

中学受験算数問題の自動分類システムの構築に向けて

岡田 優也

関西学院大学大学院 言語コミュニケーション文化研究科

ean82688@kwansei.ac.jp

1. はじめに

これまで、文書分類の技術を様々な分野に応用する試みが行われてきた。応用の例には、スパムメールのフィルタリングや Web ページの分類、特許文書への F ターム付与などがあり、文書分類の技術の適用範囲は広い[1][2]。

本研究は、中学受験算数の問題に対して文書分類の技術を応用し、問題の自動分類を行うシステムの構築を目指すものである。教育への ICT 活用が活発になりつつある現代において、これまで作成された多くの問題から自動的に情報を抽出し、分類するシステムの重要性は高い。また、一般的に算数の問題はこれまで文書分類の研究で使われてきたニュースや論文などと比べ、文書が短い傾向にある。このような特徴は、分類の精度に悪影響を及ぼす可能性が高いが、ネット上のビッグデータの解析など今後文書分類技術の応用が期待される分野においては、同様に短い文書である場合が多い[3]。そのため、短い文書の分類課題は重要な研究課題の一つである。

本研究では中学受験算数の問題に対する分類システムを作成する。また、作成した分類システムによって誤って分類された文書に対して分析を行い、今後の精度向上に向けた考察を行う。

2. 自動分類システム

2-1. 概要

本研究では、中学受験算数の問題集『受験全解算数』¹に収録されている問題から作成したデータベースを基に分類器の学習を行い、中学

受験算数問題の自動分類を行う。作成した自動分類システムの概要を図 1 に示す。

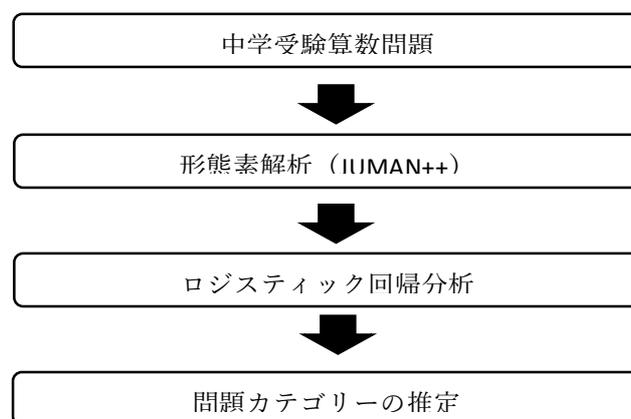


図 1: 自動分類システムの概要

本節では、以下に図 1 に示した自動分類システムについて詳説する。

2-2. データベース

データベースは、中学受験算数の問題集『受験全解算数』に収録されている問題から作成した。この問題集では、中学受験算数の問題を大きく 9 つのメタカテゴリーに分け、それをさらに 35 のカテゴリーに分類している。これらの問題について純粋な計算問題である「計算」というカテゴリーと明確にカテゴライズされていない発展問題を除いた 34 のカテゴリー、合計 1572 問の問題を抽出した。表 1 には使用した中学受験算数のカテゴリー階層とそれぞれの問題の数を示す。

¹ みくに出版 (1999 年発行)

表 1: カテゴリーごとの問題数

メタカテゴリー	カテゴリー	カテゴリー問題数	メタカテゴリー問題数
数の性質	逆数	24	167
	約数	33	
	倍数	41	
	数列の規則性	34	
	周期の規則性	35	
平面図形の性質と求積	平面図形の性質	49	204
	多角形の求積	54	
	円とおうぎ形の求積	59	
	図形の移動	42	
平面図形と比	割合からの求積	50	126
	相似比 (縮尺・かげ)	32	
	相似比と面積比	44	
立体図形の性質と体積	立体図形の性質	38	141
	図からの求積	46	
	水位と求積	31	
	水そうグラフ	26	
割合と比	割合と比	76	358
	食塩水の濃度	53	
	売り買いに関する問題	55	
	相当算	58	
	倍数算・分配算	67	
	仕事算・消去算・ニュートン算	49	
速さ	速さと進行グラフ	73	230
	旅人算と進行グラフ	68	
	通過算・時計算・流水算	69	
	図形上の点の移動	20	
場合の数・論理と推理	場合の数	72	110
	論理と推理	38	

文章題	平均算	31	106
	差集め算・過不足算	41	
	つるかめ算	34	
2量の関係	図形と規則性	54	131
	2量の関係とグラフ	44	
	統計の表とグラフ	33	
合計			1573

この合計 1572 問の問題について、図や表による表現を除き、文章による問題説明を個々の問題の文書として取り出しデータベースとした。これらの問題文書に対し、JUMAN++²による形態素解析を行い形態素に分割した。この時、活用のある形態素については原形への変換も行われている。このように分割された形態素を bag-of-words の形式に変換し、個々の問題文書の特徴ベクトルとした。

2-3. 問題カテゴリーの推定

作成した中学受験算数の問題データベースに対し、34 のカテゴリー名を教師データとするロジスティック回帰分析を行った。また、分類システムの精度を検証するために、分割数を 10 とする k-分割交差検証法を行った。

3. 実験

3.1 実験設定

2-2 節で、説明したデータベースをトレーニングデータおよびテストデータとし、ロジスティック回帰分析による分類器の構築と実験を行った。ロジスティック回帰分析における正則化パラメータは、グリッドサーチの結果によりもっとも高い精度の出た $C=1$ に設定し、分類器を作成した。

また、分類精度の検証には 10 分割の k-分割交差検証法を使用した。k-分割交差検証法とは、データを k 個の集合に分割し、そのうち $k-1$ 個の集合をトレーニングデータとし、残りの 1 個をテストデータとして一つの分類器を作成し、これを k 個の異なる組み合わせで精度

² <http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN++>

検証を行い最後にそれぞれの分類器の精度を平均することにより、分類精度を検証する手法である。

3-2. 実験結果

実験の結果、全体の正解率は平均約 67%、標準偏差約 11%という結果を得た。また、カテゴリごとの正解数、不正解数、分類精度を表 2 に示す。

表 2：カテゴリごとの判定結果

カテゴリ	正解数	不正解数	分類精度
逆数	19	13	0.59
約数	15	10	0.60
倍数	27	14	0.66
数列の規則性	19	20	0.49
周期の規則性	13	9	0.59
平面図形の性質	39	19	0.67
多角形の求積	24	31	0.44
円とおうぎ形の求積	45	27	0.63
図形の移動	27	14	0.66
割合からの求積	34	29	0.54
相似比（縮尺・かげ）	28	3	0.90
相似比と面積比	18	16	0.53
立体図形の性質	22	9	0.71
図からの求積	35	9	0.80
水位と求積	26	8	0.76
水そうグラフ	23	5	0.82
割合と比	32	56	0.36
食塩水の濃度	50	5	0.91
売り買いに関する問題	51	4	0.93
相当算	30	23	0.57
倍数算・分配算	43	31	0.58
仕事算・消去算・ニュートン算	32	12	0.73
速さと進行グラフ	54	21	0.72
旅人算と進行グラフ	54	20	0.73
通過算・時計算・流水算	54	18	0.75

図形上の点の移動	15	4	0.79
場合の数	64	16	0.80
論理と推理	7	13	0.35
平均算	27	5	0.84
差集め算・過不足算	25	20	0.56
つるかめ算	9	8	0.53
図形と規則性	40	16	0.71
2量の関係とグラフ	23	10	0.70
統計の表とグラフ	26	4	0.87
合計	1050	522	0.67

個々のカテゴリごとに分類精度を比較すると、「売り買いに関する問題」（約 93%）、「食塩水の濃度」（約 91%）、「相似比（縮尺・かげ）」（約 90%）の問題の分類精度が高く、「論理と推理」（約 35%）、「割合と比」（約 36%）、「多角形の求積」（約 44%）の問題の分類精度が低かった。また、この結果からカテゴリごとに分類精度にばらつきがでることが読み取れる。

4. 考察

なぜこのように予測するカテゴリごとに分類精度にばらつきが生じるのか考察するため、分類器において、それぞれのカテゴリを予測する際にどの形態素が重要であったかを調べた³。表 3 に分類精度の上位 3 つのカテゴリと下位 3 つのカテゴリにおいて重要度の高かった特徴量を 3 個ずつまとめて示す。

表 3：カテゴリごとにみる特徴量の重要度

上位 3 カテゴリ			下位 3 カテゴリ		
「売り買いに関する問題」			「論理と推理」		
順位	形態素	重み	順位	形態素	重み
1	品物	1.680	1	けた	1.056
2	定価	1.560	2	そのうち	0.817
3	仕入れる	1.030	3	当選	0.795

³ ロジスティック回帰分析において個々の特徴量に対して与えられる重みの大きさにより重要度を測定した。

「食塩水の濃度」			「割合と比」		
順位	形態素	重み	順位	形態素	重み
1	食塩	2.618	1	簡単だ	1.247
2	さとう	0.881	2	落とす	1.056
3	アルコール	0.655	3	入場	0.884
「相似比（縮尺・かげ）」			「多角形の求積」		
順位	形態素	重み	順位	形態素	重み
1	かげ	2.023	1	部分	1.002
2	実際	1.351	2	十二	0.976
3	地図	1.262	3	作る	0.974

表3より、上位3カテゴリと下位3カテゴリにおいて、分類に重要であると示された特徴量を比較すると、上位3カテゴリにおいては、「品物」、「食塩」、「かげ」などの個々のカテゴリに特徴的な形態素に1.5以上の重みが与えられているが、下位3カテゴリには、そのような大きな重みが与えられた形態素は存在しない。また、下位3カテゴリでは「けた」（一桁、二桁の「けた」）、「入場」、「部分」など類似する他のカテゴリにおいても頻繁に使われる形態素が重要な特徴量として上位に上がっている。上記2つの状況は、他の類似カテゴリとの識別を困難にする可能性がある。

上記の可能性を検証するため、個々のカテゴリの類似カテゴリにより形成されるメタカテゴリ内での誤りについて検証した。表4は、9つのメタカテゴリ内における誤りについてまとめたものである。

表4：メタカテゴリ内での誤り

メタカテゴリ	メタカテゴリ内の誤り	比率
数の性質	22	0.33
平面図形の性質と求積	32	0.35
平面図形と比	20	0.42
立体図形の性質と体積	13	0.42
割合と比	62	0.47
速さ	37	0.59

場合の数・論理と推理	9	0.31
文章題	10	0.30
2量の関係	1	0.03
合計	206	0.39

表4より、同じメタカテゴリ内の他のカテゴリに予測を行う誤りが、一定の割合で存在することがわかる。分類器全体では、予測の誤り(522個)のうち約39%がこの同じメタカテゴリへの誤り(206個)であるということがわかった。このことは、分類システムが極端に的外れな出力を行うことがないことを示している。

この結果から、中学受験算数の分類課題に対しては、形態素のuni-gramモデルのみでは、個々のカテゴリごとの特徴を十分に表現しきれない場合があること、そしてカテゴリの持つ階層的なカテゴリ構造を考慮する必要があることがわかった。

5. おわりに

本研究では、中学受験算数の問題を自動分類するためのシステムを提案した。また、分類システムの精度検証の結果、より精密な特徴量の設計と分類課題に内在する階層的カテゴリ構造を考慮した分類システムを作成する必要があることがわかった。これらの点に関しては、今後の課題とする。

6. 参考文献

- [1] 細川 誠介, 的場 隆一, “ディレクトリ型検索エンジンにおける Web ページのカテゴリ自動分類”, 言語処理学会第 23 回年次大会論文集, pp342-345, 2017
- [2] 佐々木 深, 細川 隆司, 西田 昌史, 西村 雅志, “特許文献中の重要語を用いた F ターム自動付与”, 言語処理学会第 23 回年次大会論文集, pp450-453, 2017
- [3] Song, Ge, et al. Short Text Classification: A Survey. *Journal of Multimedia*, 2014, 9.5.