機械学習を用いた MathML 形式の数式データの意味推定

Inferring Meanings of Mathematical Expressions Using Machine Learning

石神 佑哉[*1], 宮崎 佳典[*2]
Yuya ISHIGAMI[*1], Yoshinori MIYAZAKI[*2]
[*1] 静岡大学情報学部
[*1]Faculty of Informatics, Shizuoka University
[*2] 静岡大学学術院情報学領域
[*2]College of Informatics, Shizuoka University
Email: yuuya3182@gmail.com

あらまし: 数式の意味推定を行う技術は数式検索技術や数式ライブラリを作成するときにおいて重要な役割を担う。例えば、数式の意味推定を数式検索技術に組み込むことで、数学的な意味を考慮した検索が可能になる。本研究では、W3C の勧告を受けている数式記述言語 MathML を対象に、数式を構成する各要素に分解し、機械学習を適用することで与式の意味推定を試みる。また、先行研究の課題である学習データ不足を解消するため、学習データ作成支援のためのツール開発を行う。
キーワード：数式，意味推定，機械学習，学習データ作成支援，MathML

1. 導入

数式の意味推定とは、数式中に用いられている識別子や数値と対応する数学的な定義や説明を推定することを指す。数式の意味推定を行う際には記号や数値がどのように使われているかを把握する必要がある。また、見た目上同じ数式であっても異なるデータ構造である場合を考慮しなければならない。

本研究では数式記述言語 MathML[(1)]を対象に意味推定を行う。MathML は HTML などと同様に木構造を持ち、定義された要素を用いて構成する。また、MathML には表記情報を持つ Presentation Markup と、意味情報を持つ Content Markup の 2 種類が存在する。本研究における数式の意味付けは Presentation Markup の各要素に対して、適切な Content Markup の要素をラベル付けすることと定義する。

先行研究では機械学習を用いて数式の意味推定を行った。しかし、先行研究の主な課題として学習データの不足が挙げられている。本研究では数式の意味付けを支援するツールの開発を行う。また、このツールから出力される数式データを使用し、さらに異種特徴量を用いた機械学習を行うことで数式の意味推定の精度向上を目指す。

2. 先行研究

Nghiem らの研究では Presentation Markup データに対して、機械学習を用いた数式の意味推定を行っている[(2)]。課題として MathML データ構造の違いについて考慮していないこと、機械学習の学習データ不足が課題として挙げられている。

渡部らの研究では Presentation Markup で記述された数式データの正規化を行う手法を提案した[(3)]。数式データの正規化とは Presentation Markup の各要素を編集し、見た目上同じ数式が同じデータ構造を持つことを指す。これによって、意味推定におけるデータ構造の違いを解消できる。また、正規化された数式データの使用例として、機械学習を使った数式の意味推定を行っている[(4)]。その課題として、学習データの不足や機械学習における特徴量の設計が挙げられている。

3. 意味付け支援ツール

先行研究から十分な量の学習データが喫緊の課題であることを受け、数式の意味付けを GUI 形式で行えるツールの開発を行った。意味付け用の数式フォーマットは(4)で用いられた手法に準じ、正規化された Presentation Markup に対して意味付けを行う。図 1 は開発したツールのスクリーンショットである。

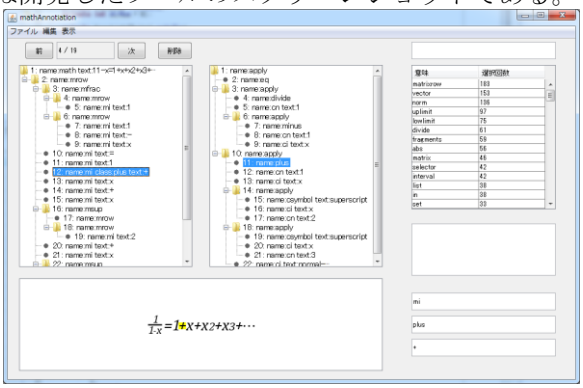

図 1 意味付け支援ツールのスクリーンショット

図 1 中の左枠には正規化された Presentation Markup が、左下部には対応する数式が表示される。また、表示されている Presentation Markup に対応する Content Markup 情報が存在する場合、中央枠内にそれが表示される。右部の表は Content Markup の要素表である。各 MathML の要素や表をマウスで選択することで数式の意味付けを視覚的に行うことができる。右下の 3 つのテキストボックスには現在選択されている Presentation Markup の要素名、ラベル、要素のテキストが各々表示される。特に、中央のテキストボックスを編集することで表などに存在しない要素を意味付けすることができる。

数式の意味付けを効率よく行うため本ツールは 2 つの機能を有し、1 つ目はルールベースで自動意味付けを行う機能である。例えば、数式中の記号「=」には、Content Markup における eq 要素を付加する。また、連続する記号「s」,「i」,「n」には、Content Markup における正弦を示す sin 要素を付加する。

2 つ目は選択要素の機械学習による意味推定である。選択した Presentation Markup の要素の特徴量を計算し機械学習を行う。結果、適した Content Markup の要素の候補が図 1 中の中央枠および右表にハイライト表示されるため、マニュアルで 1 つ 1 つ意味付けする作業に比べ、作業時間を削減できると期待される。なお、機械学習のアルゴリズムはランダムフォレストを用いた。この手法は後述する 4 節（機械学習の実験）でも採用している。

4. 機械学習の実験

本研究では機械学習による意味推定の精度を調査する実験を行った。表 1 は本研究で使用した特徴量をグループ化したもの（セット）である。

表 1 本研究で使用した特徴量のセット

ｾｯﾄ	特徴量	説明
A	1 つ前のノード	木を後順に辿った場合、1 つ（2 つ）前の各要素。
	2 つ前のノード	
B	親ノード	対象の親ノード及びその親ノード。
	親ノードの親ノード	
C	1 つ前の mi 要素（mi 要素は数値や記号を表記する Presentation Markup の要素）	木を後順に辿った場合、対象の 1 つ前に存在する mi 要素。

はじめに、本研究と(4)で用いた機械学習の手法の実験結果を比較した。実験データとして Wolfram Functions Site[5]（以降 WFS）から数式を収集した。WFS は収録している各数式に対し、Presentation Markup と Content Markup 両方のデータを提供している。WFS の Elementary Functions カテゴリの Sqrt ページに属する 195 の数式に対して意味付けを行った。実験手法は Presentation Markup の要素ごとに分類器を作成し、要素と表 1 中の全特徴量を入力する。その後、各分類器に対して 10 分割交差検定を行った。表 2 は記号と要素ごとの精度とベースラインをまとめたものである。なお、精度は分類器の出力と実際に意味付けされた Content Markup の要素が同一であった割合であり、ベースラインは各記号や要素において最も頻繁に意味付けされた Content Markup の要素の出現率である。

表 2 記号、要素ごとの精度の比較とベースライン

記号と要素	精度(%)		ベースライン(%)
	本研究	(4)の手法	
+	100.00	99.35	98.70
-	100.00	97.64	97.16
/	97.43	78.66	75.21
;	100.00	100.00	97.78
a	100.00	97.00	94.00
g	99.20	92.07	75.20
m	100.00	81.61	69.69
msup	99.05	92.01	91.98
mfrac	97.84	91.77	95.69
msub	96.50	95.10	89.51

表 2 中の msup 要素、msub 要素はそれぞれ上付き、下付き文字を表す。mfrac 要素は 2 つの要素を中央の括線の上下に表示するためのものである。mfrac 要素はほとんどが分数として使用されるが、偏微分としても使用されている。(4)では mfrac 要素の意味推定の精度がベースラインを下回ったが、本研究では改善に成功した。他の記号や要素でも(4)より精度が上回る結果となり、推定精度向上を実現した。

次に各特徴量の機械学習における貢献度を調査した。実験データは前述の実験で使用した WFS の数式データと NTCIR-12 MathIR Task Wikipedia Corpus[6]（以降、MathIR）を用いる。MathIR は数学や物理などの wikipedia のページから数式を集めたデータセットであり、数式 861 件に対して意味付けを行った。実験手法は使用する特徴量ごとに分類器を作成し、要素と特徴量を入力する。その後、10 分割交差検定を行った。なお、表 3 中の A, B, C の組み合わせは表 1 中の当該セットに属する特徴量を使用したことを示す(たとえば A∪B はセット A と B の両特徴量全体のことである)。

表 3 セットの各組み合せにおける精度

実験データ	単体セットの精度(%)		
	A	B	C
MathIR	94.20	88.65	92.81
WFS	97.75	94.03	97.23

実験データ	組み合わせセットの精度(%)			
	A∪B	A∪C	B∪C	A∪B∪C
MathIR	94.26	94.23	93.82	95.00
WFS	98.10	97.96	97.65	98.17

表 3 から、とりわけセット A、セット C の特徴量が機械学習の精度向上に貢献していることがわかる。セット B の特徴量だけ他と比べて精度が改善されなかった理由として、グループ化を行うための mrow 要素など、数式表記とは直接関係ない要素も特徴量として頻繁に使用されることが考えられる。

5. 今後の展望

1 つ目は、より効率よくデータ作成を行うための、意味付け支援ツールのインタフェース改良である。例として、ショートカットキー割り当てが考えられる。2 つ目は機械学習用の学習データを増やすことである。本研究では特定の web 上で提供されているデータのみを用いたが、今後は書籍をはじめ各種媒体を通じ、自然言語情報も加味した数式の意味推定を実現させていきたいと考えている。

参考文献

(1) MathML, https://www.w3.org/Math/
(2) Minh-Quoc Nghiem, et al.: A hybrid approach for semantic enrichment of MathML mathematical expressions, Intelligent Computer Mathematics Volume 7961 of the series Lecture Notes in Computer Science, pp. 278-287 (2013)
(3) 渡部孝幸,宮崎佳典:二次元の位置構造に着目した数式のパターンマッチング手法, 情報知識学会誌 Vol.22, No.3, pp. 253-271 (2012)
(4) 渡部孝幸,宮崎佳典:正規表現を用いた数式検索手法の提案, 情報処理学会論文誌 Vol.56, No.5, pp. 1417-1427 (2015)
(5) Wolfram Functions Site , http://functions.wolfram.com/
(6) NTCIR-12, MathIR, http://ntcir-math.nii.ac.jp/