

学習履歴データベースにおける データマイニング関連ルールに関する考察

白田 由香利

学習院大学 経済学部 経営学科

〒171-8588 東京都豊島区目白 1-5-1

yukari.shirota@gakushuin.ac.jp

本稿では、e-Learningシステムにおいて学習者の学習履歴データをデータマイニングする場合どのような関連ルールが抽出されるべきであるか、を考察する。e-Learningシステムにおいて学習者がシステムの提示したドリル問題に対して不正解であった場合、システムは学習者の状況を、データマイニングによって発見されている関連ルールを用いて分析判断したい。そのためには予め、有益な関連ルールを膨大な学習履歴データベースからデータマイニングで抽出しておくことが必要である。本論文では、個々の学習キーワードが単独で出現するドリルであれば解けるのであるが、学習キーワードが複合化して出現した場合にはドリルが解けなくなる、という状況を発見するために有効な関連ルールとはどのように定義すべきであるか、について考察する。我々の提案する設定問題は、従来のデータマイニングのバスケット解析などとは本質的に異なり、新たな抽出技法が必要となる。その技法についても考察する。

キーワード e-Learningシステム, データベース, データマイニング, 関連ルール

Data Mining Association Rules for Learning Transaction Databases

Yukari Shirota

Department of Management, Faculty of Economics, Gakushuin University

1-5-1 Mejiro, Toshima-ku, Tokyo, Japan

yukari.shirota@gakushuin.ac.jp

In the paper, we discuss which kind of association rules are effective for the data mining on an e-Learning system learner's learning transaction database. The data mining system, in advance, finds effective association rules from the learning transaction database. By using the association rules, the e-Learning system detects the learner's stalled situation and analyses its reasons. In particular, the stalled situation in which the learner understands individual keywords separately but the learner cannot solve a problem that involves compound keywords is the target of our study. Our proposed problem is different from the existing association rules for market basket analysis problem. Then the new solving methods are required. In the paper, the methods are also discussed.

Key words e-Learning system, database, data mining, association rule

1. はじめに

本稿では、e-Learningシステムにおいて学習者の学習履歴データをデータマイニングする場合どのような関連ルールが抽出されるべきであるか、を考察する。e-Learningシステムにおいて学習者

がシステムの提示したドリル問題（以下、ドリルと呼ぶ）に対して不正解であった場合、システムは学習者の状況を、データマイニングによって発見されている関連ルールを用いて分析判断したい。

そのためには予め、有益な相関ルールを膨大な学習履歴データベースからデータマイニングで抽出しておくことが必要である。問題を明確化することが容易であるという理由から、対象学科は数学とする。図1に学習履歴データベースの様子を示す。

一般に学生が数学ドリルを解く場合、複数分野に分かれた個々の公式を使うことができて、複数の公式が関連した問題になると、自分で問題を変形して公式の組合せ問題に帰着することができず、解けないことが多い。数学の問題を解く力を養うということは、個々の学習した公式や学習ポイントを複合化して活用できるようになることであり、数学のe-Learningシステムにおいてはどのような複合化によって、学生はどの箇所で見失ってしまうのか、その箇所を発見することは極めて重要なポイントとなる。

以下では、公式や学習ポイントなどの、教材における学習テーマを総称して、「学習キーワード」と呼ぶこととする。本論文で研究対象とするのは、個々の学習キーワードが単独で出現するドリルであれば解けるのであるが、学習キーワードが複合化して出現した場合にはドリルが解けなくなる、という状況を発見するために有効な相関ルールとはどのように定義すべきであるか、また、その抽出技法について考察する。

上記研究対象ルールを、簡単のため以下では、「キーワード複合化による行き詰まり発見ルール」と呼ぶことにする。正確には、行き詰まり状態を発見するためのルールであるが、簡単に上記のように呼ぶこととする。

次節では、キーワード複合化による行き詰まり発見問題とは何かを説明し、第3節では、問題解

決に利用可能な新たな相関ルールの定式化を試みる。第4節では関連研究について述べ、我々の提案する設定問題が従来のデータマイニングのバスケット解析などとは本質的に異なることを説明する。第5節では、キーワード複合化による行き詰まり問題で利用できる新たな相関ルールの抽出技法について論じる。

2. キーワード複合化による行き詰まり問題

我々が研究対象とする問題は、学習者が数学ドリルを間違えた場合、あるいは分からなくて行き詰った状態(これを総称して、以下では行き詰まり状態と呼ぶ)において、行き詰まり状態になった原因を分析して「学習キーワードの複合化により問題難度が上昇したので、学習者は解けなくなってしまった」という仮説が成立するか否かを判断するルール及びアルゴリズムを発見すること、である。

この問題の定式化に入る前に、具体的な例を示す。例えば、以下の数学ドリルを考えてみる(問題及びその解説は[1]から引用)。

キーワード複合化問題の例:

a を 0 でない実数とする時、2 次方程式

$$ax^2 - (a-1)x - a = 0$$

の実数解の個数を調べよ。

この問題の場合「2 次方程式の判別式」及び「平方完成という基本変形」という 2 つの学習キーワードが複合化している。この学習キーワードに対して、単独に個別のドリルとして与えられれば解ける学生でも、「実数解の個数を調べよ」と一挙に問題が与えられると、判別式を平方完成するこ

id	student	name	content	key1	key2	key3	kind	number	result	date	session
1296	1225	sato	対数微分法ドリル	対数微分法	対数関数	分数				2003-04-05 13:38:36	0e18fa12ee3c28e2
1297	1225	sato	対数微分法ドリル	対数微分法	対数関数	分数	回答	1	0	2003-04-05 13:38:38	0e18fa12ee3c28e2
1298	1235	sato	累乗の導関数ドリル	微分	累乗	マイナス				2003-04-05 13:40:53	0e18fa12ee3c28e2
1299	1235	sato	累乗の導関数ドリル	微分	累乗	マイナス	回答	1	1	2003-04-05 13:40:56	0e18fa12ee3c28e2
1300	1235	sato	指数関数の導関数ドリル	指数関数	合成関数	累乗				2003-04-05 13:40:58	0e18fa12ee3c28e2
1301	1235	sato	指数関数の導関数ドリル	指数関数	合成関数	累乗	回答	3	0	2003-04-05 13:41:04	0e18fa12ee3c28e2
1302	1235	sato	商の微分公式ドリル	微分の商の公式	分数					2003-04-05 13:41:16	0e18fa12ee3c28e2
1303	1235	sato	商の微分公式ドリル	微分の商の公式	分数		回答	1	0	2003-04-05 13:41:20	0e18fa12ee3c28e2
1304	1235	sato	合成関数の導関数ドリル	合成関数	微分の積の公式					2003-04-05 13:41:38	0e18fa12ee3c28e2
1305	1235	sato	合成関数の導関数ドリル	合成関数	微分の積の公式		回答	1	0	2003-04-05 13:41:48	0e18fa12ee3c28e2
1306	1235	sato	対数微分法ドリル	対数微分法	対数関数	分数				2003-04-05 13:42:04	0e18fa12ee3c28e2
1307	1235	sato	対数微分法ドリル	対数微分法	対数関数	分数	回答	1	0	2003-04-05 13:42:07	0e18fa12ee3c28e2
1308	1235	sato	行列の用語定義							2003-04-05 13:42:49	0e18fa12ee3c28e2
1309	1235	sato	行列の積							2003-04-05 13:42:52	0e18fa12ee3c28e2
1310	1235	sato	連立方程式の書き方							2003-04-05 13:42:55	0e18fa12ee3c28e2
1311	1235	sato	逆行列って何							2003-04-05 13:43:00	0e18fa12ee3c28e2
1312	1235	sato	行列式って何							2003-04-05 13:43:03	0e18fa12ee3c28e2

図1：学習履歴データベースの例

とによって常に正であることを示せばよい、と気づくに至らず解けなくなってしまうことが多い。解法の個々のテクニックを独立には理解していても、それらを組み合わせて問題に適応することができず、問題を前に所謂行き詰まり状態になってしまうのである。

数学問題は、全て学習キーワードの複合化によって問題が成立していると言える。例えば、大学生用経済数学の問題ドリルに対して、通常「分数の通分」のような基礎的な学習キーワードは付与しないが、実際にはこのような学習キーワードの積み重ねがなくては、問題解決には到らない。数学は公式や概念の積み重ねによって成立しているため、全ての問題ドリルも、多数の学習キーワードの複合化によって成立していると言える。

数学の指導において、学習キーワードの複合化によって正答率が大幅に変化するという、キーワードのペアを発見することは、学生の学習効率を高める上で非常に有効といえる。学校、予備校や塾で教え方の上手な先生というものは、数学を教えた長年の経験から「どこで学生が間違えやすいのか」という知識を獲得しており、その知識ベースにより、(1) その複合化を上手に説明する良質な教材を作成する、さらに、(2) 学生が行き詰っている場合、その状態がキーワード複合化による行き詰まり状態であることを容易に発見し、適切なアドバイスをを行う。

この人間のベテラン教師が行っている教育支援手法をマシン上でいかに実現していくかが、我々の研究テーマである。我々は教育支援システムの個人化をテーマとして研究を続けているが[2,3]、教育支援システムには、こうした行き詰まり状態を発見し、原因を分析し、適切なアドバイスを学生に与える機能が必要であると考え。そして行き詰まり原因の中でも、我々が問題とする学習キーワードの複合化によるものは重要であると考え。

では具体的にどのように教育支援システムにデータマイニングの結果を活用すればよいかについて、以下、考察する。データマイニングの結果、キーワード複合化による行き詰まり状態が発見されたとする。例えば、「判別式と平方完成が個々に理解できている学生でも、両者が複合化した問題になると正答率が低い」という事実が発見されれば、以下のような教育支援戦略を立てることが可能となる。

(A) 判別式と平方完成を複合化した問題を多く出題し、トレーニングさせる。

(B) 行き詰っている学生に対し、2つの学習キーワードが複合化されていることを気づかせるような、アドバイスをし、正解へ誘導する。マシン上で実現する場合、上記(B)の方が(A)よりも重要であり、実現困難度は高くなる。(B)の問題は、数学における理解とは何か、という問題に行き当たるからである。機能(B)の実現のため、教育支援システムに望まれる機能としては、「学生が何を分からないのかを明確にする」ための対話的カウンセリング機能。さらに、テキストの表面的意味を理解するだけではなく、テキストがなぜそう書かれているのかとか、そうした論理的展開はどうして思いつかれたのか、という深い理解[4]のレベルに学生を導くガイダンス機能が必要となるからである。

こうした人間の理解を支援するための補助システムの実現には、人工知能分野などの研究成果である、各種の学習方法(例: Explanation-based Learning)を適応するというアプローチも考えられる。また、ベテラン教師の発するアドバイスをデータベース化して、それを状況判断結果に応じて学生に提示するというアプローチもある。例えば、上記の数学問題例では「解の個数と言われたら、判別式を思い出してみよう」という教師のアドバイスが有効であると考えられる。「解の個数⇒判別式の利用」というような連想をベテラン教師のアドバイス語録から発見してデータベース化するというアプローチも、教育支援システムには有効な手法と考える。こうした学生の理解を高めるための対話的アドバイス機能を実現するためにも、その前段階として学習キーワードの複合化による行き詰まり状態発見のための相関ルールをデータマイニングで発見することが重要である。

3. キーワード複合化による行き詰まり発見ルール

本節では、教育支援システムにおいてデータマイニングを行う際の、学習キーワード複合化による行き詰まり状態を発見するための相関ルールについて定式化を行う。

3.1 教育支援システムのモデル化

対象とするe-Learningシステム及び教育支援システムには、(1) 教材データベースと、及び(2) 学習履歴データベースの2種類のデータベースがあるとする。教材データベース中の教材には、教材の内容を示すメタデータが教師サイドによって予め付加されている。このメタデータを「学習キー

ワード」と呼ぶ。

学習履歴データベースには、学習者が行ったドリル問題の回答及び、学生が回答した結果の正誤(正解不正解)情報が、学習トランザクションとして蓄積されていく。学習トランザクションは実際には、教材 ID のみが記録され、教材の学習キーワードが必要な場合は、結合操作を行って学習キーワードを求めるが、説明の簡単化のため、学習トランザクション・データベースの要素は、学習キーワード集合とドリルのタプルで表現されるとする。

問題に付加されるキーワードの集合を $K=\{k_1, k_2, \dots, k_m\}$,
学習トランザクション・データベースを $D=\{t_1, t_2, \dots, t_n\}$
と表わす。但し $t_i=(tk_i, r)$ ($tk_i \subseteq K$)。

各要素 tk_i は、その学習トランザクションで対象としたドリル問題に付与されていた学習キーワード集合である。長さ m のキーワード集合とは m 個のキーワードの組合せを指す。上記 r は回答の正誤を表す Boolean 型値を返り値とする関数である。 r は、ドリルの正誤を示す関数 f によって以下のように定義される。

$$r = f(\text{userID}, \text{time}, X) \quad (X \subseteq K)$$

ここで、引数 userID は学習者の識別子、 time はドリルに回答した日時を表わす。引数 X は、ドリルに付与されている学習キーワード集合を表わす。

3.2 理解度の定義

次に、学習キーワード集合 $X(X \subseteq K)$ に関する正解度 $\text{correct}(X)$ を定義する。正解度 $\text{correct}(X)$ は、学習トランザクション・データベース D 全体に対して、「 r が TRUE であり、かつ学習キーワード集合 X を含む」学習トランザクションの割合として定義される。 D の全レコード数が 100 であり、条件に合致するレコード数が 14 であれば、14%となる。

X に関する不正解度 $\text{incorrect}(X)$ は、 D 全体に対して、「 r が FALSE であり、かつ学習キーワード集合 X を含む」学習トランザクションの割合として定義される。任意の X に対して、 $\text{correct}(X) \cap \text{incorrect}(X) = \phi$ であり、重複はない。つまり中間点は与えないこととする。

キーワード集合 X に関する理解度 $\text{understand}(X)$ は、以下のように定義される。

$$\text{understand}(X) = \frac{\text{correct}(X)}{\text{correct}(X) + \text{incorrect}(X)}$$

学生が数学を勉強する場合、複数分野に分かれた個々の公式を使うことができて、複数の公式が関連していると、自分で問題を変形することができず解けないことが多い。これを「キーワード複合化による行き詰まり状態」と呼ぶことにした。この行き詰まり状況は、以下の3通りに分類できる。学習キーワードは X と Y の2個とする。

- (1) そうでなくても分からない学習キーワード2つ X 及び Y が、複合して XUY となったので、さらに分からなくなった。
- (2) 片方の学習キーワード X 単独(あるいは Y 単独)では分かっていたのだが、複合して XUY となったら分からなくなった。
- (3) X 及び Y は、単独では分かっていたのだが、複合して XUY となったら分からなくなった。

一般に、概念 X と概念 Y に対しては、両者に共通する基礎概念がある可能性があるが、問題の簡単化のため X と Y には重複はないとする。こうした方がデータマイニングによって発見された状態を説明する際、説明が容易であるからである。よって、常に $\text{correct}(X) \geq \text{correct}(XUY)$ 及び $\text{incorrect}(X) \geq \text{incorrect}(XUY)$ が成立する。

X 単独出現の理解度は、以下のように表現される。

$$\text{understand}(X) - \text{understand}(XUY)$$

上記3つのタイプの例を先に定義した understand を用いて表現した例を表1に示す。

表1：3つのケースの理解度の計算例

理解度	X単 独出 現	Y単 独出 現	XUY 複合出 現	複合化による行 き詰まり判定指 標値 (<0.5)
ケース(1)	30%	30%	10%	111%
ケース(2)	30%	90%	20%	74%
ケース(3)	80%	80%	30%	46%

表1に示すケース(3)、つまり複合化による難度上昇の場合を説明する。個々の学習キーワード集合 X 及び Y についての理解度は高いので、その理解度を其々 80% とした。複合化した状態での理解度は低いとするため、その理解度を 30% とした。

$$\begin{aligned} \text{understand}(X) &= 0.8, \\ \text{understand}(Y) &= 0.8, \\ \text{understand}(X \cup Y) &= 0.3 \end{aligned}$$

理解度の値による、分かった、分からないという領域の閾値を、

- ・ 80%以上であれば、分かった.
- ・ 30%以下であれば、分からない.

とする. この閾値は状況に応じて変更するものとする.

キーワード複合化による行き詰まり発見ルールは, X 単独出現, Y 単独出現, XUY の複合出現, 其々に対して上記のように閾値を設定し, その条件を満たすか否かで判断すればよい. しかし, 以下のように指標を定義し, 一つのルールとしてまとめることもできる.

$$\frac{\text{understand}(X \cup Y)}{\{\text{understand}(X) - \text{understand}(X \cup Y)\} \cdot \{\text{understand}(Y) - \text{understand}(X \cup Y)\}} \leq 0.5$$

このルールの指標値を上記3ケースに対して計算した結果を表1の最右列に示す. ケース(3)の場合のみが50%以下となり, 上記ルールを満たしていることが分かる.

4. 既存研究との比較

関連する既存研究として, 一般的データマイニングにおける相関ルール抽出技法との比較を行なう. キーワード複合化による行き詰まり発見ルール問題は, 変化率に着目する, という点において, 通常のバスケット分析手法および, 数値属性相関ルールとは異なる問題であることを説明する.

4.1 データマイニングの数値属性

データマイニングの手法として「AならばBである」といったデータベース属性間の相関関係を求める相関ルールを発見するためのアプリアリ・アルゴリズムが Agrawal により提案され, 広く普及している[5,6]. しかし現行のデータベースには性別, 血液型といったカテゴリ属性のほか, 年齢, 体重といった数値でとられる数値属性が含まれているが, アプリアリ・アルゴリズムはカテゴリ属性間の相関ルールを求めるためのものであり, 数値属性に対して直接適応することはできない[7].

そのため, 数値属性の値域を領域に分割し, 数値属性とカテゴリ属性間の相関ルールの生成を行

う手法が提案された[8,9,10]. この手法では「 $x \in R$ ならばBである」(Rはいくつかの数値属性の値の空間の部分領域)のようなルールを発見する. 福田が[8]で示した例を以下に示す. 「たとえば, 銀行の顧客のうち, 預金残高がある区間Iに入る顧客は, カードローンを利用する確率が高いとする. この知識は(預金残高 $\in I$) \Rightarrow (カードローン利用=yes)という結合ルールとして表現できる. このようなルールは, 区間Iが預金残高とカードローン利用の間の特別な性質を持つ場合についてのみ, 面白いルールといえる. その性質とは, Iに含まれる顧客の数(サポート)が十分に大きく, 条件を満たす確率(確信度)が十分に高いことである」(引用終わり). 数値属性に関する相関ルールにおける最適領域ルールを求める手法の改良として, 2つ以上の属性をもつ最適数値属性相関ルールも提案されている[7].

キーワード複合化による行き詰まり発見ルールを多値属性相関ルールとして表現すると「一つの

トランザクション中に, キーワード集合XとYがともに出現しており, かつ, そのXとYの単独出現の場合の理解度は低いならば, 回答結果は不正解である」と記述することができ, 以下のように書ける. 式中「キーワード」は学習トランザクションの教材に付加されたキーワード集合を表わす.

$$(X \subseteq \text{キーワード} \wedge Y \subseteq \text{キーワード} \wedge X \text{ 単独出現の理解度} \geq 80\%) \wedge (Y \text{ 単独出現の理解度} \geq 80\%) \Rightarrow r = \text{FALSE.}$$

この相関ルールに関して, サポートがある閾値以上で, 確信度が大きい値となるXとYの組合せを発見したい. この相関ルールの確信度がある閾値以上, という条件は, 前節で定義した「キーワード複合化による行き詰まり発見ルール」(そこでは, X単独出現の理解度 $\geq 80\%$, かつY単独出現の理解度 $\geq 80\%$, かつXUY出現の理解度 $\leq 30\%$)と同義である(分からない場合の確信度であるから, $1 - 0.3 = 0.7$ より確信度70%以上とすればよい).

しかし上記相関ルールは以下の点で, 従来のデータマイニングの相関ルールと異なる. まず, 理解度 understand は, 確信度 confidence の変形と考えられるので, 相関ルールの条件の中に, 確信度の関係式が出現していると考えられる. よって, 興味の対象となるルールとは, 理解度(あるいは確信度)の変化率に着目した条件ルールとなる.

X 単独出現及び Y 単独出現の場合と、X と Y が複合化して出現した場合の理解度の変化の様子に着目することになる。

これにより相関ルールの複雑度は増加し、計算量も増えると予測されるので、この相関ルールを満たす X 及び Y を効率的に求める新たなアルゴリズムが必要とされる。

4.2 χ^2 による検定

前節では、データマイニングの数値属性に関する相関ルールという視点から考察を行ったが、次に、データマイニングに関する統計手法の観点から見ていく。

データマイニングの相関ルールの価値基準に関する議論として、 χ^2 を使った検定を応用して、明らかに価値のない相関ルールを取り除く方法が提案されている[11]。また、Brin らは[12]で、相関ルールの価値の評価基準として確信度及びサポートによる評価のみでは必ずしも適切でないことを指摘し、初めから共起するアイテム集合自体を発見の対象としている。ここでも相関の強度は χ^2 で測られている[13]。

我々もキーワード複合化による行き詰まり発見ルールの評価において、キーワード間の関連度判別に χ^2 値を用いるという方法を採用する。同様に χ^2 値を相関ルールの価値評価に用いた研究としては、志賀らによる複合商品の適切な商品選択順序評価の研究がある[14]。志賀らは、 χ^2 値は分割表(クロス表とも呼ぶ)の度数の大小に影響を受けるため、異なる部品クラス間の関連度を直接比較できない問題点があることを指摘し、度数による影響を取り除くため下記のようなクラメールの連関係数を用いている。式に示されるように、この式は χ^2 値に対して、分割表の度数 N による調整を行った値と言える。

$$V = \sqrt{\frac{\chi^2}{N \cdot \min(k, m)}}$$

我々が求めたい関連とは、キーワード集合 X 及び Y の関連度である。統計手法においては、複数の説明変数の個々に対して、正誤の結果に応じて 2×2 の分割表(これを2重クロス表と呼ぶ)を求め、 χ^2 値が、例えば2より大きいのであれば、説明変数(この場合1個)の値によって違いがある、といえる。例えば、説明変数「クレジットカードの有無」による従属変数「ダイレクトメールへの返答の有無」との関連が言える。関連度は双対尺度法(対応分析)を行うことにより距離として示

すことができ、それを1次元的にプロットすると関連の様子が視覚化できる[15]。

説明変数を2個に増やす場合は、3重クロス表を求めて χ^2 値を計算する。この場合の判定式は、判定式= χ^2 値-2×自由度、となるので、3重クロス表の場合、 χ^2 値は、 χ^2 値-2×3 と調整される[15]。

我々が対象としている問題においては、学習キーワードが含まれているか否かが問題であり、学習キーワードは多数存在する。つまり説明変数となる学習キーワードは多数存在する。一気に度数をあげて目的の説明変数の集合を求める方法にAID(多段層別分析)があり、具体的なソフトウェア商品としてSPSSのAnswerTree3.0がある[16]。AnswerTreeでは、分析アルゴリズムとしてCHAID(chi-squared Automatic Interaction Detector, χ^2 値に基づく交互作用検出法)を採用している。CHAID(チェイドと読む)は、 χ^2 乗検定によって目的変数(従属変数)と関連の強い要因(説明変数)を自動的に選択していく手法である[17]。出力は、ツリー状に表現されるため理解しやすいという利点がある。

以上の議論より、我々の対象とする学習履歴データマイニングの問題において、CHAIDはドリル問題の回答の正解・不正解を、決定づけるキーワード集合を決定するのに有効であると考えられる。

5. キーワード複合化による行き詰まり発見ルール抽出アルゴリズム

本節では、キーワード複合化による行き詰まり発見ルールを抽出するためのアルゴリズムを提案する。

我々はキーワード複合化による行き詰まり発見ルールの評価において、キーワード間の関連度判別に χ^2 値を用いるという方法を採用する。一般にX及びYのキーワードは長さmのキーワード集合である。しかし、関連度の強い2つのキーワード集合を見つけることは、キーワード集合をどのように細分化するかという問題になり、アルゴリズムが複雑化し、計算量が増加する。そこで、方針として、キーワード集合の長さは1とする。AIDアルゴリズムにより、複数の関連する説明変数(この場合、複数のキーワード)を求める。その結果として得られた2個以上のキーワードに対して、それら複数のキーワードが複合化して出現した場合に問題の難度が上昇すると解釈することとする。

教育支援システムにおいては、その学習者の学習進度フェーズ(あるいは学習単元)に依存して、

学習ターゲットとしている学習キーワード集合が存在する。例えば、高校数学Ⅲの「微分法の方程式への応用」においては、学習キーワードとして次のようなものがあげられる。「中間値の定理」、「パラメータを含む方程式」、「解の個数」、「接線問題」。それらのキーワード集合が顕在化していない場合は、教師サイドで予め、その単元に対する重要度の高い学習キーワード集合を選別しておく。学生の進度に応じて注目すべき学習キーワード集合は変化をしていく。常識として学生が取得していると仮定してよい基礎的な学習キーワードに関しては、教材に付加しない、こととする。この学習キーワード集合を以下では K^* と表わす。

提案する、キーワード複合化による行き詰まり発見ルール抽出アルゴリズムを以下に示す。

1. K^* の要素のうち、単独出現での理解度 $\geq 80\%$ を満たさない学習キーワードは、 K^* から除外する。
2. K^* の全要素を其々説明変数として、ドリル正誤関数を目的変数として、AID(多段層別分析)を行う。AIDの結果として、細分化されたグループが得られる。
3. 細分化されたグループの中から、 χ^2 値に関するスコアが最大なグループを選択する。関連度なしのグループだけになるまで以下を繰り返す。
4. その選択されたグループに説明変数として含まれていたキーワード集合を、互いに関連度の高いキーワード集合と置く。
5. 上記ステップ4の関連度の高いキーワード集合に対して、それらが同時に出現した場合の理解度を計算する。理解度の値が30%以下であれば、このキーワード集合を含むルールを作成し、キーワード複合化による行き詰まり発見ルールとして登録する。
6. 5で処理した関連度の高いキーワード集合のグループをグループ全体から除き、ステップ3に戻る。

上記ステップ5の理解度の計算においては、3.2節で行った理解度の定義を、キーワード数 n 個に拡張する必要がある。

6. まとめ

本論文では、教育支援システムにおけるデータマイニングにおいて、個々の学習キーワードが単独で出現するドリルであれば解けるのであるが、

学習キーワードが複合化して出現した場合にはドリルが解けなくなる、という状況を発見するために有効な相関ルールとはどのように定義すべきであるか、また、その抽出技法について考察した。まず、キーワード複合化による行き詰まり発見問題とは何かを説明し、問題解決に利用可能な新たな相関ルールの定式化を行った。そこで、相関を評価するための指標として理解度 **understand** を定義し、その理解度の値を用いて、キーワードの単独出現の場合と複合出現の場合とで理解度が異なる場合を発見する問題として定式化した。

我々の提案する設定問題は、理解度の変化率に注目するものであり、従来のデータマイニングのバスケット解析などのような、一つのレコードオカレンスのみを対象とし計算できる相関ルールとは本質的に異なることを説明した。本論文で定義した問題は、学習履歴データベースに対するキーワード複合化による行き詰まり発見ルールに関するものであったが、別の応用として、医療データベースにおいて、「食品 X 及び Y は単独では病気 Z の発生率に影響を及ぼさない、しかし、食品 X と Y を一緒に食べた場合、病気 Z の発生率が増加する」といったルールの発見にも利用できるなど応用範囲は広いと考える。

本稿では又、 χ^2 値による検定手法を用いて、キーワード複合化による行き詰まり問題で利用できる新たな相関ルールの抽出技法のアルゴリズムを提案した。今後は、提案したアルゴリズムが実利用において有効であるか、検証していきたい。

謝辞 本研究の一部は、平成15年度科研費基盤研究(C)(2)「マルチメディア教育支援システム eMath における教育用データベースの構築」(課題番号: 15606014, 代表: 白田由香利), 及び、平成14年度(財)放送文化基金研究「マルチメディア教育支援システム e-Math における対話エージェントの試作」による。ここに記して謝意を表します。

参考文献

- [1] 鍵本聡, “高校数学とっておき勉強法,” 講談社ブルーバックス B1243, 1999.
- [2] 白田由香利, “データベースを核とする e ラーニングシステム構築方法,” 日本データベース学会レターズ(DBSJ Letters), Vol. 1, No. 1., pp. 43-46, 2002.
- [3] 白田由香利, “経営数学用動画付き Web 教材を低コストで開発する手法,” 日本経営数学会誌, Vol.21, No.2, pp.71-81, 2002年11月.

- [4] 銀林浩, “算数・数学における理解,” 理解とは何か, 佐伯胖 (編), pp.37-68, 東京大学出版会, 東京, 1985.
- [5] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules,” Proc. VLDB, pp.487-499, 1994.
- [6] 喜連川, “データマイニングにおける相関ルール抽出技法,” 人工知能学会誌, vol. 12, No. 4, 1997.
- [7] 全眞嬉, D.Z. Chen, 加藤直樹, 徳山豪, “高次元最適ピラミッドを用いた数値属性ルールの生成とデータマイニングへの応用,” 日本データベース学会 Letters, vol.2, no.1, pp.83-86, May 2003.
- [8] 福田剛志, “数値属性の最適結合ルールを発見する効率的アルゴリズム,” 情報処理, vol.37, no.06, pp.945-953, June 1996.
- [9] T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama, “Mining Optimized Association Rules for Numeric Attributes,” Journal of Computer and System Sciences, vol.58, no.1, pp.1-12, 1999.
- [10] T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama, “Data Mining with Optimized Two-Dimensional Association Rules,” ACM Transaction of Database Systems, vol.26, no.2, pp.179-213, June 2001.
- [11] 福田剛志, 森下真一, “相関ルールの可視化について,” 電子情報通信学会技術研究報告, vol.95, no.81, pp.41-48, May 1995.
- [12] S. Brin, R. Motwani, and C. Silverstein, “Beyond market baskets: Generalizing association rules to correlations,” Proc. the ACM SIGMOD Conf. on Management of Data, pp.265-276, 1997.
- [13] 福田剛志, 森本康彦, 徳山豪, データマイニング, 共立出版, 東京, 2001.
- [14] 志賀隆之, 岩井原瑞徳, 上林弥彦, “組み合わせ制約の大域的特徴分析による複合商品の効率的な検索支援方法,” 第14回データ工学ワークショップ (DEWS2003), 加賀, 2003.
- [15] 上田太一, “データマイニング事例集”, 共立出版, 東京, 1988.
- [16] SPSS, “AnswerTree のアルゴリズムの概要”, http://www.spss.co.jp/product/ALL/A_tree/algo.htm.
- [17] 佐藤公征, “CHAID を営業にどう活かすか”, 日経リサーチレポート, <http://www.nikkei-r.co.jp/report/0002/13data.htm>.