中		対象	的中		
50 %	307	269	88 %	41	6	15 %	9	6	67 %	0.240
30 %	307	247	80 %	41	12	29 %	43	12	28 %	0.286
13.4 %	307	214	70 %	41	21	51 %	94	21	22 %	0.311

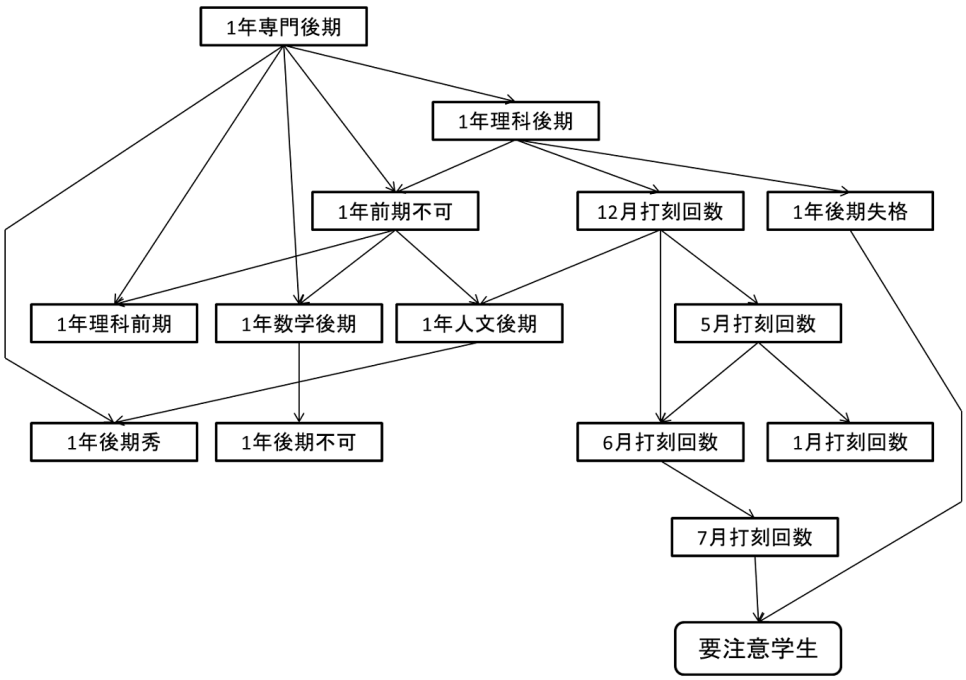

図 5.5: ウォード法による離散化を施した場合のモデル構造（クラスタ数：3）

表 5.13: Naive Bayes 構造を想定したモデルの精度一覧（クラスタ数 4 のウォード法による離散化）

閾値	Accuracy			Recall			Precision			F-measure
	対象	的中		対象	的中		対象	的中		
50 %	307	248	81 %	41	20	49 %	58	20	34 %	0.404
30 %	307	238	78 %	41	26	63 %	80	26	33 %	0.430
13.4 %	307	231	75 %	41	28	68 %	91	28	31 %	0.424

表 5.14: Free Network 構造を想定したモデルの精度一覧（クラスタ数 4 のウォード法による離散化）

閾値	Accuracy			Recall			Precision			F-measure
	対象	的中		対象	的中		対象	的中		
50 %	307	263	86 %	41	4	10 %	11	4	36 %	0.154
30 %	307	253	82 %	41	12	29 %	37	12	32 %	0.308
13.4 %	307	203	66 %	41	20	49 %	103	20	19 %	0.278

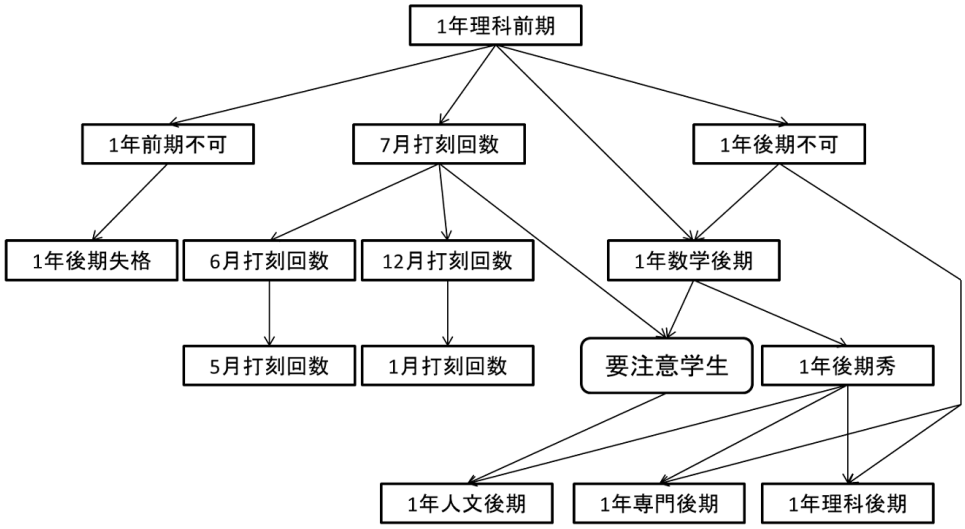

図 5.6: ウォード法による離散化を施した場合のモデル構造（クラスタ数：4）

5.4　『要注意学生』の発見モデルについての総括

本章では、発見の概要と、ベイジアンネットワークによる『要注意学生』の発見手法の提案について説明した。以下より総括を述べる。本研究では、早期に『要注意学生』を発見することを目的とし、その発見時期を 1 年次終了時とした。つまり、1 年次の情報を頼りに 2 年次以降の修学状況を予測することを試みた。

さらに、『要注意学生』の厳密な定義を、「1 年前期の GPA が 1.0 以上かつ、1 年後期の GPA が 1.0 以上であることが、将来的に留年・退学してしまう学生」とした。これにより、一見修学状況が悪化しそうにない学生であるが、実際は学習の場から脱落してしまう学生の早期発見を試みた。「1 年前期の GPA が 1.0 以上かつ、1 年後期の GPA が 1.0 以上である学生」は 307 名であり、この学生集団に含まれる『要注意学生』の人数は 41 名であった。つまり、『要注意学生』の割合は 13.4 %であった。

事前に『要注意学生』を発見する手法として、ベイジアンネットワークを採用することを提案した。ベイジアンネットワークが有する他の予測手法との差異は、出力形式が事後確率値であるため、『要注意学生』である事後確率値に閾値を設け、閾値毎の発見精度を検証した。50 %、30 %、全体の 307 名に対する『要注意学生』の割合を事前確率値とした 13.4 %をその閾値とした。確率変数に科目別 GPA データと、本研究における全データから属性選択によって抽出された変数郡を用いた。また、離散化手法にウォード法、グラフ構造に Naive Bayes 構造と Free Network 構造を採択し、多種のモデルを構築、それぞれのモデルを比較評価した。加えて、提案手法の有用性を示すため、素朴に 1 年次通年の GPA に閾値を設ける手法も検討し、両手法の比較を行った。全ての発見モデルを対象とした F-measure のランキングを 5.15 に示す。

GPA 値に閾値を設ける手法よりも、ベイジアンネットワークを用いた手法の方が高い精度を示していることが分かった。ゆえに、素朴に GPA のみを指標として修学指導の対象者を決定するよりも、本提案により指導対象者を選定した方が有用性が高いことが示された。また、全体の傾向として、Free Network 構造のモデルよりも、Naive Bayes 構造のモデルの方が各指標で良い結果を示していた。

本研究において重視された Recall に着目すると、変数を CFS により取捨選択したときのモデルが最も高い値 68 %を示していた。このときの指導対象者は 91 名であった。これだけの割合の『要注意学生』を、GPA のみを指標として発見するには、その閾値を 2.2 から 2.4 の間に設定する必要がある。この場合の指導対象者は 113 名を超えており、ベイジアンネットワークによる手法で変数を取捨選択した方がより効率良く『要注意学生』を発見できることが分かった。

このことから、様々な観点から学生の修学状況を表すことで、今後の修学状況の予測をより高精度にすることを示唆している。本研究では 3 種類のデータ、科目別 GPA データ、獲得成績数データ、打刻回数データを用いたが、他種のデータを登用することで、より有効な『要注意学生』の発

表 5.15: F-measure の値が高かったモデル（0.4 以下のものは割愛）

順位	F-measure	Recall	手法
1	0.431	54 %	NB 構造・科目別 GPA・ウォード法 4 クラスタ・閾値 50 %
2	0.430	63 %	NB 構造・変数抽出・ウォード法 4 クラスタ・閾値 30 %
3	0.430	63 %	NB 構造・科目別 GPA・GPA 別・閾値 30 %
4	0.426	63 %	NB 構造・科目別 GPA・ウォード法 4 クラスタ・閾値 30 %
5	0.424	68 %	NB 構造・変数抽出・ウォード法 4 クラスタ・閾値 13.4 %
6	0.417	59 %	NB 構造・変数抽出・ウォード法 3 クラスタ・閾値 30 %
7	0.416	51 %	NB 構造・科目別 GPA・GPA 別・閾値 50 %
8	0.410	61 %	NB 構造・科目別 GPA・ウォード法 3 クラスタ・閾値 30 %
9	0.408	51 %	NB 構造・科目別 GPA・ウォード法 3 クラスタ・閾値 50 %
10	0.404	54 %	GPA による選定・閾値 2.0
11	0.404	49 %	NB 構造・変数抽出・ウォード法 4 クラスタ・閾値 50 %
12	0.400	49 %	NB 構造・変数抽出・ウォード法 3 クラスタ・閾値 50 %

※ Naive Bayes 構造を NB 構造と表記

表 5.16: 最も良い Recall を示したモデルを用いた際の各人数

	実際に『要注意学生』である	実際に『要注意学生』でない	合計
「『要注意学生』である」と予測	58 名（TP）	64 名（FP）	122 名
「『要注意学生』でない」と予測	12 名 （FN）	204 名（TN）	216 名
合計	70 名	268 名	338 名

見が実現できると考えられる。

ここで、「NB 構造・変数抽出・ウォード法 4 クラスタ・閾値 13.4 %」のモデルを『要注意学生』の発見モデルとして採用し具体的に修学指導を行うことを想定する。このとき、「1 年前期の GPA が 1.0 より小さい、または、1 年後期の GPA が 1.0 より小さい学生」31 名と、発見モデルによって選定された 91 名の学生、計 122 名が実際の修学指導対象者となる。前者の 31 名の中には 29 名の『要注意学生』が存在し、後者の 91 名には 28 名の『要注意学生』が存在する。

つまり、全体の 338 名中、$31 + 91 = 122$ 名を『要注意学生』と認定し、残りの 216 名を指導対象外とした。さらに 338 名中に存在する本当の『要注意学生』70 名のうち、$29 + 28 = 57$ 名を『要注意学生』と指摘できた。各人数については表 5.16 に示す。したがって、本研究全体の成果（精度）は以下の式 5.4 の通りとなる。

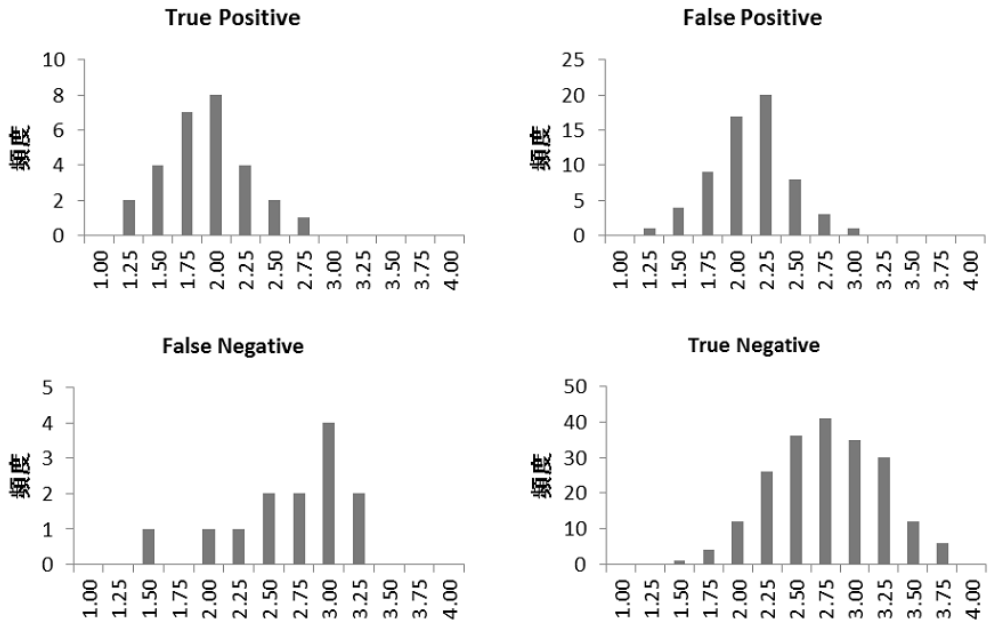

図 5.7: 各集合の GPA の分布

$$
\begin{aligned}
Accuracy &= \frac{TP+TN}{TP+FP+FN+TN} = \frac{58+204}{58+64+12+204} = 76.8\% \\
Precision &= \frac{TP}{TP+FP} = \frac{58}{58+64} = 46.7\% \\
Recall &= \frac{TP}{TP+FN} = \frac{58}{58+12} = 81.4\%
\end{aligned}
$$

つまり、本研究によって、1 年次以降に退学もしくは留年する学生の 81.4 %を全学生を指導する場合の約 3 分の 1（36 %＝ 122 ÷ 338）の時間的コストで発見できることが示された。

さらに、予測が与えられた各学生の実際の 1 年次通年の GPA を調査した。図 5.7 に、GPA のヒストグラムを示す。実際に『要注意学生』であり発見モデルによって指摘された学生（True Positive）の GPA のヒストグラムから、1 年次通年 GPA が 2.0 以上である学生の存在が確認できた。さらに、実際に『要注意学生』でなく発見モデルによって指摘されなかった学生（True Negative）の GPA のヒストグラムから、1 年次通年 GPA が 2.0 未満の学生の存在も確認できた。つまり、GPA が良くても今後の修学状況が悪化してしまう学生を拾い上げ、GPA が悪くても今後の修学状況に支障が無い学生を指導対象者と見なしていないことが分かった。ゆえに、本研究で提案した発見モデルは、単純に GPA のみを指標とした指導対象者の限定よりも、より柔軟な発見ができていると言える。

また、実際に『要注意学生』であったにも関わらず発見モデルによって指摘されなかった学生（False Negative）の GPA は、2.5 以上を上回る学生が多かった。1 年次の GPA が 2.5 以上を上回るも、2 年次以降で留年もしくは退学になるその理由について、今後追究する必要がある。

第6章　むすび

本章では、本研究の結果、今後の課題、本研究の展望と将来像について述べる。

6.1　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 338 名を対象に、在学時の各種データを用いた修学状況分析と、修学指導の対象者となり得る『要注意学生』の事前発見の手法を提案した。

修学状況の分析では、まず、修学指導の対象者となるべき学生の指摘を試みた。その結果、調査対象の 338 名の内、70 名が 1 年次以降に退学もしくは留年していることが分かった。さらに、1 年次の前期と後期の GPA について調査すると、GPA が 1.0 を下回る学生のほぼ全員が 1 年次以降に退学もしくは留年していることが分かった。この結果を受け、調査対象を 1 年前期と 1 年後期の GPA が共に 1.0 以上の学生に限定し、更なる調査を進めるためデータマイニングの手法を適用した。データマイニングの手法として、主成分分析と K-means 法におるクラスタリングを採用した。主成分分析により多くの変数が持つ情報を縮約し、クラスタリングによって学生集団を群分けし、各クラスタの傾向や退学・留年者の割合を調査した。この調査により、各観点における『要注意学生』の傾向を得ることができ、単一の観点ではなく複数の観点から学生を評価することの有用性を示すことができた。

『要注意学生』の発見では、その発見の手法としてベイジアンネットワークを提案した。本提案の有用性を示すため、GPA のみを指標とした方法と、ベイジアンネットワークを用いた手法の発見精度を比較した。さらに、用いるデータから、CFS によって変数を取捨選択し、精度の変動を検証した。その結果、GPA のみを指標とした方法よりも、ベイジアンネットワークを用いた手法の方が発見精度が良いことを示すことができた。さらに、多種の変数から CFS によって変数を取捨選択したモデルの方がより多くの『要注意学生』を発見できることも確認できた。最終的には、本研究によって構築された「NB 構造・変数抽出・ウォード法 4 クラスタ・閾値 13.4 %」のモデルを用いて、1 年次以降に退学もしくは留年する学生の 81.4 %を全学生を指導する場合の約 3 分の 1（36 %＝ 122 ÷ 338）の時間的コストで発見できることを示すことができた。

6.2　今後の課題と展望

本研究は主に、データの準備・変換・正規化、修学状況の調査・分析、『要注意学生』の定義と発見、といった構成に大別できる。各項目及びそれに関連する事柄、そして本研究の展望について述べる。

6.2.1　データの準備・変換・正規化

本研究では、講義別成績データ、打刻データ、学生修学状況データを元データとして用いた。これらデータから、各学生の GPA、獲得成績数、月別打刻数などのデータを生成した。特に GPA データ（科目別 GPA も含む）をより多く用いたが、データマイニングによる修学状況分析や発見モデルの構築の際に、獲得成績数や月別打刻数の適用にも有用性があることが確認できた。これは様々な観点から学生やその修学状況を評価することに意義があることを示唆している。つまり、データを別形式に拡張するか新たなデータ形式を導入する余地があると言える。例えば、本研究では打刻データから月別の打刻回数を算出したが、週間別の打刻回数や曜日別の打刻回数、もしくはそれらの分散を計算し新たなデータを獲得することで、学生の新たな傾向を得られる可能性がある。

データの表現方法も今後議論の対象となり得る。GPA データや打刻データは数値型の変数であったが、ベイジアンネットワークに適用する際に離散化を施す必要があった。数値型のデータであるならば数値型変数に適した手法を用いるか、より情報量を損失させない離散化手法を考慮する必要がある。

データの形式や表現方法以外に着目すると、今回は名古屋工業大学に在籍していた学生 2 年度分のデータを使用したことも今後検討すべき点だと言える。全部で 338 名分のデータが得られたが、クラスタリングやベイジアンネットワークなどの一般的傾向を抽出するような手法に適用するには、より多量のデータを用いたい。また、別の問題として、年度における修学状況の変化も挙げられる。例えばある数学系の講義において、その講義の最終成績は、教員の講義の上手さや試験の難易度に大きく左右される。今後は研究を継続させつつ年度単位でデータを増量し、年度毎の変化に影響を受けない普遍的傾向の調査や、汎用性の高い発見モデルの構築をする必要がある。

6.2.2　修学状況の調査・分析

データマイニング手法の 1 つである、主成分分析とクラスタリングを用いた分析を行った。主成分分析は情報の縮約、クラスタリングは学生の群分けに用いた。どのような手法を用いて分析を行うかは、調査の目的やデータの形式に左右されるため、その都度議論を要する。例えば、前述した打刻データを週間別打刻数データに拡張すれば、そのデータは時系列データとして扱われ

るため、Naive Bayes などの時系列データに適した手法を適用する必要がある。手法の選択について、より議論が必要だと言える。

6.2.3　『要注意学生』の定義と発見

本研究では『要注意学生』の定義を「1 年前期の GPA と 1 年後期の GPA が共に 1.0 以上であり、かつ、1 年次以降に留年もしくは退学してしまう学生」とした。しかし、何故留年もしくは退学したかにについて言及していない。退学の理由が積極的なもの（別の大学への転学など）なのか消極的なもの（大学生活に対する精神的拒絶など）なのかによって、修学指導の内容や対応を変える必要がある。より詳細のデータ獲得を目指さなければならない。

『要注意学生』の発見にはベイジアンネットワークを採用したが、ベイジアンネットワークの各種要因、確率変数の定義、有向グラフ構造の学習、条件付き確率の推定、それぞれをより検討する必要がある。有向グラフの学習、条件付き確率の推定には、より精度を高める別手法も多く報告されているため、データ形式に応じて各種調整したい。

また、本研究の結果として、Free Network 構造よりも Naive Bayes 構造の方がより高い精度を得られることが示されたが、Free Network 構造はその構造自体が確率変数間の因果関係を表しており、データマイニングの手法としても活用できる。今後はベイジアンネットワークについて調査し、その特性を最大限に活かしていきたい。

6.2.4　展望

本研究の最終目標は、実際の修学指導に提案した手法が適用されることである。そのためには、本研究による提案、『要注意学生』の発見が実用に適うかどうかを検証しなければならない。そのため、データの更なる準備が必要だと言える。目下の目標は、違う年度のデータを適用しても十分な精度を示すモデルの構築にあると言える。これが実現されれば、本研究を実用化する有用性の根拠が増すため、最終的な実用化に向けて研究を進めることができる。

謝辞

本研究を進めるにあたり、日頃から多大なご尽力を頂き、ご指導を受け賜りました名古屋工業大学 舟橋健司 准教授、伊藤宏隆 助教に深く感謝申し上げます。

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 巨樹、情報基盤センター教職員の皆様に深く感謝いたします。

そして、本研究に対し御討論、御協力頂きました名古屋工業大学中村研究室の皆様ならびに中部大学岩堀研究室の皆様に深く感謝いたします。

最後に、舟橋研究室のゼミにおいて舟橋研究室諸氏に多大な助言を頂きました。この場でお礼申し上げます。

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] 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)

[3] 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)

[4] 山本洋雄, 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”, 電子情報通信学会技術研究報告. ET, 教育工学 108 巻, 247 号, pp. 39-44 (2008)

[5] 石井一夫, “図解よくわかるデータマイニング”, 日刊工業新聞社 (2004)

[6] 元田浩, 山口高平, 津本周作, 沼尾正行, “データマイニングの基礎”, オーム社 (2006)

[7] 原圭司, 高橋健一, 上田祐彰, “ベイジアンネットワークを用いた授業アンケートからの学生行動モデルの構築と考察”, 情報処理学会論文誌, 情報処理学会論文誌 51 巻, 4 号, pp. 1215-1226 (2010)

[8] 伊藤暁人, 舟橋健司, 伊藤宏隆, “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)

[9] 寿真田崇志, 松本哲也, 大西昇, “e-Learning におけるベイジアンネットワークを用いた学習者特性の推定”, 電子情報通信学会技術研究報告.ET, 教育工学 106 巻, 583 号, pp. 203-208 (2007)

[10] 内田千代子, “大学における休・退学、留年学生に関する調査 第 31 報”, 学生の心の悩みに関する教職員研修会, 第 32 回全国大学メンタルヘルス研究会 報告書, pp. 80-94 (2011)

[11] 佐藤和彦, 大川輝人, “事前対応型の修学指導支援システムの提案”, 電子情報通信学会技術研究報告. ET, 教育工学 107 巻, 205 号, pp. 57-60 (2007)

[12] R. Kohave, G. H. John, “The Wrapper Approach”, *Feature Extraction, Construction and Selection*, pp. 33-50 (1998)

[13] 河口至商, “多変量解析入門 1”, 森北出版 (1973)

[14] J. R. Quinlan, “Induction of decision tree”, *Machine Learning*, Vol. 1, No. 1, pp. 81-106 (1986)

[15] J. R. Quinlan, “C4.5”, *Programs for Machine Learning*, Mogan Kaufmann (1986)

[16] Duan, S. and Babu, S, “Processing Forecasting Queries”, *Proc. 2007 Intl, Conf. on Very Large Data Bases*, pp. 711-722 (2007)

[17] A.K.Jain, M.N.Nurty, P.J.Flymn, “Data Clustering: A Review”, *ACM Computing Surveys*, Vol. 31, No. 3 (1999)

[18] 鈴木恒一, “大量データから知識を抽出するベイジアンネットワークの学習技術とその応用”, Softechs, L3765A, 32 巻, 1 号, pp. 14-17 (2011)

[19] 鈴木護, “ベイジアンネットワーク入門”, 培風館 (2009)

[20] 田中和之, “ベイジアンネットワークの統計的推論の数理”, コロナ社 (2009)

[21] N. Friedman, D. Geiger, M. Goldszmidt, “Bayesian Network Classifiers”, Machine Learning, pp. 131-163 (1997)

[22] M. Ramoni, P.Sebastiani,“Parameter estimation in Bayesian networks from imcomplete data”, Intelligent Data Analysis Journal, Vol. 2, No. 1 (1998)

[23] 細川和仁, “大学教育の質保障からみた GPA 精度”, 秋田大学教養基礎教育研究年報, 14 号, pp. 13-22 (2012)

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

付 録A　発見における諸情報

本研究では『要注意学生』の発見モデルを構築する過程において、科目別 GPA（14 変数）と、CFS によって取捨選択された変数群（14 変数）を用いている。また、その際、変数を離散化する工程を経ている。本節では、CFS によって取捨選択される前の変数一覧（36 変数）と、各離散化の結果を記載している。

値域の表記は、[4.00, 3.00) のようになされている。この表記が持つ意味は、「属性値が 4.0 以下であり、3.00 よりも大きい」である。

表 A.1: 科目別 GPA：GPA 値別の離散化（クラスタ数は 4）

科目	項目	1	2	3	4
英語前期	離散幅	[4.0, 3.0)	[3.0, 2.0)	[2.0, 1.0)	[1.0, 0.0]
	人数	58	169	75	5
英語後期	離散幅	[4.0, 3.0)	[3.0, 2.0)	[2.0, 1.0)	[1.0, 0.0]
	人数	69	134	91	13
人文前期	離散幅	[4.0, 3.0)	[3.0, 2.0)	[2.0, 1.0)	[1.0, 0.0]
	人数	60	135	68	44
人文後期	離散幅	[4.0, 3.0)	[3.0, 2.0)	[2.0, 1.0)	[1.0, 0.0]
	人数	55	91	98	63
数学前期	離散幅	[4.0, 3.0)	[3.0, 2.0)	[2.0, 1.0)	[1.0, 0.0]
	人数	47	108	101	51
数学後期	離散幅	[4.0, 3.0)	[3.0, 2.0)	[2.0, 1.0)	[1.0, 0.0]
	人数	40	93	101	73
体育前期	離散幅	[4.0, 3.0)	[3.0, 2.0)	[2.0, 1.0)	[1.0, 0.0]
	人数	171	111	22	3
体育後期	離散幅	[4.0, 3.0)	[3.0, 2.0)	[2.0, 1.0)	[1.0, 0.0]
	人数	169	116	17	5
理科前期	離散幅	[4.0, 3.0)	[3.0, 2.0)	[2.0, 1.0)	[1.0, 0.0]
	人数	27	90	100	90
理科後期	離散幅	[4.0, 3.0)	[3.0, 2.0)	[2.0, 1.0)	[1.0, 0.0]
	人数	25	70	122	90
専門前期	離散幅	[4.0, 3.0)	[3.0, 2.0)	[2.0, 1.0)	[1.0, 0.0]
	人数	84	171	47	5
専門後期	離散幅	[4.0, 3.0)	[3.0, 2.0)	[2.0, 1.0)	[1.0, 0.0]
	人数	52	127	113	15
その他前期	離散幅	[4.0, 3.0)	[3.0, 2.0)	[2.0, 1.0)	[1.0, 0.0]
	人数	113	175	14	5
その他後期	離散幅	[4.0, 3.0)	[3.0, 2.0)	[2.0, 1.0)	[1.0, 0.0]
	人数	88	128	84	7

※離散幅を 1.0 に設定

※全て 1 年次の GPA であるため、表記から接頭の「1 年」を省略

表 A.2: 科目別 GPA：ウォード法による離散化（クラスタ数は 3）

科目	項目	1	2	3
英語前期	離散幅	[4.00 , 2.75)	[2.75 , 2.25)	[2.25 , 0.00]
	人数	145	82	80
英語後期	離散幅	[4.00 , 3.25)	[3.25 , 2.25)	[2.25 , 0.00]
	人数	134	104	69
人文前期	離散幅	[4.00 , 3.50)	[3.50 , 2.50)	[2.50 , 0.00]
	人数	135	112	60
人文後期	離散幅	[4.00 , 2.50)	[2.50 , 1.50)	[1.50 , 0.00]
	人数	146	98	63
数学前期	離散幅	[4.00 , 2.30)	[2.30 , 1.70)	[1.70 , 0.00]
	人数	116	108	83
数学後期	離散幅	[4.00 , 2.30)	[2.30 , 0.90)	[0.90 , 0.00]
	人数	149	103	55
体育前期	離散幅	[4.00 , 3.67)	[3.67 , 2.50)	[2.50 , 0.00]
	人数	169	113	25
体育後期	離散幅	[4.00 , 3.25)	[3.25 , 2.50)	[2.50 , 0.00]
	人数	169	116	22
理科前期	離散幅	[4.00 , 2.75)	[2.75 , 1.75)	[1.75 , 0.00]
	人数	147	88	72
理科後期	離散幅	[4.00 , 2.75)	[2.75 , 1.25)	[1.25 , 0.00]
	人数	162	90	55
専門前期	離散幅	[4.00 , 2.88)	[2.88 , 2.38)	[2.38 , 0.00]
	人数	134	97	76
専門後期	離散幅	[4.00 , 2.88)	[2.88 , 1.88)	[1.88 , 0.00]
	人数	142	85	80
その他前期	離散幅	[4.00 , 3.50)	[3.50 , 2.50)	[2.50 , 0.00]
	人数	175	113	19
その他後期	離散幅	[4.00 , 3.25)	[3.25 , 2.25)	[2.25 , 0.00]
	人数	128	91	88

※全て 1 年次の GPA であるため、表記から接頭の「1 年」を省略

表 A.3: 科目別 GPA：ウォード法による離散化（クラスタ数は 4）

科目	項目	1	2	3	4
英語前期	離散幅	[4.00 , 3.25)	[3.25 , 2.75)	[2.75 , 2.25)	[2.25 , 0.00]
	人数	87	82	80	58
英語後期	離散幅	[4.00 , 3.25)	[3.25 , 2.75)	[2.75 , 2.25)	[2.25 , 0.00]
	人数	104	72	69	62
人文前期	離散幅	[4.00 , 3.50)	[3.50 , 2.50)	[2.50 , 1.50)	[1.50 , 0.00]
	人数	135	68	60	44
人文後期	離散幅	[4.00 , 3.50)	[3.50 , 2.50)	[2.50 , 1.50)	[1.50 , 0.00]
	人数	98	91	63	55
数学前期	離散幅	[4.00 , 3.10)	[3.10 , 2.30)	[2.30 , 1.70)	[1.70 , 0.00]
	人数	108	83	69	47
数学後期	離散幅	[4.00 , 2.30)	[2.30 , 1.70)	[1.70 , 0.90)	[0.90 , 0.00]
	人数	103	77	72	55
体育前期	離散幅	[4.00 , 3.67)	[3.67 , 2.50)	[2.50 , 1.50)	[1.50 , 0.00]
	人数	169	113	22	3
体育後期	離散幅	[4.00 , 3.25)	[3.25 , 2.50)	[2.50 , 1.50)	[1.50 , 0.00]
	人数	169	116	17	5
理科前期	離散幅	[4.00 , 2.75)	[2.75 , 1.75)	[1.75 , 1.25)	[1.25 , 0.00]
	人数	90	88	72	57
理科後期	離散幅	[4.00 , 2.75)	[2.75 , 1.75)	[1.75 , 1.25)	[1.25 , 0.00]
	人数	102	90	60	55
専門前期	離散幅	[4.00 , 3.38)	[3.38 , 2.88)	[2.88 , 2.38)	[2.38 , 0.00]
	人数	97	85	76	49
専門後期	離散幅	[4.00 , 2.88)	[2.88 , 2.42)	[2.42 , 1.88)	[1.88 , 0.00]
	人数	85	80	79	63
その他前期	離散幅	[4.00 , 3.50)	[3.50 , 2.50)	[2.50 , 1.50)	[1.50 , 0.00]
	人数	175	113	14	5
その他後期	離散幅	[4.00 , 3.25)	[3.25 , 2.75)	[2.75 , 2.25)	[2.25 , 0.00]
	人数	91	88	65	63

※全て 1 年次の GPA であるため、表記から接頭の「1 年」を省略

表 A.4: 変数混合：ウォード法による離散化（クラスタ数は 3）

科目	項目	1	2	3
1 年人文後期	離散幅	[4.00 , 2.50)	[2.50 , 1.50)	[1.50 , 0.00]
	人数	146	98	63
1 年数学後期	離散幅	[4.00 , 2.30)	[2.30 , 0.90)	[0.90 , 0.00]
	人数	149	103	55
1 年理科前期	離散幅	[4.00 , 2.75)	[2.75 , 1.75)	[1.75 , 0.00]
	人数	147	88	72
1 年理科後期	離散幅	[4.00 , 2.75)	[2.75 , 1.25)	[1.25 , 0.00]
	人数	162	90	55
1 年専門後期	離散幅	[4.00 , 2.88)	[2.88 , 1.88)	[1.88 , 0.00]
	人数	142	85	80
1 年前期「不可」	離散幅	[6.00 , 1.50)	[1.50 , 0.50)	[0.50 , 0.00]
	人数	196	71	40
1 年後期「秀」	離散幅	[10.00 , 4.50)	[4.50 , 1.50)	[1.50 , 0.00]
	人数	129	110	68
1 年後期「不可」	離散幅	[8.00 , 2.50)	[2.50 , 0.50)	[0.50 , 0.00]
	人数	171	77	59
1 年後期「失格」	離散幅	[6.00 , 1.50)	[1.50 , 0.50)	[0.50 , 0.00]
	人数	278	22	7
5 月打刻回数	離散幅	[109 , 82)	[82 , 71)	[71 , 0]
	人数	146	106	55
6 月打刻回数	離散幅	[108 , 77)	[77 , 63)	[63 , 0]
	人数	170	106	31
7 月打刻回数	離散幅	[102 , 74)	[74 , 63)	[63 , 0]
	人数	151	82	74
12 月打刻回数	離散幅	[103 , 65)	[65 , 58)	[58 , 0]
	人数	169	75	63
1 月打刻回数	離散幅	[129 , 81)	[81 , 62)	[62 , 0]
	人数	168	83	56

表 A.5: 変数混合：ウォード法による離散化（クラスタ数は 4）

科目	項目	1	2	3	4
1 年人文後期	離散幅	[4.00 , 2.50)	[2.50 , 1.50)	[1.50 , 0.00)	[0.00 , 0.00]
	人数	146	98	63	0
1 年数学後期	離散幅	[4.00 , 2.30)	[2.30 , 0.90)	[0.90 , 0.00)	[0.00 , 0.00]
	人数	149	103	55	0
1 年理科前期	離散幅	[4.00 , 2.75)	[2.75 , 1.75)	[1.75 , 0.00)	[0.00 , 0.00]
	人数	147	88	72	0
1 年理科後期	離散幅	[4.00 , 2.75)	[2.75 , 1.25)	[1.25 , 0.00)	[0.00 , 0.00]
	人数	162	90	55	0
1 年専門後期	離散幅	[4.00 , 2.88)	[2.88 , 1.88)	[1.88 , 0.25)	[0.25 , 0.00]
	人数	142	85	80	0
1 年前期「不可」	離散幅	[6.00 , 1.50)	[1.50 , 0.50)	[0.50 , 0.00)	[0.00 , 0.00]
	人数	196	71	40	0
1 年後期「秀」	離散幅	[10.00 , 4.50)	[4.50 , 1.50)	[1.50 , 0.00)	[0.00 , 0.00]
	人数	129	110	68	0
1 年後期「不可」	離散幅	[8.00 , 2.50)	[2.50 , 0.50)	[0.50 , 0.00)	[0.00 , 0.00]
	人数	171	77	59	0
1 年後期「失格」	離散幅	[6.00 , 1.50)	[1.50 , 0.50)	[0.50 , 0.00)	[0.00 , 0.00]
	人数	278	22	7	0
5 月打刻数	離散幅	[109 , 82)	[82 , 71)	[71 , 25)	[25 , 0.0]
	人数	146	106	55	0
6 月打刻数	離散幅	[108 , 77)	[77 , 63)	[63 , 20)	[20 , 0.0]
	人数	170	106	31	0
7 月打刻数	離散幅	[102 , 74)	[74 , 63)	[63 , 11)	[11 , 0.0]
	人数	151	82	74	0
12 月打刻数	離散幅	[103 , 65)	[65 , 58)	[58 , 10)	[10 , 0.0]
	人数	169	75	63	0
1 月打刻数	離散幅	[129 , 81)	[81 , 62)	[62 , 11)	[11 , 0.0]
	人数	168	83	56	0

発表論文リスト

投稿論文

1. H. Itoh, K. Itoh, K. Funahashi, “Forecasting Students’ Grades Using Bayesian Network Models and an Evaluation of Their Usefulness”, The Journal of Information and Systems in Education, Vol. 11, No. 1, pp. 32-41 (2013)

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

口頭発表

1. 伊藤圭佑, 伊藤宏隆, 舟橋健司, “データマイニングによる要注意学生の発見法の提案”, 人工知能学会第 69 回先進的学習科学と工学研究会資料集, pp. 35-40 (2013)

2. 伊藤宏隆, 伊藤雄真, 伊藤圭佑, 舟橋健司, “IC カード出欠データと成績データを用いた学生の成績予測”, 2013 年電子情報通信学会総合大会論文集, D-15-6 (2013)

3. 伊藤宏隆, 伊藤圭佑, 舟橋健司, “ベイジアンネットワークを用いた学生の成績予測”, 情報処理学会第 74 回全国大会講演論文集第 4 分冊, pp. 419-420 (2012)