# 決定木を用いた
# 文理選択に関する知識発見
# Knowledge Discovery on Selection of Humanities and Sciences Using Decision Tree

野津田　雄太
Yuta Notsuda
広島市立大学大学大学院
情報科学研究科
Email: notsuda@cm.info.hiroshima-cu.ac.jp

髙橋　健一
Kenichi Takahashi
広島市立大学大学大学院
情報科学研究科
Email: takahasi@hiroshima-cu.ac.jp

***Abstract*— This paper describes experimental results to extract features of students majoring in humanities and sciences respectively, in order to show factors that influence the selection when students select their course, i.e. humanities or sciences. In the research, we collect data through questionnaires to students. The questionnaire includes questions such as the lifestyle, subjects studied in schools. Then we analyze the data using the decision tree.**

## Ⅰ. はじめに

近年，国内で理系離れが問題視されている．理系離れとは理科離れとも言い，理科に対する生徒・児童の興味・関心の低下，授業における理解力の低下，日常生活において必要な基礎的な科学的知識を持たない人々が増加しているなどの状態を指す言葉である．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指摘されている[1].

現在，理系離れの要因として明確なものは存在しない．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伴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．また，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いる．しかし，近年の調査で理系出身者は文系出身者に比べ平均収入が多いことが報告されており，社会的通念が変化しつつあると思われる．

理系離れに対する様々な対策や議論[1]が行われている．例えば，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が行われている．いくつかの大学では，理系には数少ない女子学生の獲得のために，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を行っている．しかし，どの活動も高い効果を上げているとは言えない．また，物理や化学など各科目での研究はあるものの，理系・文系全体を対象とした包括的な研究は少ないと言える．

そこで，本研究では高校生・大学生を対象にアンケートを実施し，決定木学習を用いて，学生の生活習慣，科目の履修・嗜好等の傾向から文理選択に影響を与える要因を考察する．決定木とはデータマイニングで良く用いられるモデルであり，意思決定や物事の分類を多階層で繰り返し実行する場合，その分岐の繰り返しを階層化して樹形図で表現したグラフである[2].

本研究では，学生に対するアンケートをもとに，高校生・大学生それぞれにおいて，生活習慣，履修科目を属性とした決定木生成，文理選択の理由等に対して統計をとることによって分析を行う．

## Ⅱ. 決定木

決定木とは，データマイニング手法のひとつである．データマイニングとは，蓄積された膨大な量のデータから，意味のあるパターンやルールを発見する技術のことである．データを一見しただけでは想像が及びにくい，発見的な知識獲得が可能であるという期待を含意している．1990 年代頃，デジタル社会の発展に伴い，年々増加する膨大な量のデータを処理するための手法としてデータマイニングの概念が現れた．現在では，小売店の販売データや電話の通話履歴やクレジットカードの利用履歴などの大量に蓄積されるデータを解析し，その中に潜む項目間の相関関係やパターンなどの発見に用いられている．

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． 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．代表的なアルゴリズムとして ID3, C4.5 などがある．

本研究では，決定木の生成に Weka[3]というデータマイニングソフトを使用する．Weka とは様々な分類器

を実装したデータマイニングソフトであり，機械学習の研究と普及の為に開発されているフリーのソフトウェアである．Weka による決定木の生成は J48 アルゴリズムを用いて行う．J48 アルゴリズムは，C4.5 アルゴリズムを Weka に実装したものである．C4.5 アルゴリズムとは，情報利得比が最大になる属性を木のノードとして採用する操作を再帰的に行うことで木を生成するアリゴリズムである．式(1)に情報利得比 $(Gain\ Ratio)$の計算を示す．

$$Gain\ Ratio(x_i) = \frac{Gain(x_i)}{Split\ Info(x_i)} \qquad (1)$$

ここで$x_i$は事例に対する属性の集合$X$の各属性であり，情報利得比は情報利得$(Gain)$と分割情報量$(Split\ Info)$の商で表される．式(1)により，全ての属性に対し情報利得比を求め，値が最大となる属性$x_i$を木のノードに採用する．また，本研究では，生成された決定木は 10-交差検定法を用いて評価する．

## Ⅲ．入力データ

本研究では，決定木生成に用いるデータを収集する為に，アンケートを実施した．本章では実施したアンケート及び，アンケート結果から抽出した属性，属性値について述べる．

### A. アンケート

まず学生に対してアンケートを実施した．回答者は本学学生 260 名(理系:151，文系:109)と広島市内の高校生 110 名(理系:34，文系:76)である．

アンケートの質問項目は大きく分けて以下の 3 つに分類される．

＜1.生活習慣＞

学生の生活習慣が文理選択に影響を与えているか調べるための項目である．大まかに分類すると日常生活，コミュニケーション，親族，性格，その他の 5 つに関する項目に分類される．以下にそれらの説明を示す．

- 「日常生活」
  睡眠時間，テレビ視聴時間，PC 使用時間，朝型か夜型かなど 12 属性
- 「コミュニケーション」
  異性と遊ぶか，対話が得意か，Chat などが得意かの 3 属性
- 「親族」
  親の文理，兄弟の文理など 10 属性
- 「性格」
  社交性，感性，理性，計画に関する 4 属性
- 「その他」
  部活動，血液型など 6 属性

＜2.履修科目＞

高校時に履修する科目の選択傾向について調べるための項目である．科目の種類は文部科学省の「高等学校学習指導要領」に基づく 48 科目である．

＜3.科目の嗜好等＞

科目の好き・嫌い，得意・不得意，及びそれらの理由等を調べるための項目である．

### B. 属性値とクラス

本研究ではアンケートの各設問を属性，回答結果を属性値とし，文理選択をクラスとして決定木を生成する．入力データは CSV 形式で保存されたファイルを使用する．

## Ⅳ. 実験・考察

収集したアンケートから生成した決定木等を用いて，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．本研究では，「生活習慣」，「履修科目」，「文理選択の理由等」の 3 つの観点から分析・考察を行う．

### A. 生活習慣に基づく決定木

生活習慣に関するアンケート項目を属性，文理選択をクラスとし，J48[2]を用いて決定木を生成した．使用した属性は 35 個である．

大学生に対するアンケートから生成された決定木を図 1 に示す．終端ノードの括弧内に示される数字は，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．図 1 より，ゲームをする学生は理系である傾向が見られる．また，文系学生は睡眠時間が少なく，理系学生は本を読まないといった特徴がある．さらに，文系学生は外向的であり，理系学生は内向的であるという結果が伺える．

次に高校生に対するアンケートから生成された決定木を図 2 に示す．図 2 より，睡眠時間の多い学生は理系である傾向が見られる．また，親の文理で多くの事例が分類されているので，親の影響が強いと考えられる．一部の理系学生において PC の使用時間が著しく多いといった結果も見られる．ここで表 1 は図 1，図 2 における事例の割合を示したものである．どちらの決定木も 7 割弱の精度を示している．

表 1　決定木の精度

| | 大学生 | 高校生 |
|---|---|---|
| 正しく分類された事例の割合 | 68.5% | 66.4% |
| 誤って分類された事例の割合 | 31.5% | 33.6% |

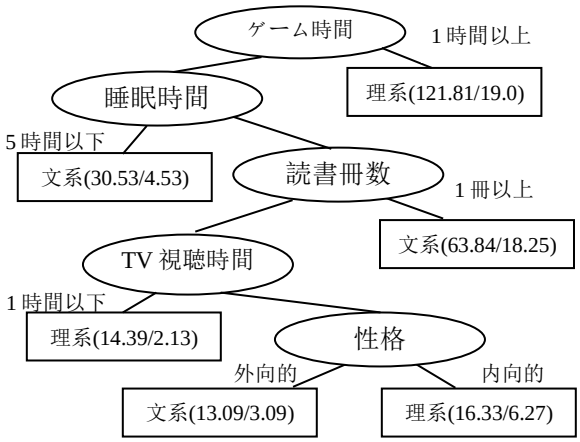


図 1　大学生データから生成した決定木

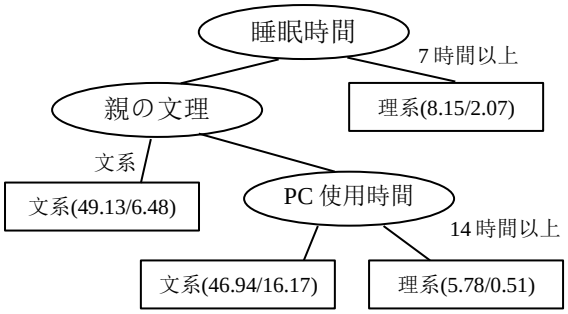


図 2　高校生データから生成した決定木

高校生と大学生の結果を比較してみると，理系学生は睡眠時間が多く，文系学生は少ない傾向が見られる．また，理系学生はテレビや本を見ない代わりにゲームやパソコンを使用する時間が多いということがわかった．しかし，全体的に共通した部分は少なく，このことから文理選択の基準は年代によって多種多様であり，本実験で得られた結果が全ての年代において共通であると解釈することはできない．

また，性別を属性に加えると，男性なら理系，女性なら文系といったようにほとんどの事例が男女の違いのみで文理を識別される結果となった．これは女性研究者に対する社会的通念などが影響しているものと考える．

B. 履修傾向に基づく決定木

次に，履修科目を属性として決定木を生成し，分析した．本実験では文理選択及び履修科目が確定した大学生に対するアンケート結果を使用している．その結果，「数学 C」の履修の有無のみによって文理を識別する木が生成された．さらに「数学 C」を属性から除外して決定木を生成した結果，「数学Ⅲ」，「化学Ⅱ」，「物理Ⅱ」の履修の有無で文理を識別する木が生成された．しかし多くの理系学部の受験生は，これらの科目を履修していることが前提となっているため，価値のある結果とはいえない．そこで，文理共通科目で選択可能であり，かつ必修でない社会･芸術系科目のみで決定木を生成した．生成した決定木を図 3 に示す．図 3 より，文系学生の多くが「日本史 B」を履修する傾向があり，理系学生は「日本史 B」，「世界史 B」を履修しない傾向にあることが分かった．

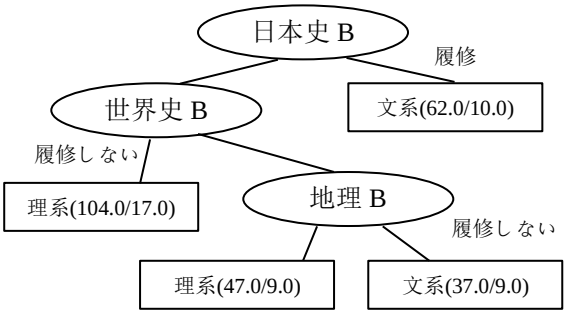


図 3　履修科目による決定木

また，属性を理科科目(物理，化学，生物，地学)のみで生成した場合の決定木を図 4 に示す．図 4 より，「物理Ⅰ」の履修の有無でほとんどの事例が分類された．物理の問題は計算が多く数学に近いことから，文系学生は履修を避けているのではないかと考える．また，理系学生は「生物Ⅰ」をあまり履修しない傾向がみられる．これは，理系学部で生物を受験必須科目とする学部が少ないこと，物理化学に比べて文系科目寄り(暗記が多い)ことが理由であると考える．

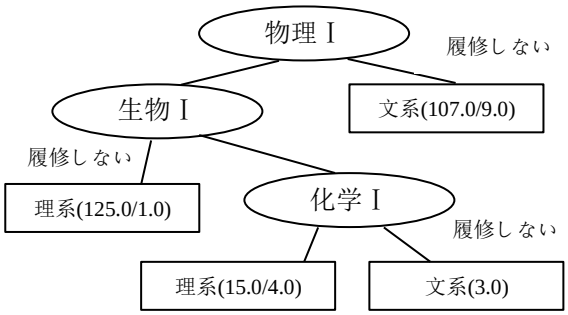


図 4　理科科目による決定木

C. 文理選択の理由に基づく分析

最後に，文理選択の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多種多様であり，そのままでは扱うのが困難である．その為，属性「文理選択の理由」の属性値から，文理選択の理由を「科目の嗜好等」，「学校」，「興味」，「将来」，「その他」の 5 つに分類した．以下にそれぞれの説明を示す．

●「科目の嗜好等」
科目の好き嫌い,得意不得意を理由とする場合
●「学校」
学校のクラス分けや先生を理由とする場合
●「興味」
文理に関する興味を理由とする場合
●「将来」
将来やりたいことなどを理由とする場合
●「その他」
上記以外(「なんとなく」や「わからない」などの消極的な回答を多く含む)

文系・理系それぞれでの統計結果を円グラフで示す．図 5 に大学生，図 6 に高校生，図 7 に大学生と高校生を合わせたデータを使用したものを示す．

図 5，図 6 および図 7 から，大学生，高校生共に「科目の嗜好等」を文理選択の理由とする学生の割合が最も多いことがわかる．特に文系の大学生は 6 割以上もの学生が「科目の嗜好等」を理由としている．そのため，科目の好き嫌い，得意不得意が文理選択に大きく影響していると考えられる．一方で，「将来」，「興味」を理由とする学生が全体的に少ないこと，特

に高校生において「その他」の割合が多いことから，学生の進路に関する無関心が懸念される．

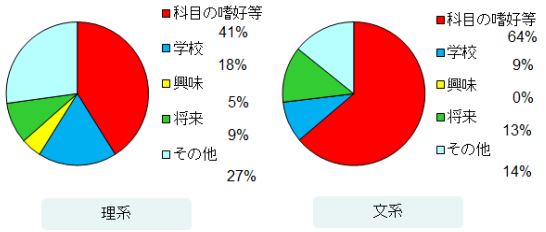


図 5　大学生の文理選択の理由

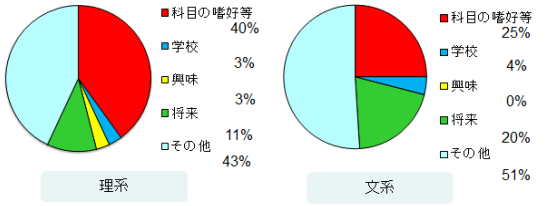


図 6　高校生の文理選択の理由

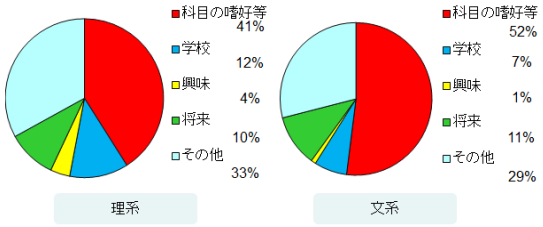


図 7　学生の文理選択の理由

「科目の嗜好等」を文理選択の理由とする学生が多いことから，科目ごとの嗜好について調査した．図 8 に文系大学生の科目の嗜好をまとめたグラフを示す．横軸が科目，縦軸が学生の割合である．文系高校生においても図 8 とほぼ同様なグラフとなった．結果として多くの文系学生は「数学」が苦手または嫌いであり，このことが理系への進学の妨げとなっていると考えられる．

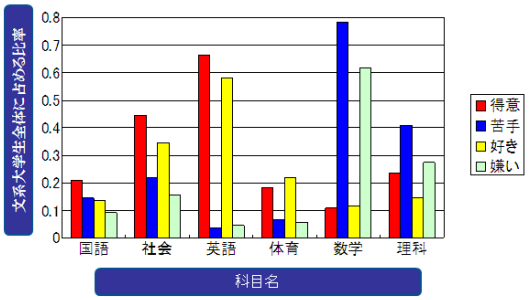


図 8　文系大学生の科目の嗜好

また，文系学生の数学が苦手または嫌いな理由は「難しい」「理解できない」が大半を占めており，苦手または嫌いになる時期は，小中学校など早い時期に多い傾向が見られる．これらのことから，小中学校で学ぶ基礎的知識が身につかず，進学してからも授業についていけないことが理系離れに影響を与えていると考えられる．

## V．おわりに

本研究では，高校生・大学生を対象にアンケートを実施し，決定木学習を用いて，学生の生活習慣，科目の履修・嗜好等の傾向から文理選択に影響を与える要因について考察を行った．生成した決定木から文系・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．

本研究では，高校生・大学生を対象にアンケートを実施し，決定木学習を用いて，学生の生活習慣，科目の履修・嗜好等の傾向から文理選択に影響を与える要因について考察を行った．生成した決定木から文系・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から，大学生及び高校生の文理それぞれでの特徴を挙げた．しかし，大学生と高校生での共通性を発見することができなかった．このことから，文理選択に影響を与える要因には年代による差異があると考えられる．また，文理選択の理由としては科目の嗜好，特に数学に関して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苦手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．従って，理系離れの対策として小学校などの初等教育で数学を苦手にさせない授業展開・指導案が必要であると考えられる．

本実験で使用したデータは，収集地域，アンケート実施対象が限定的である為，得られた結果は限定的なものであるといえる．その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，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，御指導，御教授頂いた上田祐彰准教授に深く感謝致します．また，アンケート収集などにご協力いただいた方々に感謝致します．

参考文献

[1] 鶴岡森昭 ほか，“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），第 1 号(1996)，pp105-115.

[2] 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

[3] “Weka3 - Data Mining with Open Source Machine Learning Software in Java”，
http://www.cs.waikato.ac.nz/ml//index.html

問い合わせ先
〒731-3194
広島市安佐南区大塚東 3 丁目 4 番 1 号
広島市立大学大学院　情報科学研究科
野津田　雄太