相関ルールを用いた誤答パターンの分析

Analysis of the Pattern of the Erroneous Answers Utilizing Association-rule Algorithm .

松河秀哉[*1]・北村智[*2]・山内祐平[*2]・中野真依[*3]・金森保智[*3]・宮下直子[*3]
Hideya MATSUKAWA[*1]・Satoshi KITAMURA[*2]・Mai NAKANO[*4]・
Yasutomo KANAMORI[*4]・Naoko MIYASHITA[*4]

大阪大学大学教育実践センター[*1]・東京大学大学院情報学環[*2]・株式会社ベネッセコーポレーション[*3]

Institute for Higher Education Research and Practice*, Osaka University[*1], Interfaculty Initiative in Information Studies, The University of Tokyo[*2], Benesse Corporation[*3]

〈あらまし〉 本研究では、データマイニングの一手法である相関ルールを用いて、中学 1 年から高校 1 年までの数学の模擬試験の誤答データの時系列的な分析と、高校 1 年向けの英語の模擬試験の回答データの分析を行った。その結果、いずれの分析でも、生成されたルールから、誤答の原因をある程度推測可能であることが示唆された。

〈キーワード〉 相関ルール・誤答パターン・データマイニング

1. はじめに

教育工学の関連領域において、誤答分析はこれまで、算数などの教科を中心に、学習者が持っている誤りのモデルを明らかにしようとするバグ研究一環として盛んに行われてきた(Wenger 1987)。また、CAI 教材の作成の前段階として、教材が対象とする領域の典型的な誤答を収集することも数多く試みられている(洲脇ら 2001)。

こうした研究においては、誤答の分析は、例えば引き算の方法など、ごく限られた領域に焦点を当て、その範囲の誤答を研究者自身が分類して詳細に調べることが多く、労力の観点から、より広い領域間の関連はほとんど注目されることがなかったと言える。

一方、近年では、情報技術の急速な進歩に伴い、膨大な情報の中から有用なデータを発見するための、データマイニングの手法が発展してきた。

自動的に膨大な情報を処理するデータマイニングの手法を利用すれば、これまで困難であった、より広い領域にわたる、誤答の関係性を分析することが可能であると考えられる。

そこで、本研究では、中学 1 年の 1 学期から高校 1 年 1 学期まで 3 年間にわたる数学の学力テストの正誤データ、及び、高校 1 年時の英語の学力テストの回答データに対して、相関ルールを用いた分析を行うことにより、中学段階と高校段階での誤答の関連性や、様々な問題のタイプを横断して現れる特徴的な誤答など、広範囲のデータおける誤答のパターンを明らかにすることを目的とする。

2. 分析方法

2. 1. 相関ルールと評価の尺度

相関ルールは、A ならば B であるの様に、ある事象と別の事象の共起関係の起こりやすさを表すものである。例えば、ある店でパンとバターを同時に購入する人の 90 %は、ミルクも購入するといったものが、相関ルールのひとつの例である。マーケティング分野では、レジで収集される POS データを用いて、こうした分析が日常的に行われており、商品の効率的な陳列などに応用されている。

最近では、教育の分野においても、大量の学習履歴データの中から、「問題 A と B に正答すれば、90 %の確率でで合格」といったように、最終成績を予測するルールを探索する研究(Minaei-Bidgoli 2003)などが行われるようになってきている。

相関ルールの「よさ」を計る尺度としては、支持度(support)と確信度(confidence)がよく用いられる。例えば、ある模擬試験の受講者数を N、その中で、問題 1 と問題 2 を間違った

人の数をNa、問題3を間違った人の数をNb、問題1，2，3を同時に間違った人の数をNabとすると、問題1と問題2に間違うならば、問題3も間違うというルールの支持度は、Nab/Nと表すことができる。またそのルールの確信度は、Nab/Naと表すことができる。一般に、支持度が高いルールほどよく通用するルールであり、確信度が高いほどそのルールは信頼できる。

確信度は、有効なルールの判断基準となり得るが、誤答を分析する場合、結論部分に難問が含まれていると、確信度だけでは不十分となる場合もある。例えば上のルールで、問題3の誤答率が元々90%近くあったと仮定すると、「問題1と問題2に違うならば、問題3も間違える」というルールは確信度が90%あっても、「問題1と問題2に間違える」という条件部分はあまり特別な意味があるといえなくなってしまう。

そこで利用されるのが、「リフト」という指標である。リフトはルール全体の確信度を結論部分の確信度で割ったもので、(Nab/Na)/(Nb/n)と表すことができる。例えば、問題3単体の誤答率が10%なのに、「問題1と問題2に違うならば、問題3も間違える」というルール全体の確信度が90%ならばリフトの値は9である。これは、「問題1と問題2に間違える」という条件がつくことで誤答が9倍起こりやすくなったことを意味する。反対に、問題3単体の誤答率が89%で、ルール全体の確信度が90%の場合は、リフトは90/89 ≒ 1であり、確信度が同じでも、「問題1と問題2に間違える」という条件は、あまり重要でないことが判断できる。

本研究の分析においては、相関ルールの絞込みに、上述した、確信度、支持度、リフトを尺度として用いた。

2.2. 分析対象

分析対象とするデータは、株式会社ベネッセコーポレーションの進研ゼミ中学講座と高校講座の全国学力模擬試験のデータである。

数学の誤答分析には、中学1年の1学期から中学3年2学期までの毎学期と、高校1年1学期の計9回分の模擬試験（2004年度から2007年度）を全て受験している4444名の各設問に対する正誤データを用いた。分析対象となった変数の数は模試の設問の数と同じ389であった。

英語の誤答分析には、高校1年1学期のマークシート形式の模擬試験（2007年度）の各設問に対する25919件の回答データを用いた。模試の設問数は117問であったが、複数の選択肢をダミー変数化したため、分析対象となった変数の数は917であった。

2.3. 分析ツール

2.3.1. 分析に用いたハードウエア

分析を行ったPCのスペックは以下のとおりである。

CPU:Xeon 5365 3.0Ghz（4core）
Memory:DDR-2 677 FB-DIMM 16GByte
HDD:1TByte（500GByte × 4 Raid1+0）
OS:WindowsXP x64 Edition

2.3.2. 分析に用いたソフトウエア

相関ルールの分析には、Christian Borgeltによって開発された、apriori.exeを用いた。本ソフトは、相関ルールを求めるための代表的なアルゴリズムである、apriori アルゴリズムを実装したもので、フリーウエアとして、以下のサイトで公開されている。

http://www.borgelt.net/apriori.htm

このサイトで公開されているapriori.exeは32bit版であるが、本ソフトウエアを用いてWindowsXP上で、試験的に分析を行ったところ、変数が多いため、プロセスがメモリ空間のに収まりきらず、分析が中断された。そこで、本ソフトウエアのソースコードを、VisualStudio2005を用いて64bit環境(x64)向けにコンパイルして、分析に利用した。

分析の結果として生成された相関ルールの格納と検索を行うためのデータベースにはMS-SQL Server2005を用いた。

2.4. 分析手順

1. apriori.exeを用いて相関ルールを求めるにあたって、数学のデータに関しては、確信度70%以上、ルールの支持度0.1%以上、リフト1.5以上、条件の個数3以下という制約を与えた。英語のデータに関しては、確信度90%以上、ルールの支持度0.1%以上、リフト1.5以上、条件の個数3以下という制約を与えた。

2. apriori.exeの出力結果は、自作のスクリプトを用いてタブ区切りのデータに整形した

上で、DB にインポートした。

3. DB に格納したデータに対して、SQL 文を発行し、確信度やリフト値によるソートをかけながら、誤答間の関係を解釈可能なデータを筆者が単独で探索した。

3. 結果

数学のデータに関しては、所与の条件で、420742255 件のルールが作成された。このうち、ルールの結論部分に、高 1 の設問のみを含み、条件部分に中学の設問のみを含むルールは 48641199 件あった。ルールの件数が膨大であるため、そのすべてを精査することは不可能であった。したがって、ここでは、結論部の誤答率とルールの支持度が高く、最も明確に誤答間の関係を解釈可能であった 4 問の問題の組み合わせを図 1 から図 4 に示す。

図 1 から図 3 は中学時代の誤答を表し、図 4 は高校時代の誤答を表している。問題 A、問題 B、問題 C に同時に間違えたのは 351 人(全体の 7.9%)であり、そのうち問題 D も間違えたのは 250 人(全体の 5.6%)だった。つまり問題 A と問題 B と問題 C に間違えれば問題 D も間違えるというルールの確信度は 71.2%であった。ルールのリフト値は 2.694 であり、問題 A と問題 B と問題 C を同時に間違えるという条件がつくことによって、問題 D は約 2.7 倍間違えやすくなることが示されている。

次に英語の誤答に関して所与の条件で分析したところ、152352 件のルールが生成された。このうち、誤答の原因として未回答を含まないものは 123 件あった。その中で、結論部の誤答率とルールの支持度が高く、最も明確に誤答間の関係を解釈可能であった 4 問の問題の組み合わせを図 6 に示す。

この結果のなかで、問題 A の回答が 3,2 かつ問題 B の回答が 6,2 かつ問題 C の回答が 3 であった学習者は 37 人であり、このうち、問題 D の回答が 1 なのは 35 人(全体の 0.14 %)であった。したがって「問題 A,B,C に間違うならば問題 D も間違う」というルールの確信度 94. 6%であり、リフト値は 2. 34 であることから、問題 A,B,C を同時に間違う学習者は、そうでない学習者より、2 倍以上問題 D を間違えやすいことが示されている。

4. 考察

4.1. 数学の結果について

まず、数学の分析結果についての誤答間の関連性について考察する。はじめに、相関ルールの結論部分である高校時点の問題 D の性質からみてみると、この問題を解くには、図 5 の様な図示を行い、次にその情報を用いて方程式を立式し、方程式を計算して解を出すと

問11　a でわると商が b で余りが c となる正の整数 n があります。整数 n を a, b, c を用いて表しなさい。

① $n=a+b+c$
② $n=a(b+c)$
③ $n=ab+c$
④ $n=\frac{b}{a}+c$

図1 中学時代の問題A(全体の誤答率29.5%)

問28　右の図で，直線 ℓ の式は $y=x+1$，直線 m の式は $y=-2x+10$ です。直線 ℓ と m の交点を A(3, 4)とし，直線 ℓ，m と x 軸との交点をそれぞれ B，C とします。点 A を通り，△ABC の面積を 2 等分する直線の式を求めなさい。

① $y=2x-5$　② $y=2x+2$
③ $y=4x-8$　④ $y=-\frac{3}{2}x-3$

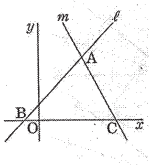

図2　中学時代の問題B(全体の誤答率21.9%)

次のそれぞれの場合について，y を x の式で表しなさい。

問17　周の長さが 30cm である長方形の縦の長さが x cm，横の長さが y cm

① $y=\frac{30}{x}$
② $y=-x+30$
③ $y=-x+15$
④ $y=15x$

図3　中学時代の問題C(全体の誤答率35.1%)

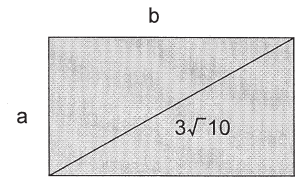

図5　問題Dを解くのに必要な図示

(6)　長さが24cm の針金を折り曲げて長方形を作ると，対角線の長さが $3\sqrt{10}$ cm になった。

この長方形の面積は，[サシ] cm^2 である。

図4　高校時代の問題D(全体の誤答率26.4%だが、上記3題とも間違えると誤答率は71. 2%)

問題A（正答率61.1%、このタイプの誤答率1.7%）
二つめの角を左に曲がると、白い建物がある。
・You will find a white building ① __ __ ② __ the second
1. and 2. at 3. if 4. left 5. turn 6. you
正答：You will find a white building if you turn left at the second
誤答：You will find a white building if you turn at left the second

問題B（正答率47.2%、このタイプの誤答率7.2%）
父はたくさんの助言を私にしてくれた。
・My father ① __ __ ② __.
1. advice 2. a lot of 3. gave 4. me 5.to 6. was given
正答：My father gave a lot of advice to me.
誤答：My father was given to me a lot of advice.

問題C（正答率14.6%、このタイプの誤答率41.9%）
・We finally ① the top of the mountain.
1. reached 2. reached at 3. reached on 4. reached to
正答：We finally reached the top of the mountain.
誤答：My father reached on the top of the mountain.

問題D（正答率47.8%、このタイプの誤答率38.8%
→ABC誤答がそろうと94.6%に高まる）
・He showed ①.
1. me to his computer 2. his new computer me
3. me of his computer 4. his new computer to me
正答：He showed his new computer to me.
誤答：He showed me to his new computer.

図6 英語の誤答パターン例

いう手順を踏む必要があることが分かる。

次に、相関ルールの条件部分に含まれる中学時点の各設問の性質をみると、問題Aは文章から立式を行う問題、問題Bは文章の図示、図からの立式、式の計算を行う問題、問題Cは文章の図示と図からの立式を行う問題である。

このことから、相関ルールの条件部分に含まれている問題は、結論部分の問題を解くために必要なスキルの、一部、もしくは全部が必要な問題であることが分かる。即ち、このルールあてはまる学習者は、文章で書かれた問題から図示や立式をすること、その上で計算を行うことのどちらか、もしくは両方に以前からつまずいていた可能性が考えられる。

このように、相関ルールを用いることで、中学時代と高校時代のように時系列的に離れた誤答間の関連性の分析から、その学習者がある問題に誤答する原因を過去にさかのぼって探れる可能性が示唆され、その原因に応じて学習者に効果的な処遇を行うことなどが期待される。

4.2. 英語の結果について

図6中の各設問について、その性質を見てみると、問題Aと問題Cは助詞を伴わない動詞の扱いに関連する誤答、問題Bと問題Dは二重目的語を取る動詞に関連する誤答、であり、一見したところ、全体としての関連は薄そうに思われる。しかし、日本語に直して考えてみると、「左に曲がる」、「私にしてくれた」、「山頂に着いた」、「私に見せた」というように、全て「に＋動詞」という訳が含まれる問題であることが分かる。即ち、このルールにあてはまる学習者は、日本語に引きずられて、動詞の扱いを誤っている可能性が考えられる。

このように、分析対象が1回の模擬テストの結果であっても、選択肢の内容まで含めて相関ルールを求めることで、比較的詳細に誤答の原因を推測できる可能性が示唆された。

5. まとめ

本研究では、データマイニングの一手法である相関ルールを用いて、中学1年から高校1年までの数学の模擬試験の誤答データの時系列的な分析と、高校1年向けの英語の模擬試験の回答データの分析を行った。その結果、いずれの分析でも、生成されたルールから、誤答の原因をある程度推測可能であることが示唆された。

相関ルールから誤答の原因を推測するには、ルールの解釈が不可欠であるが、今後は、この作業に教科の専門家の協力を求めることで、より重要な誤答間の関連が発見されることが期待される。

参考文献

Minaei-Bidgoli, B., Kashy, D.A., Kortemeyer G., Punch, W.F., （2003）Predicting Student Performance: An Application of Data Mining Methods with an educational Web-based System, IEEE comf. on Frontier In Education FIE 2003

岡本敏雄・溝口理一郎（監訳）（1990）知的CAIシステム．オーム社，東京（Etienne Wenger,（1987）Artificial Intelligence and Tutoring Systems. Los Altos, CA）

洲脇史朗・宮地功（2001）中学校における「一次方程式」の誤答分析．電子情報通信学会技術研究報告．ET，教育工学，Vol.101, No.397（20011019）pp. 1-8

* 本研究は、東京大学大学院情報学環ベネッセ先端教育技術学講座のプロジェクトとして、（株）ベネッセコーポレーションと共同で行われている．