情報処理学会論文誌　教育とコンピュータ　Vol.1　No.4　83–92 (Dec. 2015)

論　文

大学生アンケートからの文系理系学生の特徴に関する分析

野津田 雄太[1,†1]　高橋 健一[1,a)]　稲葉 通将[1,b)]

受付日 2014年9月12日，再受付日 2015年1月9日 / 2015年4月10日 / 2015年7月6日，
採録日 2015年9月5日

概要：1990 年頃から国内で理系離れが問題視されている．しかし，理系離れの要因として明確なものは存在しない．その分析に寄与するため，文系・理系の大学生の日常生活の習慣や科目の得意・不得意などの状況を調査した．本研究では，著者が所属する大学の学生を対象にアンケート調査を実施し，データマイニング手法を用いて，科目の履修・嗜好などの傾向から文系・理系への進路選択と相関のある要因と，学生の生活習慣から文系・理系学生の日常の傾向について考察を行う．用いたデータマイニング手法は決定木，主成分分析，ベイジアンネットワークである．分析の結果，小学校・中学校からの得意・不得意科目や日常生活における学生の嗜好と学生の文理に相関があることが分かった．

キーワード：知識獲得，データマイニング，決定木，主成分分析，ベイジアンネットワーク，履修科目，文系理系，生活習慣

Analysis on Features of University Students in Humanity and Science Courses from the Questionnaire

Yuta Notsuda[1,†1]　Kenichi Takahashi[1,a)]　Michimasa Inaba[1,b)]

Received: September 12, 2014, Revised: January 9, 2015/April 10, 2015/July 6, 2015,
Accepted: September 5, 2015

Abstract: From around the 1990s, undergraduate students majoring in sciences have decreased, and indifference to sciences has been spread. However, specific factors of the indifference to sciences are not clear. In order to contribute to the analysis, investigation on lifestyle and subjects studied in schools by students majoring in humanities and sciences is done. This paper describes experimental results to extract features of undergraduates majoring in humanities and sciences respectively to show factors that correlate with their course, i.e., humanities or sciences. In the research, data are collected through questionnaires to undergraduates. The questionnaire includes questions such as the lifestyle, subjects studied in schools. Then the data are analyzed by using data mining methods, namely decision trees, Principal Component Analysis, and Bayesian network. The analysis shows a correlation between students in sciences and humanities and lifestyle and a correlation between those students and subjects studied in elementary and junior high schools.

Keywords: knowledge acquisition, data mining, decision tree, principal component analysis, Bayesian network, subject, humanity and science courses, lifestyle

1. はじめに

1990 年頃より大学工学部への入学志願者が減少し，国内での理系離れが問題視されている．理系離れが進行すると，科学的思考力や計算力の低下により，特に高等教育において授業の内容を理解できない学生が増え，専門的知識・技能を有する人材の育成が難しくなることが問題としてあげられる．理系離れの一因として，学習指導要領の変遷により学習内容が希薄になったこと，自然と触れ合う機会の減少にともなう自然科学への興味の低下など，子供を取り巻く環境の変化が考えられている．理系離れの原因を探るため，理科教育に注目した報告 [1] や理科系科目に注目したモデル作成の試み [2]，理系離れ解消のための対策の提言 [3] がなされてきた．また，理系への進路のための

[1] 広島市立大学大学院情報科学研究科
Graduate School of Information Sciences, Hiroshima City University, Hiroshima 731–3194, Japan
[†1] 現在，NEC ソリューションイノベータ株式会社
Presently with NEC Solution Innovators, Ltd.
[a)] takahasi@hiroshima-cu.ac.jp
[b)] inaba@hiroshima-cu.ac.jp

特集号も発刊されている [4]．その「教育編」の記事では，理系に進ませるための子供時代の日頃の育て方や理系に進学した子の小学時代の習慣などが書かれている．さらに，理系への進学者を増やす試みとして，たとえば，大学の教員が小学校で理科の授業を行い，小学生に理科にもっと興味を持ってもらおうという活動や，理系には数少ない女子学生の獲得のために，いくつかの大学では女子学生のための説明会やトイレなどの学内施設の改装などが行われている [5]．これまで理系は文系に対して不遇であるといった社会的通念などが要因であると考えられてきたが，近年の調査でその社会的通念が変化しつつあると思われる．また，文系・理系の大学生を対象に，日常的な生活における習慣，嗜好，高校での履修科目に関してアンケート調査を行い，教科，特に数学の好き嫌い，得意・苦手が文理選択に影響を及ぼしていることが報告されている [6], [7]．さらに近年，中学校・高等学校における理系進路選択に関する研究が行われ，教科の好き嫌いと重要性の意識が影響を与えることが報告されている [8], [9]．しかし，理系離れの要因として明確なものはなく，どの活動もすぐには高い効果をあげているとはいえず，長い期間をかけて効果を検証していくことが必要である．また，理系・文系全体を対象とした包括的な研究は少ないといえる．

そこで，本研究では著者が所属する大学の学生を対象にアンケート調査を実施し，データマイニング手法を用いて，学生の生活習慣から文系・理系学生の特徴について，科目の履修・嗜好などの傾向から文理選択と相関がありそうな要因について考察を行う．対象の学生は文系学部（国際学部）および理系学部（情報科学部）に所属する学生である．これらの学生が高校で所属していた文系・理系コースにより文系・理系に分類している．文系・理系への進路を選ぶ際，得意科目・苦手科目は明示的に考慮されるが，一方で日常的な嗜好や趣味も暗黙的に進路に影響を与えていると考えられる．このことは，出前授業・実験も小学生から理系の科目に興味を持たせることで，理系の進路の可能性を上げることとも相通じるものがあると思われる．そのため，科目に関してだけでなく日常生活についてもアンケート調査を実施し，文系・理系学生の傾向を考察する．本研究では，学生の日常的な嗜好や趣味が，理系と文系学生で違いがあるか調査し，文系・理系への進路選択との相関を検討する．分析には，データマイニング手法の中でも決定木，主成分分析，ベイジアンネットワークを使用する．まず，日常生活に関するアンケート結果に対して，機械学習の分野で用いられる決定木を適用し，文系・理系の学生を分類する主要な属性を抽出し，それぞれの学生を特徴付ける要因を考察する．また，統計的手法としてよく用いられる主成分分析を適用し，寄与度の高い属性を抽出することでより特徴的な決定木の作成を試みる．さらに，ベイジアンネットワークでは確率変数の依存関係がネットワークで表現されるので，ベイジアンネットワークにより依存関係の抽出を試みる．このように本研究では，学生に対するアンケートをもとに，学生の生活習慣，履修科目を属性としたモデルの生成，文理選択の理由などに対して，機械学習の分野の手法および統計的手法を適用して分析を行う．なお，モデルの生成にはデータマイニングソフト Weka [10] を使用する．

本論文の構成は以下のとおりである．2 章では分析に用いたデータマイニング手法について簡単に説明する．3 章で具体的なアンケート項目を列挙し，4 章でデータマイニング手法による分析結果を述べる．5 章で本研究で得られた結果および今後の課題をまとめる．

2. データマイニング手法

本研究で使用するデータマイニング手法について解説する．2.1 節で決定木，2.2 節で主成分分析，2.3 節でベイジアンネットワークについて述べる．

2.1 決定木

決定木は，意思決定や物事の分類を多段階で繰り返し実行する場合，その多段の分岐過程を階層化して樹形図で表現したグラフである [11]．決定木は計算の速さ，結果の読みやすさ，説明のしやすさなどから様々な分野で応用されている．決定木生成の代表的なアルゴリズムに ID3，C4.5 などがある [12]．

本研究では，Weka の J48 を用いて決定木を生成する．J48 は，Weka における Quinlan の C4.5 アルゴリズムの実装で，枝刈りあり・なし両方の決定木を生成する [13]．C4.5 は，データを様々な条件を基準に，木の枝葉のように分類していく手法である．C4.5 は事例の集合を一括して受け取ると，事例の属性（X）とクラスに対し，各属性（x_i）で事例集合を分割した場合の平均情報量を求める．ここで，X は属性数を m とした場合の属性集合（$x_i \in X$ $(i = 1, \cdots, m)$）である．次に情報利得を求め，情報利得比が最大となる属性を，事例集合を分割するノードとして採用する．この処理を分割された事例集合に再帰的に適用することで決定木を生成する．

2.2 主成分分析

主成分分析（Principal Component Analysis: PCA）は，多数の変数で記述されるデータが観測されたとき，情報の損失をできるだけ抑えられるように変数の数を減らし，少数個の無相関な合成変数（主成分）で分析を行う手法である [14]．PCA を行うことで主成分という新しい指標が求められ，変数やデータの類似性を可視化したり，影響力の大きい変数を発見したりできる．M 個の変数 x_m を持つデータに PCA を行うと M 個の主成分が得られる．各主成分 Z_m（$m = 1, \cdots, M$）は各変数の値 x_m と各変数の重

み L_{mi}（$i = 1, \cdots, M$）の合成変数で表され，式 (1) のようになる．

$$Z_m = \sum_{i=1}^{M} L_{mi} x_i \tag{1}$$

M 個の変数の相関係数行列の固有値を $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_i \geq \cdots \geq \lambda_M \geq 0$ としたとき，λ_i に対応する固有ベクトルを重みとした合成変量が Z_i に対応し，Z_i の分散が λ_i に等しくなる．最も分散の大きい Z_1 が第 1 主成分，以下分散の大きさ順に第 2，⋯，第 M 主成分と呼ばれる．各主成分ともとの変数間の相関係数は因子負荷量と呼ばれる．因子負荷量は，第 i 主成分の各変数に対する重み $L_{i1}, L_{i2}, \cdots, L_{iM}$ に，対応する固有値の平方根をかけたものである．たとえば，Z_i と変数 x_1 の相関係数は $a_{1i} = L_{i1}\sqrt{\lambda_i}$ となる．これは主成分に強く影響を及ぼす変数を特定するための指標として扱うことができる．また，1 つの主成分がどの程度データの特徴を表しているか示す指標を寄与率と呼ぶ．寄与率は 0 以上 1 以下の値をとり，V_m（$m = 1, \cdots, M$）を主成分 Z_m の分散とすると主成分 Z_m の寄与率 C_m は式 (2) から求められる．

$$C_m = \frac{V_m}{\sum_{i=1}^{M} V_i} \tag{2}$$

寄与率は第 1 主成分が最も高く，第 2 主成分，第 3 主成分と徐々に低くなるため，上位の主成分のみを用いることで，少ない次元にデータを集約してデータの特徴を表すことができる．本研究では主成分によって学生の文系・理系の判別を試みるとともに，決定木やベイジアンネットワークを生成する際に使用する属性選択の 1 つとして主成分分析を使用する．

2.3　ベイジアンネットワーク

ベイジアンネットワークは，確率変数の依存関係をネットワークとして表した確率モデルの 1 つである [15], [16], [17], [18]．ベイジアンネットワークモデルは，確率変数，確率変数間の依存関係，確率変数間の条件付き確率，の 3 つによって定義されるネットワーク構造の確率モデルである．確率変数はノードによって表される．確率変数は目的変数と説明変数に分けられ，目的変数はベイジアンネットワークから推定したい変数，説明変数は推定するための変数となる．本研究では，目的変数が学生の文理，説明変数が各アンケート項目となる．確率変数間の依存関係は，原因から結果となる変数の向きを持つ有向リンクによって表される．確率変数 X_i，X_j 間の条件付き依存関係を $X_i \rightarrow X_j$ と表し，リンクの先にくるノードを子ノードと呼び，リンクの元にあるノードを親ノードと呼ぶ．親ノードが複数ある場合，子ノード X_j の親ノードの集合を $Pa(X_j)$ と表す．確率変数間の条件付き確率は，親ノードの値を観測したとき，子ノードのとる値の条件付き確率を列挙した表（条件付き確率表）によって表現する．また，あるノードの親ノードやそこから先の親ノードのすべてのいずれかを先祖ノードと呼ぶ．構築されるベイジアンネットワークの構造数は，確率変数の数に対して爆発的に増加する．この探索問題は NP 困難問題である．そのため，探索数を減らすための様々な構造探索手法が提案されており，一般的な山登り法やタブー探索，焼きなまし法のほかに，K2 アルゴリズム，遺伝アルゴリズムによる探索がある．

3.　アンケート調査

本研究では，モデル生成に用いるデータを収集するために，アンケート調査を実施した．本章では実施したアンケートおよびアンケート結果から抽出した属性および属性値について述べる．アンケート調査では大学 1 年生を対象とした．回答者は大学生 536 名（情報科学部：325 名，国際学部：211 名）であり，基本的に高等学校普通科の卒業生であると想定している．以下では，大学での所属学部にかかわらず，アンケートで回答した高校時での文系・理系コースをそれぞれ文系・理系として扱う．**表 1** にアンケート調査対象学生の人数の内訳を示す．

3.1　アンケート項目

アンケートでは，学生が高校のときに文系コースか理系コースのどちらに所属していたかを尋ねている．これを本分析ではクラスとして扱う．それに続くアンケートの質問項目は大きく分けて，生活習慣，履修科目，科目の嗜好などの 3 つに分類し，分析では属性として扱う．以下に，アンケート項目と図で表記する際のクラス名，属性名を括弧書きで示す．

＜クラス＞「学生が文系か理系か（学生の文理）」

＜1. 生活習慣＞　学生の生活習慣を調べるための項目であり，以下の 5 つ（●）に関する項目に分類される．

- 「日常生活」　12 属性
  「1 日の睡眠時間（睡眠時間）」,「携帯電話の使用目的（携帯電話）」,「よくするゲームの種類（ゲーム種類）」,「1 週間あたりのゲーム時間（ゲーム時間）」,「1 カ月あたりの読書数（読書）」,「テレビと動画配信（テレビ・動画）」,「1 日あたりのテレビ視聴時間（テレビ時間）」,「朝型か夜型か（朝夜型）」,「車を運転するか（車）」,「PC の所有（PC 所有）」,「1 週間あたりの PC 使用時間（PC 時間）」,「PC 使用目的（PC

表 1　アンケート対象の大学生
Table 1　Students for the questionnaire.

	2010 年度	2011 年度	合計人数
理系	151	179	330
文系	109	97	206
合計	260	276	536

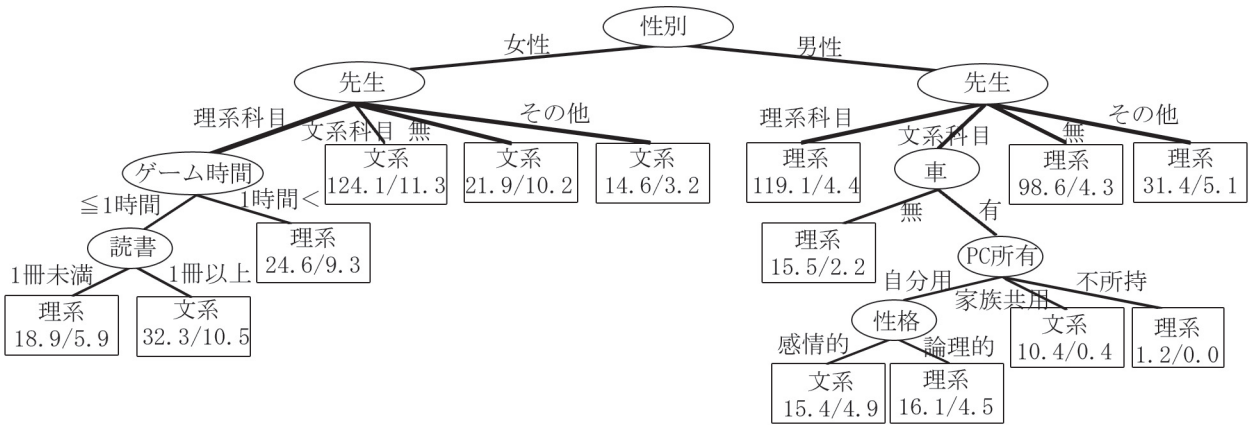


図 1　生活習慣の属性により生成した決定木
Fig. 1　A decision tree generated from the attributes relating to lifestyle.

目的)」

- 「コミュニケーション」　3 属性
  「異性と遊ぶか（異性)」,「対話が得意か（対話)」,「チャットや SNS などが得意か（チャット)」
- 「親族」　10 属性
  「祖父母と暮らしているか（祖父母)」,「親は文系か理系か（親の文理)」,「自分が何番目の兄弟姉妹か（兄弟順)」,「兄弟姉妹の人数（兄弟姉妹人数)」,「理系姉妹の有無（理系姉妹)」,「文系姉妹の有無（文系姉妹)」,「理系兄弟の有無（理系兄弟)」,「文系兄弟の有無（文系兄弟)」,「理系兄弟姉妹の有無（理系兄弟姉妹)」,「文系兄弟姉妹の有無（文系兄弟姉妹)」
- 「性格」　4 属性
  「外向的か内向的か」,「直感的か感覚的か」,「論理的か感情的か」,「計画的か柔軟か」
- 「その他」　8 属性
  「子供のころ外で遊んでいたか（外)」,「血液型（血液型)」,「高校での所属クラブ（課外)」,「好きな先生の担当科目（先生)」,「奨学金をもらっているか（奨学金)」,「進路を決めた時期（時期)」,「文系か理系かを決定した理由（文理選択理由)」,「性別（性別)」

＜ 2. 履修科目＞　高校時に履修した科目の選択傾向について調べるための項目である．科目として，普通教育に関する科目を 48 科目にまとめたものを使用した．ここでは，平成 11 年 3 月文部科学省告示の「高等学校学習指導要領（旧学習指導要領）[19]」において，「数学」「理科」の以外の教科では，I，II，III がつく科目をひとまとめに，たとえば「音楽 I」「音楽 II」「音楽 III」は「音楽」とした．

＜ 3. 科目の嗜好など＞　科目の好き・嫌い，得意・不得意，およびそれらの理由とそうなった時期である．

3.2　属性値とクラス

本研究ではアンケート項目の回答を属性値とし，「学生が文系か理系か（学生の文理)」をクラスとしてモデルを生成する．入力データは CSV 形式で保存されたファイルを使用する．また，入力値として欠損値を許す．

4.　分析結果と考察

収集したアンケートから生成したモデルなどを用いて，学生の文系・理系と相関の可能性がある要因について考察を行う．本研究では，「生活習慣」,「履修科目」,「文理選択の理由など」の 3 つの観点から分析・考察を行う．ただし，＜生活習慣＞の「その他」に含まれる属性「文系か理系かを決定した理由（文理選択理由)」は記述式であるため，内容が様々に記述されており，決定木やベイジアンネットワークの作成には使用しない．

4.1　決定木を用いた知識獲得

収集したアンケートから生成した決定木を用いて，クラス「学生が文系か理系か（学生の文理)」と相関のある要因について考察を行う．＜生活習慣＞に関するアンケート項目を属性，「学生の文理」をクラスとし，決定木を生成した．**図 1** に生成された決定木を示す．終端ノードに示される数字は，交差検定の際に正しく分類された事例数と誤分類数である．たとえば，18.9/5.9 は正しく分類された平均事例数が 18.9，誤分類の平均事例数が 5.9 であることを示している．図 1 の分類精度は 83.2%であった．図 1 から分かるように，根ノードの属性「性別」によりほとんどの事例が正しいクラスに分類されている．根ノードの次の段には属性「先生」が使用されており，「性別」が女性で「先生」が文系科目のとき文系学生の約 124 事例が，「性別」が男性で「先生」が理系科目のとき理系学生の約 120 事例が正しく分類されている．このように，「好きな先生の担当科目」の文理がそのまま「学生の文理」と一致する事例が多いことから，好きな先生もしくはその担当科目の文理は「学生の文理」と相関があると考えられる．しかし本分析で使用している属性数は，記述式の「文理選択理由」を除いた 36

個で数が多いため，不要で冗長な情報が含まれている可能性がある．そこで，属性選択を行う必要があると考え，「学生の文理」と関連の強い属性を抽出するために主成分分析を行った．

4.2　主成分分析

主成分によって「学生の文理」の判別を試みるとともに，決定木やベイジアンネットワークを生成する際に使用する属性選択の 1 つとして主成分分析を行う．また主成分の導出には統計解析アドインソフトであるエクセル統計 2010 を使用する．全事例に対して主成分分析を行った結果の第 1 から第 10 主成分までの固有値と寄与率，累積寄与率を**表 2** に示す．表 2 より，各主成分の寄与率が低く，説明能力は低いということが分かる．それゆえ，元のデータより次元を減らせるものの，主成分を用いてもまだ次元数は多い．また，第 1，第 2 主成分空間における事例の分布を**図 2** に示す．図 2 から分かるように，第 1，第 2 主成分空間では「学生の文理」の分布に偏りがあり，第 2 主成分の値の正負によりある程度判別できそうであるが，明確に区別するのは困難である．

まず，主成分分析の結果から，決定木やベイジアンネットワークを生成する際に使用する属性を選択する．因子負荷量から危険率 5%で属性選択を行った．主成分分析の結果から選択した 18 属性を**表 3** に示す．表 3 から「性別」を除いた属性を使用して生成した決定木を**図 3** に示す．終端ノードに示される数字は，交差検定の際に正しく分類された事例数と誤分類数である．ここで**表 4** は図 3 における事例の分類精度を示したものである．図 3 ではまずゲームをする時間で二分され，ゲームを週に 1 時間以上する学生は理系である可能性が高いと推定された．ゲームをあまりしない学生は性格によって分割される傾向があった．また，図 1 や図 3 などの決定木のノードとしてほとんど使われてはいないが，表 3 では「親の文理」や「理系姉妹の有無」など家族に関する属性が多く選ばれており，親兄弟の影響もあると考えられる．

次に，CFS（Correlation based Feature Selection）[20], [21] を評価指標とする最良優先探索によって，「性別」「ゲーム時間」「先生」を除いた属性から属性選択を行った．選択された 10 属性を**表 5** に示す．これらの属性を用いて，精度 66.4%で生成された決定木の主要部分を**図 4** に示す．ここでは，特に多くの事例が分類されている部分を示している．図 4 から分かるように，「論理的か感情的か」という属性でほとんどの学生が分類された．次に，図 4 から「論理的か感情的か」を除外した場合に生成された決定木の主要部分を**図 5** に示す．精度は 69.2%であった．図 5 から，「1 カ月あたりの読書数」や「PC 使用目的」のノードが複

表 2　主成分
Table 2　Principal components.

主成分	固有値	寄与率（%）	累積寄与率（%）
1	5.022	13.95	13.95
2	2.827	7.85	21.80
3	1.916	5.32	27.13
4	1.716	4.77	31.89
5	1.425	3.96	35.85
6	1.390	3.86	39.71
7	1.318	3.66	43.37
8	1.263	3.51	46.88
9	1.218	3.38	50.27
10	1.187	3.30	53.56

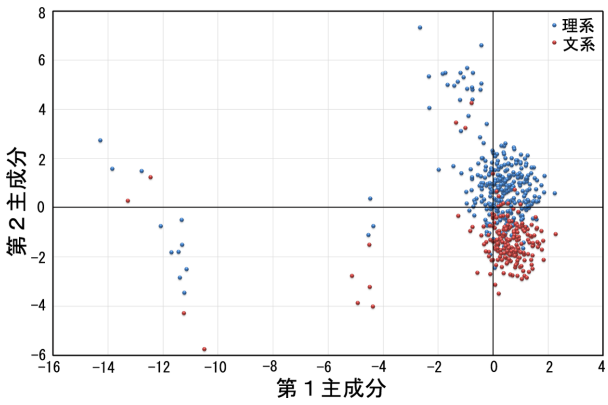


図 2　第 1 主成分，第 2 主成分の文理分布
Fig. 2　Distribution of the 1st and 2nd principal components.

表 3　主成分分析により選択した 18 属性
Table 3　18 attributes chosen with principal components.

No	属性
1	性別
2	1 週間あたりのゲーム時間
3	PC の所有
4	異性と遊ぶか
5	対話が得意か
6	チャットや SNS などが得意か
7	親は文系か理系か
8	兄弟姉妹の人数
9	理系姉妹の有無
10	文系姉妹の有無
11	理系兄弟の有無
12	文系兄弟の有無
13	祖父母と暮らしているか
14	外向的か内向的か
15	直感的か感覚的か
16	論理的か感情的か
17	計画的か柔軟か
18	進路を決めた時期

表 4　図 3 の決定木の精度
Table 4　Accuracy of the decision tree in Fig. 3.

内容	精度（%）
正しく分類された事例の割合	71.8
誤って分類された事例の割合	28.2

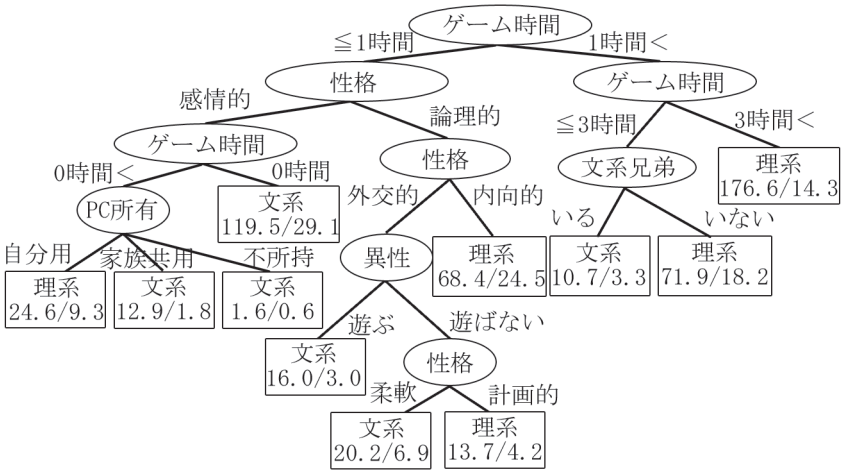


図 3　主成分分析により選択した属性を用いて生成した決定木
Fig. 3　A decision tree generated from the attributes selected by PCA.

表 5　CFS により選択した 10 属性
Table 5　10 attributes chosen with CFS.

No	属性
1	1 カ月あたりの読書数
2	朝型か夜型か
3	車を運転するか
4	1 週間あたりの PC 使用時間
5	PC 使用目的
6	奨学金をもらっているか
7	親は文系か理系か
8	文系兄弟の有無
9	外向的か内向的か
10	論理的か感情的か

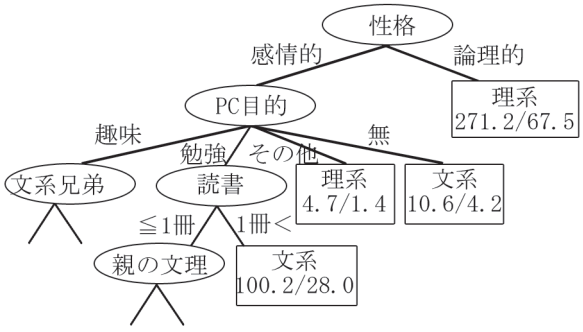


図 4　CFS により選択した属性を用いて生成した決定木の一部
Fig. 4　A part of the decision tree generated from the attributes selected by CFS.

数個表れ，それらによって多くの事例が分類される木が生成された．特に「PC 使用目的」が「趣味」の場合は理系である事例が多い．これはアンケートの対象の理系学生のほとんどが情報系学部の学生であるため，ふだんから PC を使用する頻度が高いためだと考えられる．

4.3　ベイジアンネットワーク

収集したアンケートから生成したベイジアンネットワークを用いて，学生の文理選択に影響を与える要因について考察を行う．ベイジアンネットワークの生成には Weka に実装されている山登り法を用いる．生活習慣に関する属性を用いてベイジアンネットワークを構築した場合，不要なノードや「学生の文理」のノードとはつながりのないノードが多くネットワークが複雑になった．決定木同様，属性選択を行い，選択された属性を用いてモデルの構築を行う．主成分分析の結果から選択された属性を用いて構築したベイジアンネットワークの一部を**図 6** に示す．図 6 から，「PC の所有（PC 所有）」と「進路を決めた時期（時期）」が「学生の文理」の先祖ノードとなっていることが分かる．この場合の条件付き確率表を**表 6** に示す．表 6 から次のような傾向があることが分かる．

- 自分用の PC を持っていて，進路を決めた時期が小学校の学生は理系
- PC の所有が家族共有で，進路を決めた時期が小学校または中学校の学生は文系
- PC を持っておらず，進路を決めた時期が小学校または中学校の学生は理系

次に，CFS を評価指標とする最良優先探索によって選択

表 6　主成分分析により選択された属性から得られた条件付き確率表
Table 6　Conditional probability table obtained from attributes by PCA.

PC 所有	時期	理系の確率	文系の確率
自分用	高校	0.63	0.37
自分用	中学校	0.67	0.33
自分用	小学校	0.87	0.13
家族共用	高校	0.58	0.42
家族共用	中学校	0.25	0.75
家族共用	小学校	0.23	0.77
不所有	高校	0.46	0.54
不所有	中学校	0.83	0.17
不所有	小学校	0.75	0.25

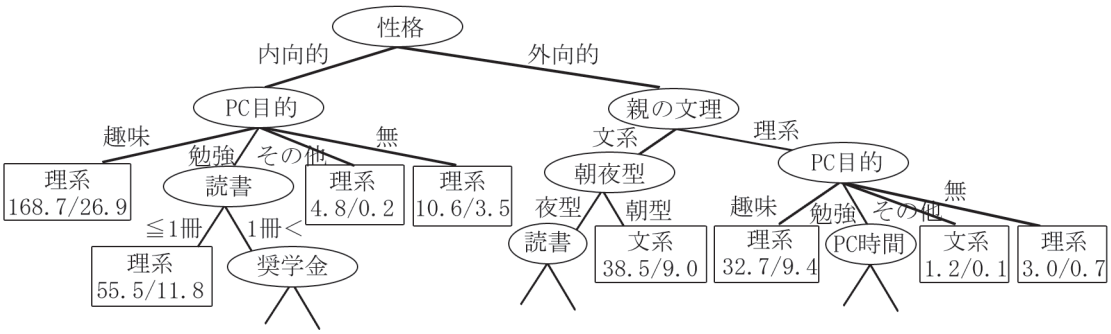


図 5　図 4 の根ノードの属性を除いて生成した決定木の一部

Fig. 5　A part of a decision tree generated from the attributes in Fig. 4 excluding the root node.

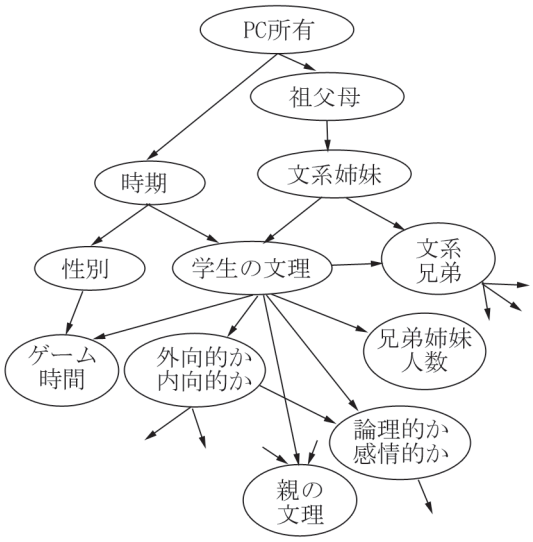


図 6　主成分分析により選択された属性を用いて生成されたベイジアンネットワークの一部

Fig. 6　A part of Bayesian network generated from the attributes selected by PCA.

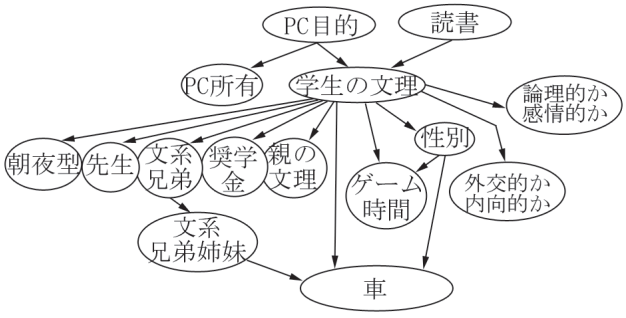


図 7　CFS により選択された属性を用いて生成されたベイジアンネットワーク

Fig. 7　Bayesian network generated from the attributes selected by CFS.

された属性を用いてベイジアンネットワークを構築した．構築したベイジアンネットワークを**図 7** に示す．図 7 より，「PC 目的」と「読書」が「学生の文理」の親ノードとなった．このことから，パソコンや本といった嗜好品に対する興味や関心が，学生の文理選択と相関のある要因の 1 つとして考えられる．また，得られた条件付き確率表を**表 7** に示す．表 7 から次のような相関があることが分かる．

表 7　CFS による属性から得られた条件付き確率表

Table 7　Conditional probability table obtained from attributes by CFS.

PC 目的	読書	理系の確率	文系の確率
趣味	1 冊未満	0.83	0.17
趣味	1 冊以上	0.77	0.24
勉強	1 冊未満	0.67	0.33
勉強	1 冊以上	0.40	0.60
その他	1 冊未満	0.50	0.50
その他	1 冊以上	0.70	0.30
不使用	1 冊未満	0.85	0.15
不使用	1 冊以上	0.35	0.65

- 「PC 使用目的（PC 目的）」が「趣味」は理系
- 「PC 使用目的（PC 目的）」が「勉強」で，本をあまり読まない学生は理系，本を読む学生は文系
- PC を使用せず，本をあまり読まない学生は理系，本を読む学生は文系

「PC 目的」や「読書」は決定木から得られた特徴にも表れており，「学生の文理」を識別するのに重要な属性と考えられる．しかし，その他の特徴に関しては決定木とベイジアンネットワークの 2 手法での共通性はみられなかった．また，図 1，図 3 の決定木でノードに使用された属性，たとえば，「ゲーム時間」，「文系兄弟」，「車」，「外向的か内向的か」が，ベイジアンネットワークを用いた場合では図 7 のように「学生の文理」の子ノードとなっている場合が多く，因果関係が正しく表されていない可能性がある．そのため，文理選択に影響を与えている要素についてはさらなる検討が望まれる．

4.4　履修科目に基づく分析

高校での履修科目が文系・理系学生でどのような違いがあり文理選択にどのような影響を与えているかを調べるた

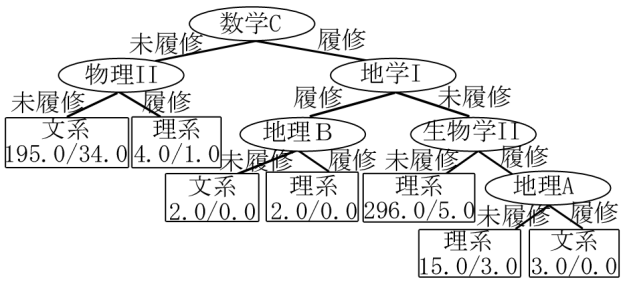


図 8　全履修科目により生成した決定木

Fig. 8　A decision tree generated from all subjects.

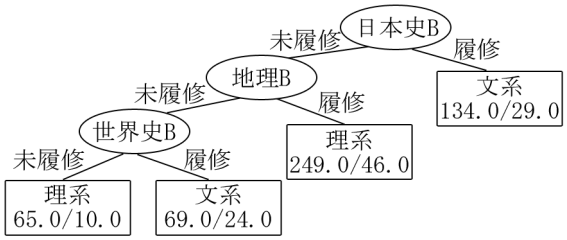


図 9　社会科の科目により生成した決定木

Fig. 9　A decision tree generated from social studies.

表 8　図 9 の決定木の精度

Table 8　Accuracy of the decision tree in Fig. 9.

内容	精度
正しく分類された事例の割合	77.8%
誤って分類された事例の割合	22.2%

めに，高校における全履修科目を属性として決定木を生成した．なお，以下ではアンケートの回答における履修状況のみを取り上げ，学習指導要領における必履修などを考慮せずに分析している．**図 8** に生成された決定木を示す．ほとんどの事例が，「数学 C」と「物理 II」を履修しなければ文系，「数学 C」を履修して「地学 I」と「生物 II」を履修しなければ理系と分類される木が生成された．しかし「数学 C」と「物理 II」はほとんどの文系学部の受験には必要がない．また，「地学 I」と「生物 II」を履修しないで「物理」か「化学」を履修していることは本学の入試科目からも，また多くの理系学部の受験で前提となっているため，当然の結果といえる．

そこで，文理共通の社会科（世界史 A，B，日本史 A,B，地理 A，B，現代社会，倫理，政治・経済）の科目で決定木を生成した．生成された決定木を**図 9** に示す．ここで**表 8** は図 9 における事例の分類精度を示したものである．図 9 の決定木から，「日本史 B」を履修する学生は文系である傾向が，また，「日本史 B」を履修せず，「地理 B」を履修する学生は理系である傾向がみられる．「世界史 B」については，文系・理系学生ともに同程度の人数が履修しており，文理の差が出ていない．本学の理科の入試では，物理もしくは化学を受験する必要があり，文献 [22] に述べられているように，「地理をとっている人は物理・化学それから物理だけとか，あるいは化学・生物，化学だけとか，そういう物理・化学系をとっている人が多い」ことがあてはまっている．

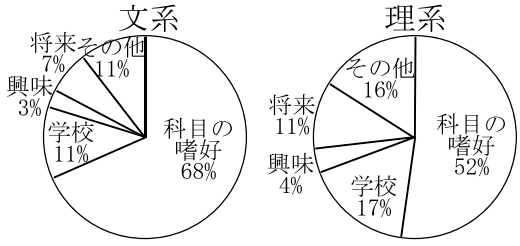


図 10　文系・理系選択の理由

Fig. 10　Reasons for selection of courses.

表 9　「科目の嗜好」の内訳

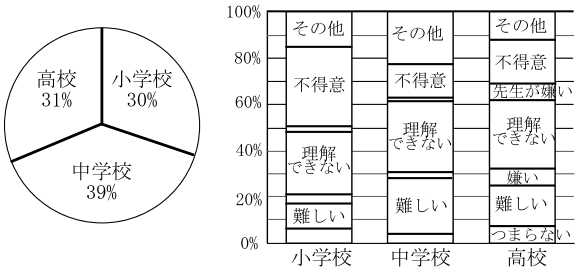
Table 9　Detailed reasons for selection of courses.

嗜好	文系（%）	理系（%）
得意・好き	26.1	56.8
不得意・嫌い	73.9	43.2

4.5　文理選択の理由に基づく分析

最後に，文系か理系かを決定した文理選択理由について分析を行った．アンケートから得た文理選択の理由は記述式のため，そのままでは扱うのが困難である．そのため，属性「文理選択理由」の属性値から，文理選択の理由を「科目の嗜好など」，「学校」，「興味」，「将来」，「その他」の 5 つに分類した．「科目の嗜好など」は科目の好き・嫌い，得意・不得意を，「学校」は学校のコース分けや先生を，「興味」は文系・理系に関係した興味を，「将来」は将来やりたいことやなりたい職業を理由としている場合である．「その他」は前述の 4 つ以外の理由（「なんとなく」や「分からない」など）である．**図 10** に文系・理系それぞれの結果を示す．

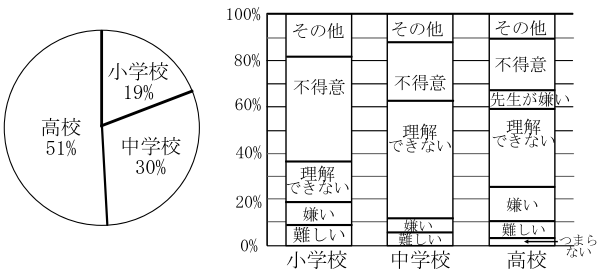
図 10 より，文系，理系ともに「科目の嗜好」を文理選択の理由とする学生の割合が最も多いことが分かる．文系の学生の約 7 割が「科目の嗜好」を文理選択の理由としている．さらに，「科目の嗜好」の内容として，科目が得意または好きで文理を選択したのか，不得意または嫌いで文理を選択したのかをその科目名とともにアンケートに記述してもらった．**表 9** に，文理選択の理由を「科目の嗜好」とした学生のうち，科目が得意・好き，不得意・嫌いかを理由としている学生の割合を示す．表 9 から，「科目の嗜好」を理由とする文系学生の約 74%が「科目が不得意・嫌い」という消極的な理由を文理選択の理由としていることが分かる．文系学生の「不得意・嫌い」とする科目として記述されていたのは，「理系科目」，「数学」であった．一方，理系学生では，約 57%が「理系科目」，「数学」が「得意・好き」を理由としていた．このことから，理系科目が得意・好きである学生を増やすことで，理系に進む学生数に好影



(a) 時期　(b) 理由

図 11　文系学生が数学を不得意・嫌いになった時期と理由

Fig. 11　Time when students are weak or dislike Mathematics and its reasons.



(a) 時期　(b) 理由

図 12　文系学生が理科を不得意・嫌いになった時期と理由

Fig. 12　Time when students are weak or dislike Science and its reasons.

響を与える可能性が考えられる．

まず，文系と理系の学生で特に嗜好に差のある「数学」に関して，不得意，嫌いになった時期および理由について考察を行う．文系学生全体のうち，数学が不得意・嫌いと回答した学生は延べ 280 人である．複数回答のため文系実学生数より多くなっている．数学が不得意・嫌いになった時期（小学校・中学校・高校）ごとの学生の割合および理由を**図 11** に示す．数学が嫌いになった時期は，中学校が 39%と最も多く，小学校が 30%，高校が 31%とほぼ同数であり，小学校の割合も少なくない．特に数学に関しては小中学校の時期から不得意・嫌いになる学生が多い傾向がみられた．数学が不得意，嫌いな理由として，小学校や中学校など早い時期から「難しい」,「理解できない」と回答する学生が多いことが分かった．「理解できない」の割合は，小学校では約 25%であったが，中学校では約 30%に増加している．「不得意」の割合は小学校の約 30%が中学校で約 15%に半減している一方，「難しい」の割合は小学校の約 10%から中学校で約 20%に倍増している．高校に至っては，「先生が嫌い」「嫌い」「つまらない」が増え，数学への興味や関心が損なわれていることがうかがえる．中学校の「理解できない」「難しい」の割合が増えている理由として，中学校からの数学では解に至るまでの過程，物事を論理的に考えることが重視され，小学校の算数との間にギャップがあるためと考えられる．

次に，「理科」に関して，不得意，嫌いになった時期および理由について考察を行う．文系学生のうち，理科が不得意・嫌いと回答した学生は比較的少なく，延べ 66 人であった．理科が不得意・嫌いになった時期（小学校・中学校・高校）ごとの学生の割合および理由を**図 12** に示す．理科を不得意・嫌いになった時期は，数学に比べて，小学校では 19%と比較的少なく，高校では 51%とかなり多くなっている．また，中学校から「理解できない」と回答している学生の割合が，小学校では約 20%であったが，中学校では約 50%に急増している．中学校からの理科では科学的な概念や思考力といったより抽象化された内容が入ってきて難しくなるためと考えられる．

5. おわりに

本研究では，大学生を対象にアンケート調査を実施し，データマイニング手法を用いて，学生の生活習慣，科目の履修・嗜好などの傾向から文系理系と相関のある要因について考察を行った．生成したモデルから文系・理系を選択した学生間で生活習慣および履修科目に差がみられた．

本論文では，生活習慣に関する属性を用いて生成した決定木やベイジアンネットワークから，大学生の文理それぞれでの特徴をあげた．決定木を用いた場合，主に「好きな先生の担当科目」や「1 週間あたりのゲーム時間」，学生の性格に関する属性などが学生の文理と相関があることが分かった．ベイジアンネットワークを用いた場合，パソコンや本など学生の嗜好が文理と相関を示す結果となった．

履修科目に基づく分析では，文理選択の理由として科目の嗜好，特に数学に関しての嗜好が顕著に表れる結果となった．これは，文系学生は小学校，中学校など早い時期から数学が不得意または嫌いになる学生が多いことが原因だと考えられる．したがって，理系離れの対策として小学校などの初等教育で数学を不得意にさせない授業展開・指導案が必要であると考えられる．

本分析で使用したデータは，収集地域，アンケート調査対象が 1 大学の学生（情報科学部と国際学部）と限定的である．得られた結果をより一般的なものとするため，今後の課題として，データ収集範囲の拡大，アンケート実施対象の検討があげられる．また，文理の識別をより正確に明示する属性を発見するために，文系的な属性をよりバランスよく含むなどアンケートの設問に関して熟考する必要があると考えられる．

謝辞　本研究において，アンケート収集などにご協力いただいた方々，また本論文に関し貴重なご示唆，ご意見をいただいた査読者の方々に感謝いたします．なお，本研究は，一部広島市立大学特定研究（一般研究 No.0216）および JSPS 科研費 24501141 の助成を受けた．

参考文献

[1] 鶴岡森昭，永田敏夫，細川敏幸，小野寺彰：大学・高校理科教育の危機—高校における理科離れの実状，高等教育ジャーナル（北大），Vol.1, pp.105–115 (1996).
[2] 斉藤浩一，高橋郷史：「理科離れ」の原因帰属に関するモデル作成の試み—高校生の意識調査をもとに，東京情報大学研究論集，Vol.9. No.1, pp.1–9 (2005).
[3] 増田貴司：「理科離れ」解消のために何が必要か，TBR 産業経済の論点，No.07-06, 株式会社東レ経営研究所 (2007).
[4] 週刊東洋経済，理数力で決める!学校&就職，Vol.6315, pp.38–89 (2011).
[5] たとえば，朝日新聞記事
「科学の出前」受講 3 万人 (2014.06.26).
理科って楽しい (2014.08.01).
女子学生獲得に大学が力 (2014.02.26).
4 女子大「理系に興味を」(2012.08.03).
大学が「リケジョ」獲得作戦 (2010.10.04).
[6] 野津田雄太，高橋健一：決定木を用いた学生の文理選択に関するアンケートからの知識獲得，電子情報通信学会技術研究報告，AI，人工知能と知識処理，Vol.111, No.310, pp.7–12 (2011).
[7] 野津田雄太，高橋健一，稲葉通将：学生の文理選択に関するアンケートからの知識獲得，電子情報通信学会技術研究報告，AI，人工知能と知識処理，Vol.112, No.435, pp.17–22 (2013).
[8] 特集「文理選択と大学入試」，Guideline, 河合塾 (2013.11), 入手先 ⟨http://www.keinet.ne.jp/gl/13/11/toku_1311.pdf⟩.
[9] 後藤顕一：中学校・高等学校における理系進路選択に関する研究，国立教育政策研究所 (2013.03).
[10] Weka3 – Data Mining with Open Source Machine Learning Software in Java, 入手先 ⟨http://www.cs.waikato.ac.nz/ml/index.html⟩.
[11] 元田　浩，津本周作，山口高平，沼尾正行：データマイニングの基礎，オーム社 (2006).
[12] Quinlan, J.R.: *C4.5: Programs for machine learning*, Morgan Kaufmann Publishers (1993).
[13] J48, 入手先 ⟨http://www.opentox.org/dev/documentation/components/j48⟩.
[14] R と主成分分析，入手先 ⟨https://www1.doshisha.ac.jp/~mjin/R/24.pdf⟩.
[15] 芳賀麻誉美，本村陽一：ベイジアンネットワークの確率推論による商品開発とマーケティング戦略—バニラカップアイスの設計と意思決定支援への適応を通して，人工知能学会人工知能基本問題研究会資料，Vol.60, pp.59–64 (2005).
[16] 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，ET2006-141, pp.203–208 (2007).
[17] 本村陽一：ベイジアンネットワークの基礎と応用における新展開，人工知能学会誌，Vol.22, No.3, pp.302–305 (2007).
[18] 原　圭司，高橋健一，上田祐彰：ベイジアンネットワークを用いた授業アンケートからの学生行動モデルの構築と考察，情報処理学会論文誌，Vol.51, No.4, pp.1215–1226 (2010).
[19] 文部科学省高等学校学習指導要領（旧学生指導要領），入手先 ⟨http://www.mext.go.jp/a_menu/shotou/cs/1320224.htm⟩.
[20] Hall, M.A. and Smith, L.A.: Feature Selection for Machine Learning: Comparing a Correlation-based Filter Approach to the Wrapper, *The 12th AAAI International Florida Artificial Intelligence Research Society Conference*, May 1–5, 1999, Orlando, Florida, USA, pp.235-239 (1999).
[21] Duan, S. and Babu, S.: Processing Forecasting Queries, *VLDB*, pp.711–722 (2007).
[22] 大学入試センター：大学入試研究の動向，p.110, 右側上から 4 段落，No.28 (2011).

野津田 雄太

2013 年広島市立大学大学院情報科学研究科博士前期課程修了．在学中，データマイニング，知識獲得の研究に従事．現在，NEC ソリューションイノベータ株式会社勤務．

高橋 健一　（正会員）

1979 年名古屋工業大学大学院工学研究科修士課程修了．同年名古屋工業大学工学部助手．同大学講師，助教授を経て，1994 年広島市立大学情報科学部教授．現在，同大学大学院情報科学研究科所属．知識処理，エージェント，e ラーニング等の研究に従事．工学博士．IEEE，電子情報通信学会各会員．

稲葉 通将　（正会員）

2012 年名古屋大学大学院情報科学研究科博士後期課程修了．同年広島市立大学大学院情報科学研究科助教，現在に至る．対話システム，対話処理に関する研究に従事．博士（情報科学）．電子情報通信学会，人工知能学会各会員．