

オンライン手書き数式パターンの収集

徳田創^{†1} レ・ドゥック・アイン^{†2} 斉福利^{†3}
蔡文杰^{†4} 中川正樹^{†3}

タブレットやスマートフォンの爆発的普及とともに、学習ソフトウェアで利用するオンライン手書き数式認識のニーズが高まっている。そして、この手書き数式認識技術の研究開発には標本パターンのデータベースが不可欠である。データベースとして、出現する手書き数式パターンをなるべくカバーすることと、できるだけ自然な手書きパターンを収集するために、数式を模写する Copy Writing と問題を解く過程で筆記する Free Writing という2つの方式で収集する。筆記時間と文字や記号の出現頻度を考えて、一人あたり、小学生で合計60問、中学生で45問、の数式を設定した。また、現在研究室で16人、小学生64人からの収集を行った。中学生のパターン収集も今後行う予定である。本稿ではオンライン手書き数式パターン収集の設計と実現について述べる。

キーワード：データベース、手書き数式パターン、オンライン認識

Collection of On-line Handwritten Mathematical Expressions

SO TOKUDA^{†1} ANH DUC LE^{†2} FULI QI^{†3}
WENJIE CAI^{†4} MASAKI NAKAGAWA^{†3}

Due to the rapid spread of touch interface devices such as tablets and smartphones, automatic on-line recognition of handwritten mathematical expressions (MEs) for educational applications is highly demanded. In order to develop effective recognition systems, however, a large amount of sample patterns is essential. In order to cover the vocabulary and syntax in MEs and to collect natural patterns, we employ two strategies: Copy Writing and Free Writing. In Copy writing, a writer copies a ME presented one by one. In Free Writing, he/she writes a mathematical expression while solving a problem. The mathematical equations to be collected from each student are set 60 equations for an elementary school student and 45 for a junior high school student, considering the occurrence of numerals, characters and symbols. Handwritten patterns have been already collected from 16 people from our laboratory and 64 elementary school students. Collection from junior high school students will be made planned soon. This paper presents the design and practice of collecting on-line handwritten MEs.

Keyword: Database, Handwritten Mathematical Expressions, On-line Recognition

1. はじめに

近年、タブレットの普及が進み4人に1人が持つようになった。これに伴い手書きでの入力が実用的になってきた。教育現場でのタブレットの導入も進んでいる。その際に算数や数学の課題の出題方式を選択式でなく、数式の筆記問題にすることができれば、児童・生徒のより深い理解を問うことができる。そして、それを自動採点できれば、迅速に採点することができ、児童・生徒の記憶が新しいうちに採点をフィードバックできる。また、自動採点は、児童・1生徒の自習にも活用できる。この実現のために手書き数式認識が必要である。

一般に、パターン認識システムの研究開発には、大量の標本パターンが欠かせない。オンライン手書き数式認識で

も同じである。この課題への関心の高まりから、手書き認識・文書認識の国際会議では毎年コンテストが開催され、標本パターンデータベースが公開されるようになってきている[1][2]。我々もこれらを活用している。しかし、それらは高等数学にまでわたる一般の数式であり、我々が研究開発している初等中等教育で学ぶ数式パターンは非常に少ない。

したがって、初等中等教育で利用される可能性ある教育ソフトウェアのためには、実際の小中学生あるいは高校生が答案などで筆記する手書き数式パターンの収集が不可欠である。そして、児童・生徒が躓きやすい分数や代数を含む手書き数式パターンを十分に含めることも重要である。

^{†1} 東京農工大学工学府産業技術専攻
Department of Industrial Technology and Innovation,
Tokyo University of Agriculture and Technology(TUAT)
^{†2} 東京農工大学工学府電気情報工学専攻
Department of Electronic and Information Engineering, TUAT

^{†3} 東京農工大学工学府情報工学専攻
Department of Computer and Information Sciences, TUAT
^{†4} 株式会社アイラボ
iLabo Ltd.

本稿では、初等中等教育での学習ソフトウェアで利用する手書き数式認識エンジンの研究開発のためのオンライン手書き数式パターン収集の設計と実現について述べる。

2. 設計方針

数式認識の研究に利用するためには、より自然なデータの収集が必要である。これを実現するために次の項目を設計の視点とした。

2.1. Copy Writing

Copy Writing では、よく出現する数式記号や構造をなるべくカバーするために、に収集する数式をまるごと模写してもらう。模写のため、筆記者の数学能力をあまり考慮する必要がなく、難易度の高い数式を設定することも可能である。

しかし、数式の模写を行っているため、実際に問題を解いている時の数式とは違ったパターンになる危険性はある。

2.2. Free Writing

上記 2.1 で述べたように Copy Writing では、出現頻度の高い記号や構造をきれなく収集することには向いているが、自然なパターンの収集には適していない。このため、より自然なパターンの収集を可能とする出題方式を用意する。

Free Writing は自然な手書き数式パターンを収集するための出題方式である。文章題を設定し、実際に筆記者には問題を解きながら数式を書いてもらう。そのため、より現実的な数式のオンライン手書きパターンの収集が期待できる。

しかしながら、文章題のため数式の筆跡パターンの収集は筆記者の数学能力に依存する。このため、入力される数式が望ましいものにならない可能性がある。

また、問題の出題の仕方によっては、意図していない非現実的な数式パターンの収集が行われてしまうこともある。

2.3. 収集する数式の表示

数式を表示することで筆記者に筆記する数式を表示する。

手書き文字パターンを採集する場合、その文字パターンを表示しないことで収集することでより自然なパターンが収集できるとされている。その時に筆記する文章を音声で筆記者に聞かせることでより自然なパターンが収集できる[3]。しかし、特に数式は音声のみでは筆記者が違う数式を筆記することや単純に数式を書けないということが予想される。

このため問題となる数式や模写する数式を入力枠上に表示する。

2.4. 個人情報の収集と取り扱い

筆記者には、統計処理と後の分析のために、プロフィール入力を依頼する。つまり、オンライン手書きパターンとして、氏名、ふり仮名、生年月日、性別、利き手、筆記手、母語、筆記の動機、学年を入力してもらう。システム側では、筆記開始日と終了日時を記録する。

この情報は、統計処理と分析のためだけに利用し、個人が特定できる情報は、一切公開しない。

3. 収集対象数式の設定

収集する数式パターンに出現する数字や記号の頻度を平均化する。このことで、各数字・記号の学習と評価を安定化することができる。一方、認識が難しい小数点などは多く含める。

また、手書き数式パターンの収集が目的であるため、数学的能力を問う難題の出題は控え、なるべく簡単な数式問題を用意する。

将来的に数式単体だけでなく単位を含んだ数式の認識も見込まれるため、単位を加えた数式も収集する数式に含める。

3.1. 小学生に対する問題

新学習指導要領を参照し、小学生の各学年で出現する数式および数式記号、単位を挙げ、それらを含むオンライン

手書き数式パターンの収集を行う。筆記時間を1時間に想定し、収集する Copy Writing 形式の問題を30問、Free Writing 形式の問題を30問作成した。

次にBNF記法を用いて含まれる数式を記述する。

$\langle NUM \rangle ::= \{0 \dots 9999\}$ $\langle ALB \rangle ::= \{a b x y\}$ $\langle DEC \rangle ::= \{0.01 \dots 100.99\}$ $\langle EQU1 \rangle ::= \langle NUM \rangle \{+ - \times \div\} \langle NUM \rangle \{= \} \langle NUM \rangle$ $\langle EQU2 \rangle ::= (\langle EQU1 \rangle) \{+ - \times \div\} \langle NUM \rangle \{= \} \langle NUM \rangle$ $\langle EQU3 \rangle ::= \langle NUM \rangle \langle ALB \rangle \{+ - \} \langle NUM \rangle \{+ - \}$ $\langle NUM \rangle \langle ALB \rangle \{= \} \langle NUM \rangle$ $\langle EQU4 \rangle ::= \langle NUM \rangle \{+ - \times \div\} \langle NUM \rangle \{< >\} \langle NUM \rangle$ $\langle EQU5 \rangle ::= \langle DEC \rangle \{+ - \times \div\} \langle DEC \rangle \{= \} \langle DEC \rangle$ $\langle EQU6 \rangle ::= \langle NUM \rangle \{+ \} \langle NUM \rangle \{= \} \langle NUM \rangle \{あまり\}$ $\langle NUM \rangle$ $\langle EQU7 \rangle ::= \langle NUM \rangle \frac{\langle NUM \rangle}{\langle NUM \rangle} \{+ - \} \frac{\langle NUM \rangle}{\langle NUM \rangle} \{= \} \langle NUM \rangle$ $\frac{\langle NUM \rangle}{\langle NUM \rangle}$ $\langle EQU8 \rangle ::= (\langle NUM \rangle \{+ - \} \langle NUM \rangle) \{ \times \div \} \{ \langle NUM \rangle \{= \} \langle NUM \rangle \langle NUM \rangle \{+ - \} \langle NUM \rangle \{= \} \langle NUM \rangle \frac{\langle NUM \rangle}{\langle NUM \rangle} \}$ $\langle EQU8 \rangle ::= (\langle DEC \rangle \{+ - \} \langle DEC \rangle) \{ \times \div \} \{ \langle DEC \rangle \langle DEC \rangle \{+ - \} \langle DEC \rangle \{= \} \langle DEC \rangle \}$	数学単位 km, m, cm, mm L, dL, mL g, Kg, t cm², m², km², ha, a 日, 時, 分 °, % cm³, m³
---	---

Figure 1 小学生問題のBNF記法

3.2. 中学生に対する問題

小学生の手書き数式パターン採集の結果を顧みて、時間内に回答可能な問題量を定めた。中学になり、数式問題がより高度になることを考慮し、問題数を減らした。その結果、Free Writing の数式問題の出題数を15問とした。また、中学になって出現する図式記号を加えた。

小学生のものに加え、新しく収集する数式は次のとおりである。

$\langle ALB \rangle ::= \{a b c A B C x y\}$ $\langle EQU \rangle ::= \langle NUM \rangle \{< >\} \langle NUM \rangle \sqrt{\langle NUM \rangle}$ $\langle EQU \rangle ::= \sqrt{\langle NUM \rangle} = \langle NUM \rangle$ $\langle EQU \rangle ::= \langle NUM \rangle \langle ALB \rangle \langle NUM \rangle \{+ - \} \langle NUM \rangle \langle ALB \rangle \{+ - \}$ $\langle NUM \rangle$ $\langle EQU \rangle ::= (\langle NUM \rangle \langle ALB \rangle \{+ - \} \langle NUM \rangle) (\langle NUM \rangle \langle ALB \rangle \{+ - \} \langle NUM \rangle) = \langle EQU \rangle$	数式記号 $\equiv, \cong, \neq, \leq, \geq$ $\pm, \sqrt{\quad}$ $', \perp, \parallel, \frown$ $\Delta, \angle, \square$
---	---

Figure 2 中学生問題のBNF記法

3.3. 問題の評価

作成した問題の評価を行う。収集対象数式の設定で述べたとおり、数式問題は出現する数字や記号の回数を平均化することでよりよい標本パターンデータベースになる。表1は、各学年の問題での数字出現回数を示す。

Table 1 小学生問題の数字出現回数

学年	Free		Copy	
	平均回数	標準誤差	平均回数	標準誤差
1	10.13	0.97	10.1	1.06
2	13.23	0.81	16.99	1.04
3	15.63	11.10	22.6	1.83
4	18.15	20.87	21.32	2.61

5	15.15	5.12	26.27	8.83
6	24.67	9.57	13.51	1.58

表1で、3、4年の標準偏差は他学年と比べて大きい。これは教育指導要領上、使える小数の範囲が限定されていることで0の出現回数が増えたためである。

Table 2 中学生問題の数字出現回数

学年	Free		Copy	
	平均回数	標準偏差	平均回数	標準偏差
1	4.81	0.96	15.62	5.70
2	9.48	1.62	24.50	3.05
3	6.52	2.99	21	4.066

4. 数式収集ツール

実際の収集の流れについて記述する。

4.1. 収集ツール

収集ツールは Windows で起動する。図に示す収集画面で収集を行う。数式の入力枠は入力する数式の長さに比例している。

筆記した数式に間違いがある場合の修正は 2 通りある。1 つは筆記した数式すべてを消すボタンである。もう 1 つは最後から 1 画ずつ消していくことができるボタンである。

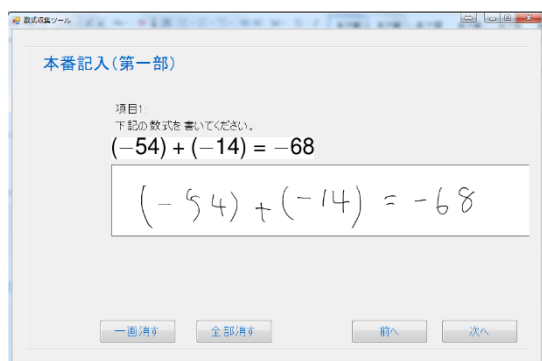


Figure 3 手書き数式パターン収集画面

4.2. 実際の収集

筆者には、収集システムを起動した際に図4に示す手

書き数式パターン収集の目的を読んでもらう。この目的に同意が得られた場合、次に記入の手順の説明を行い、プロフィールの入力を行ってもらう。

これらが終わった後に収集を開始する。ただしいきなり本番を行うのではなく、本番と同じ形式の筆記練習を行う。筆記者が筆記の練習を十分に行った後に本番へ移行する。

収集に要する時間は限られている。時間が経過、もしくはすべての問題の筆記が終わることで終了とする。

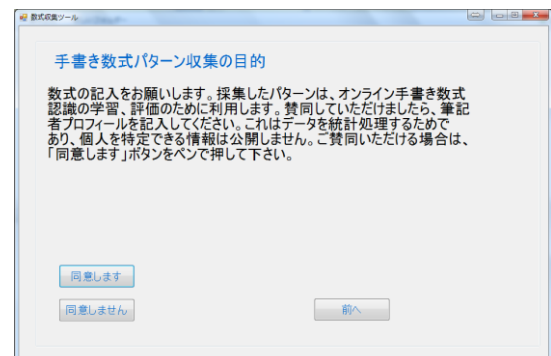


Figure 4 手書き数式パターン収集の目的

4.3. 収集パターンの後処理

収集した手書き数式パターンに対して後処理を行う。後処理では得られたパターンに対して、正解をタグ付けすることで、数式認識方式の学習や評価に利用できるようになる。収集した手書き数式パターンを数式識別器[4]にかける。出題問題から収集されることが予想される数式の正解と筆跡パターンの認識器での識別結果を照合し、正しく照合できたものはデータベースに追加し、できなかったものは手でタグ付けを行う。

4.4. 収集対象

収集対象は小学生と中学生である。また、当研究室の学生からの収集も行った。

5. 収集の実際と公開に向けて

本稿執筆段階で小学生 64 人からの収集を終えている。収集に要した時間は早い生徒で 30 分、遅い生徒で 45 分であった。また、当研究室の学生から小学生の数式パターン

を16人分収集した。

作成した中学生の数式問題の手書きパターン収集は今後開始する予定である。

手書き数式パターンの収集は今後も続けていく必要がある。また、今回は数式だけの収集であったが、認識技術の進歩によって、今後は図や文章の混在した証明問題のような数学問題の収集も考えられる。

ちなみに、収集した手書きパターンを用いて現状のオンライン手書き数式認識システム[5]の認識実験を行った。収集手書きパターンの集合を5セットに分割し、学習を4セットで行い、残りの1セットで評価する。5セットすべてが1回はテストセットになるように役割を変えて、テストセットに対する認識率の平均を求める。認識精度は85.15%であった。

6. おわりに

数式パターンの収集には数式問題を解くことが必要であり、時間がないと終わらないなどの課題がある。それは、算数、数学の問題が難しくなるからである。ただし、今回手書き数式パターンを収集した問題は、反復練習によって学習できる問題であったため、この問題は軽微であった。

まずは、反復練習の問題から始め、数式パターンの収集、手書き数式認識の開発、そして、自動採点の高度化が進めば、各人の数学能力の向上につながる。手書き数式パターンの収集と数式認識はポジティブフィードバックを構成して、手書きによる自学自習システム、あるいは、筆記答案の自動採点などの実用化につながることを期待したい。

今まで、児童・生徒はノート、教科書、そして参考書で勉強していた。しかし、タブレット1つで勉強する世代が今後登場するだろう。

今後、手書き数式認識を教育面で使用するにあたり、今回得た手書き数式パターンはその一助になることを願っている。

謝辞 本手書きパターン収集には東京書籍㈱のご協力を頂き、また、複数の小学校にご参加頂いた。中川研究室

の学生にも手書き数式パターンを提供いただいた。ここにあらためて深謝する。

参考文献

- 1) H.Mouchere, C.Viard-Gaudin, D.H.Kim,J.H.KimandU.Garain,"ICDAR 2013 CROHME - Third International Competition on Recognition of Online Mathematical Expressions", International Conference on Document Analysis and Recognition, ICDAR 2013, USA
- 2) H.Mouchere,C.Viard-Gaudin,D.H.Kim,J.H.KimandU.Garain,"ICHFR 2012 competition on recognition of online handwritten expressions (crohme2012)", International Conference on Frontiers in Handwriting Recognition, ICFHR 2012, Italy
- 3) 中川, 東山、山中、澤田、レーパン、秋山: "文章形式事態制限なしオンライン手書き文字パターンの収集と利用", 電子情報通信学会技術研究報告. PRU, パターン認識・理解 95(278), 43-48, 1995-09-28
- 4) Anh Duc Le, M.Nakagawa,:"A ground-truthing tool for making a database of on-line mathematical expression", IEICE PRMU technical report. 112(495), 147-150, 2013
- 5) Anh Duc Le, Van Phan T, Nakagawa M, "A System for Recognizing Online Handwritten Mathematical Expressions and Improvement of Structure Analysis", 11th IAPR International Workshop on Document Analysis Systems, pp.51,55, 7-10 April 2014