NDLSH の細目付き件名標目に対して代表分類記号を機械学習によって付与できるか

谷口祥一[†]　　木村麻衣子[‡]
†慶應義塾大学文学部　　‡慶應義塾大学非常勤講師
taniguchi@z2.keio.jp　　mayizi@keio.jp

［抄録］NDLSH において、件名標目に対して概念上で対応する NDC 新訂 9 版の分類記号が「代表分類記号」として示されている。ただし、その範囲は限られており、細目を伴った殆どの件名（主標目＋細目）には代表分類記号が付与されていない。本研究は、NDL 作成の書誌レコードに付与された細目付き件名標目と NDC 分類記号の組み合わせの中から、代表分類記号となりうるものを機械学習によって同定することを試みた。人手によって妥当性を判定したデータ集合を準備し、複数の属性集合と機械学習法を組み合わせて実験を行い、性能を評価した。

１．はじめに

国立国会図書館件名標目表（NDLSH）において、一部の件名標目に対して、概念上で対応する日本十進分類法（NDC）新訂 9 版の分類記号が「代表分類記号」として示されている。たとえば、件名標目「メタデータ」は、「014」（資料の収集、整理、保管）と「014.3」（目録法；記述目録法）という代表分類記号をもつ。この代表分類記号による件名標目と分類記号との対応づけは、多様な活用法が期待できるが、現時点では代表分類記号の付与範囲は限られており、細目を伴った件名（「主標目＋細目」）には原則的に代表分類記号が付与されていない。本研究は、国立国会図書館（NDL）作成の書誌レコードに付与された件名標目と NDC 分類記号の組み合わせの中から、細目付き件名の代表分類記号となりうるものを機械学習によって同定することを試みる。

なお、個別の書誌レコードに付与された件名標目と分類記号の組み合わせが, 当該件名標目の代表分類記号とどの程度一致するのかについて詳細を調べた先行研究があるが [1]、本研究はそこで対象外とされた代表分類記号をもたない件名標目と分類記号の組み合わせに対する調査でもある。機械学習を件名標目と分類記号の対応づけに適用した石田による研究はあるが、その研究の射程や方式は異なる [2]。

２．対象データ

1997 年から 2014 年 3 月分までの NDL 作成の書誌レコードから、NDC 新訂 9 版の分類記号と NDLSH 件名標目のペアで、かつ普通件名であるものを抽出した。地名件名、固有名件名などは、NDC による代表分類記号が殆ど付与されていないため、対象から外した。また、NDLSH 件名に対する代表分類記号は、Web NDL Authorities から取得した。

その結果、細目付き普通件名（主標目＋細目）で代表分類記号をもたない、かつ主標目には代表分類記号があるものは、63,578 件名（7,484 主標目）、件名・分類記号ペア数（異なり数）99,264（平均 1.56、SD 2.17）が抽出された。

ここからさらに、人手によりペアの妥当性判定（細目付き件名の代表分類記号の同定）を行うため、主標目ごとに細目付きのペアをまとめた単位で系統抽出による標本抽出を実施し、主標目異なり数 474、件名（主標目＋細目）異なり数 3,266、分類記号とのペア異なり数 5,022 を抽出した。この標本集合は、件名ごとに平均 1.54（SD 2.18）の分類記号とペアを形成し、主標目ごとには平均 6.89（SD 20.69）の細目をもち、平均 10.59（SD 39.02）の分類記号とペアを形成していた。

この標本抽出された件名・分類記号ペアに対して、発表者 2 名それぞれが妥当性を判定し、その後、判定が分かれたケースについては協議し最終的な判定を確定した。判定はそれぞれの件名と分類記号の組み合わせが概念的に対応するか、つまり代表分類記号としてよいかという判定である。換言すれば、主標目には既に代表分類記号が付与されているため、細目が付加されたときに、それらとは異なる代表分類記号が対応づけられるべきか否かという判定でもある。なお、判定は NDL の分類基準、件名作業指針、および一部の付与実績に従って行った。

判定結果は、「適切」、「準適切」、「不適」の 3 区分とした。準適切とは、直接的な対応づけは不自然であるが、適用を拡大して捉えたときには適切とも考えられるケースを割り当てた。判定結果の集計を表 1 に示した。併せて、主標目の代表分類記号との完全一致・前方一致・不一致のクロス集計も示した。完全一致のものす

べてが「適切」と判定されているわけではなく、56 ペアは「不適」であった。

３．機械学習の適用実験

機械学習ツール Weka を用いて、複数の機械学習法を上記の判定済みデータに対して適用し、その性能値を求める実験を行った。

(1)学習用データ、評価用データ

人手による判定済みデータ（件名・分類記号ペア）を、系統抽出法により主標目の単位で 3 分割し、学習用データと評価用データとする 3 交差検証法を採用した。代表分類記号をもつデータ集合（102,647 ペア；例外を除いて件名は細目なし）を学習用データに追加し用いた「学習データ方式 1」と、人手による判定済みデータのみを学習用データに用いる「学習データ方式 2」とを採用した。

(2)属性（特徴素）集合

個々の件名・分類記号ペアに対する属性集合は、最も広範な「属性集合 1」（37 属性）から、最小限の属性に限定した「属性集合 3」（12 属性）まで 3 段階を設けた。いずれの属性値も機械的に生成できるものである。たとえば、属性集合 3 とは、ペア ID、主標目 ID、細目 ID、NDC 分類記号、主標目の代表分類記号との一致区分（完全一致、前方一致、不一致）、ペア出現回数、ペア出現率、ペア出現回数×ペア出現率、レコード内先頭出現ペア出現回数、先頭出現ペア出現率、先頭出現ペア出現回数×先頭出現ペア出現率、それに正解情報とした（いずれも数値属性）。正解情報は、当該件名・分類記号ペアが「適切」か「不適」という 2 クラスとした。属性集合 1 と 2 は、上記の属性群にさらにペア共起率（Jaccard 係数、Dice 係数）、件名ベースの平均情報量など、多様な値を属性として加えたものとした。最小属性集合である属性集合 3 の決定は、Weka の属性選択機能を用いながら行った。

(3)適用する機械学習法

下記の代表的な学習法 7 つを採用した（名称はいずれも Weka におけるもの）。

a) AdaBoostM1：アンサンブル学習のうち、ブースティングに属する学習法
b) ConjunctiveRule：single conjunctive rule（連言ルール）学習法
c) J48：決定木 C4.5
d) Logistic：ロジスティック識別
e) NaiveBayes：ナイーブベイズ識別
f) RandomForest：アンサンブル学習のうち、ランダムフォレスト学習法
g) SMO：サポートベクタマシン（SVM）

なお、それぞれの学習法の適用においては、Weka のデフォルト設定のまま実行し、個別のパラメータ等の調整はしていない。

４．実験結果と考察

(1)学習データ方式 1

代表分類記号をもつデータ集合を学習用データに追加した学習データ方式 1 において、a)属性集合 1,2,3 と、b)人手による判定「準適切」を正解クラス「適切」・「不適」のいずれに入れるかという組み合わせで実験を行った。表 2 に機械学習法ごとに、得られた精度、再現率、F 値を示した。属性集合 2 については、属

表１．人手による判定結果

	主標目の代表分類記号との一致・不一致							
判定結果	合計ペア数	完全一致	前方一致	不一致	ペア出現回数	完全一致	前方一致	不一致
「適切」	2,489 49.6%	1,322 26.3%	813 16.2%	354 7.0%	9,260 65.8%	4,704 33.4%	3,667 26.0%	889 6.3%
「準適切」	104 2.1%	0 0.0%	10 0.2%	94 1.9%	225 1.6%	0 0.0%	42 0.3%	183 1.3%
「不適」	2,429 48.4%	56 1.1%	348 6.9%	2,025 40.3%	4,592 32.6%	110 0.8%	733 5.2%	3,749 26.6%
合計	5,022 100%	1,378 27.4%	1,171 23.3%	2,473 49.2%	14,077 100%	4,814 34.2%	4,442 31.6%	4,821 34.2%
参考：								
代表分類記号なし全数	99,264 100%	25,995 26.2%	28,556 28.8%	44,713 45.0%	333,406 100%	113,246 34.0%	123,977 37.2%	96,183 28.8%
代表分類記号あり全数	102,647 100%	14,088 13.7%	7,175 7.0%	81,384 79.3%	675,642 100%	383,277 56.7%	67,321 10.0%	225,044 33.3%

表２．学習データ方式１の実験結果

	Ada BoostM1	Conjunctive Rule	J48	Logistic	Naïve Bayes	Random Forest	SMO	機械学習法の多数決
①属性集合 1、かつ人手による判定「準適切」は正解クラス「不適」とする								
精度	0.831	0.837	0.895*	0.873*	0.599	0.887*	0.885*	0.884*
再現率	0.897*	0.855	0.802	0.798	0.982*	0.829	0.851	0.861*
F 値	0.862*	0.846	0.846	0.834	0.744	0.857*	0.867*	0.872*
②属性集合 1、かつ人手による判定「準適切」は正解クラス「適切」とする								
精度	0.880*	0.841	0.898*	0.854*	0.608	0.873*	0.874*	0.888*
再現率	0.810	0.825	0.793	0.821	0.986*	0.817	0.835*	0.823
F 値	0.844*	0.833	0.842*	0.837*	0.752	0.844*	0.854*	0.854*
③属性集合 3、かつ人手による判定「準適切」は正解クラス「不適」とする								
精度	0.831	0.838	0.898*	0.821	0.507	0.881*	0.829	0.840*
再現率	0.899*	0.858	0.828	0.816	0.999*	0.847	0.798	0.911*
F 値	0.864*	0.848	0.862*	0.818	0.672	0.863*	0.813	0.874*
④属性集合 3、かつ人手による判定「準適切」は正解クラス「適切」とする								
精度	0.880*	0.842	0.884*	0.839	0.520	0.873*	0.853*	0.860*
再現率	0.812	0.827	0.809	0.815	0.999*	0.823	0.793	0.862*
F 値	0.845*	0.834	0.845*	0.827	0.684	0.847*	0.822	0.861*

*： 代表分類記号との一致区分のみによる予測性能よりも高い値

性集合 1,3 と大きな性能値の変動がないため掲載を省略した。

属性集合 1 かつ人手による判定「準適切」を正解クラス「不適」とした場合（ケース①）、ナイーブベイズ以外はほぼ同等の性能を示しており、精度 0.831~0.895、再現率 0.798~0.897、F 値 0.834~0.867 であった。他方、ナイーブベイズは精度が低く（0.599）、再現率が高い（0.982）結果となった。人手による判定「準適切」を正解クラス「適切」に入れた場合（②）にも、ほぼ同様な結果であるが、単純な判定とはならないものが増えた結果、再現率そして F 値が低下した学習法が増えている。

属性集合 3 を採用したときには（③と④）、属性集合 1 の場合に比べて性能値が上昇した学習法もあれば、減少した学習法も見られた。全体的な傾向および性能値の範囲は、属性集合 1 の場合と比べて大きく変化していない。

これらの結果は性能値の値自体としてはかなり高い数字に見える。しかし、表 1 に示した通り、人手による判定結果は、主標目の代表分類記号との一致区分（完全一致、前方一致、不一致）と相当程度に相関が見られるため、この単一属性のみによる正解の予測性能と比較することが適切であろう。a)人手による判定「準適切」を正解クラス「不適」とした場合、上記の単一属性による予測性能は精度 0.838、再現率 0.858、F 値 0.848 となった。なお、この値は前方一致を含めて主標目の代表分類記号と一致と見なしている（完全一致のみ一致と見なしたときには性能値が低下する）。同様に、b)人手による判定「準適切」を正解クラス「適切」に入れた場合、上記の単一属性による予測性能は精度 0.842、再現率 0.827、F 値 0.834 となった。ケース①・③と a)、②・④と b)を比較し、機械学習による性能値が上回ったときには、表 2 において「*」を付した。これを見ると、上記の単一属性による予測性能を上回った学習法も多いとはいえ、性能上昇は限られたものであった。

(2)学習データ方式 2

人手による判定済みデータのみを学習用データに用いる学習データ方式 2 においても、属性集合 1 かつ人手による判定「準適切」を正解クラス「不適」とした場合（ケース⑤）、ナイーブベイズとそれ以外では異なる傾向となった（表 3）。ナイーブベイズでは、精度が高く、再現率が低いという結果、すなわち①とは逆の結果となった。それ以外の学習法においては、①と比べて性能値が上昇しているものもあれば、減少しているものもある。

人手による判定「準適切」を正解クラス「適切」に加えた場合（⑥）、⑤に比べて性能値が上昇している学習法が多い。他方、属性集合 3 の場合にも、人手による判定「準適切」を正解クラス「不適」から「適切」に変更したとき（⑦から⑧へ）、性能上昇が見られるものが多い。属性集合 1 と 3 の比較（⑤と⑦、⑥と⑧）では、属性集合 3 が全般的に性能低下となる場合が多く見られた。

表３．学習データ方式２の実験結果

	Ada BoostM1	Conjunctive Rule	J48	Logistic	Naïve Bayes	Random Forest	SMO	機械学習法の多数決
⑤属性集合 1、かつ人手による判定「準適切」は正解クラス「不適」とする								
精度	0.855*	0.823	0.906*	0.864*	0.828	0.877*	0.863*	0.881*
再現率	0.873*	0.875*	0.819	0.818	0.397	0.837	0.821	0.844
F 値	0.864*	0.848	0.860*	0.840	0.536	0.857*	0.842	0.862*
⑥属性集合 1、かつ人手による判定「準適切」は正解クラス「適切」とする								
精度	0.864*	0.875*	0.893*	0.872*	0.846*	0.863*	0.862*	0.886*
再現率	0.870*	0.788	0.800	0.827	0.419	0.858*	0.845*	0.834*
F 値	0.867*	0.829	0.844*	0.849*	0.560	0.861*	0.853*	0.859*
⑦属性集合 3、かつ人手による判定「準適切」は正解クラス「不適」とする								
精度	0.851*	0.838	0.895*	0.817	0.860*	0.868*	0.779	0.871*
再現率	0.859*	0.858	0.829	0.834	0.294	0.836	0.842	0.852
F 値	0.855*	0.848	0.861*	0.825	0.439	0.852*	0.809	0.861*
⑧属性集合 3、かつ人手による判定「準適切」は正解クラス「適切」とする								
精度	0.852*	0.875*	0.867*	0.836	0.889*	0.860*	0.844*	0.863*
再現率	0.896*	0.788	0.835*	0.829*	0.368	0.836*	0.795	0.820
F 値	0.874*	0.829	0.851*	0.832	0.520	0.848*	0.819	0.841*

*： 代表分類記号との一致区分のみによる予測性能よりも高い値

表４. 実験結果と主標目の代表分類記号との一致・不一致：多数決方式の実験結果①の場合

			完全一致			前方一致			不一致		
	ペア数	予測誤り数	ペア数	予測正解	予測誤り	ペア数	予測正解	予測誤り	ペア数	予測正解	予測誤り
「適切」	2,489	345	1,322	1,322	0	813	706	107	354	116	238
「準適切」	104	12	0	0	0	10	7	3	94	85	9
「不適」	2,429	270	56	0	56	348	222	126	2,025	1,937	88
合計	5,022	627	1,378	1,322	56	1,171	935	236	2,473	2,138	335

学習データ方式 1 から方式 2 に切り替えたときには（①と⑤、②と⑥、③と⑦、④と⑧）、対応するそれぞれの機械学習適用結果において、性能上昇が見られる場合もあれば低下を見せているものもある。これらから、総じて、代表分類記号があるデータ（例外を除き細目をもたない）は、細目付き件名の学習用データとしては有効とは言い難い。

また、主標目の代表分類記号との一致区分のみによる予測性能を上回ったものは、表 3 において「*」を付した。

(3)機械学習法の多数決方式

各ペアに対する個々の予測結果を 7 つの機械学習法の間で多数決を行い、ペアごとの最終的な予測結果とする方式で性能評価を行った（表 2 および表 3 の最右欄）。7 つの機械学習法のうち 4 つ以上が「適切」と予測したときに最終予測結果「適切」とし、人手による判定結果と照合する方式である。実験の結果、個別学習法の性能値（特に F 値）を上回る結果が大半となったが、全体的には大きな変化とは認めがたい。また、この多数決方式どうしの性能比較（たとえば①と②、③と④の比較など）においては、性能が低下する場合が大半であった。

併せて、代表分類記号との完全一致、前方一致、不一致それぞれにおいて、この多数決方式の予測結果が正解となる場合および不正解となる場合を集計した。表 4 に、実験結果①のケースについて集計結果を示した。

以上の結果から、機械学習による代表分類記号の同定は、今回の学習用データ量では、かなり困難と結論づけられよう。

注

1) 谷口祥一, 尾形沙由美. NDLSH における NDC 代表分類記号と書誌レコードの分類記号はどの程度一致するのか. *Library and Information Science.* No.75, 2016, p.37-66.

2) 石田栄美. 日本十進分類法と基本件名標目の相互マッピングの試み. 文化情報学. Vol.12, No.1, 2005, p.1-11.