機械学習によってNDLSH細目付き件名標目に対するNDC代表分類記号を同定する試み

谷口 祥一（慶應義塾大学文学部）taniguchi@z2.keio.jp
木村 麻衣子（日本学術振興会特別研究員 RPD(東京大学東洋文化研究所)）mayizi@keio.jp

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）と国立国会図書館分類法の分類記号が「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，NDC 新訂 9 版の「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，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示されていない。

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。具体的には，既出の組み合わせから，教師あり機械学習によって適切な組み合わせ，すなわち細目付き件名に対応する適切な分類記号を同定することを試みる。

なお，本研究の中間段階で，実験結果の概略を既に報告している[1]。本発表は学習・評価用データ件数をその後倍増させ，信頼性を増した上での実験結果，さらには新たな実験条件を追加した上での実験結果の公表を意図している。

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（件名当たり平均 1.6，SD 2.2）であった。

ここからさらに，人手によりペアの適切性判定（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，主標目数 845，件名（主標目＋細目）数 6,502，件名・分類記号ペア数 10,053 を抽出した。この標本集合は，件名ごとに平均 1.6（SD 2.1）の分類記号とペアを形成し，主標目ごとには平均 7.7 (SD 21.8) の細目をもち，平均 11.9 (SD 38.9）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 名それぞれが適切性を判定し，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事例を割り当てた。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 2,557 ペアすべてが「適切」と判定されているわけではなく，117 ペアは「不適」であった。

他方，機械学習の適用実験においては，「適切」か「不適」かの 2 区分による予測および評価とするため，人手による判定「準適切」とされたペアは「不適」とされたペアに組み入れた（すなわち，いずれも負例とする）実験とする。

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。a)この人手による判定済みデータを学習用データに用いた「学習データ方式 1」に加えて，b)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

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」	4,977	2,434	1,918	625	27,326	9,743	15,992	1,591
	49.5%	24.2%	19.1%	6.2%	72.2%	25.7%	42.3%	4.2%
「準適切」または「不適」	5,076	123	905	4,048	10,515	222	2,500	7,793
	50.5%	1.2%	9.0%	40.3%	27.8%	0.6%	6.6%	20.6%
「準適切」	411	6	113	292	961	12	245	704
	4.1%	0.1%	1.1%	2.9%	2.5%	0.0%	0.6%	1.9%
「不適」	4,665	117	792	3,756	9,554	210	2,255	7,089
	46.4%	1.2%	7.9%	37.4%	25.2%	0.6%	6.0%	18.7%
合計	10,053	2,557	2,823	4,673	37,841	9,965	18,492	9,384
	100%	25.4%	28.1%	46.5%	100%	26.3%	48.9%	24.8%
参考：								
代表分類記号なし全件	99,264	25,995	28,556	44,713	333,406	113,246	123,977	96,183
	100%	26.2%	28.8%	45.0%	100%	34.0%	37.2%	28.8%
代表分類記号あり全件	102,647	14,088	7,175	81,384	675,642	383,277	67,321	225,044
	100%	13.7%	7.0%	79.3%	100%	56.7%	10.0%	33.3%

細目なし）を学習用データに追加して用いた「学習データ方式 2」，あるいは逆に，c)学習用データの件数を削減した「学習用データ方式 3」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），その中間にある「属性集合 2」（25 属性）という 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り，正解クラス情報以外はすべて数値属性として表現した。

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，それに正解クラス情報とした。正解クラス情報は，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。

「先頭出現ペア」とは，書誌レコード内で先頭に出現した件名と分類記号の組み合わせに限定したペアであり，これらにかかわる属性群の採用にとどまらず，これらの属性群のみから構成する属性集合 4（24 属性）と属性集合 5（9 属性）を併せて設けた。

(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) BayesNet：ベイズ識別のうち，ベイジアンネットワークを用いた学習法
c) DecisionTable：ルールベースの学習のうち，決定表を用いた学習法
d) J48：決定木 C4.5
e) Logistic：ロジスティック識別
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）

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，個別の設定値等の調整はしていない。なお，b)と c)は前回報告した実験から変更している。

４．実験結果と考察

(1)実験 1：学習データ方式 1

学習用データに人手による判定済みデータのみを用い，まず属性集合 1～3 における性能値の変化および機械学習法による性能の相違を検証した。各属性集合において個々の機械学習法がもたらした性能値である正解率（accuracy）と F 値のマイクロ平均を，表 2 に示した。

正解率の最大値は 0.876（RandomForest かつ属性集合 1）であり，最小値は 0.805（SMO かつ属性集合 3）であった。7 つの機械学習法のうち 5 つにおいて，属性集合 1 が他に比べて高い値を示したが，残り 2 つは属性集合 3 が

表２．実験１（学習データ方式 1）の結果

	Ada BoostM1	Bayes Net	Decision Table	J48	Logistic	Random Forest	SMO	多数決方式
①属性集合 1								
正解率	0.868*	0.815	0.831	0.856*	0.833	0.876*	0.838*	0.872*
F 値	0.870*	0.820	0.830	0.854*	0.828	0.872*	0.834	0.871*
②属性集合 2								
正解率	0.852*	0.826	0.833	0.834	0.830	0.873*	0.830	0.863*
F 値	0.853*	0.828	0.835	0.830	0.825	0.869*	0.826	0.863*
③属性集合 3								
正解率	0.855*	0.840*	0.848*	0.838*	0.827	0.840*	0.805	0.857*
F 値	0.855*	0.838	0.844*	0.828	0.829	0.856*	0.809	0.857*
④属性集合 4								
正解率	0.861*	0.808	0.809	0.869*	0.822	0.873*	0.822	0.868*
F 値	0.858*	0.815	0.794	0.864*	0.815	0.868*	0.818	0.865*
⑤属性集合 5								
正解率	0.863*	0.835	0.857*	0.857*	0.816	0.850*	0.828	0.859*
F 値	0.862*	0.835	0.848*	0.852*	0.819	0.846*	0.824	0.857*

*： 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

高い値を示した。F 値を見ると，最大値は 0.872（RandomForest かつ属性集合 1），最小値は 0.809（SMO かつ属性集合 3）であり，正解率と同じケースが該当した。相対的には，正解率，F 値の両者において RandomForest と AdaBoostM1 の性能値が高いといえよう。逆に，極端に性能が低い機械学習法は見られないといってよかろう。

属性集合による性能の変化を見てみると，全体的には最も多くの属性を採用した属性集合 1 の性能が僅かながら高く，次に最小の構成をとる属性集合 3 の性能が高い結果となった。

実験には，各ペアに対する個々の機械学習法による予測を 7 つの機械学習法で多数決し最終的な予測とする方式（多数決方式）を加えた（表 2 の最右欄）。個別学習法の性能値を上回るケースが大半となったが，他と比べて最高値が得られたケースは限られていた。

それぞれの機械学習法は，いずれも高い性能値を示したように見える。しかし，表 1 に示した通り，人手による判定結果は，件名を構成する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た。そのため，この単一の属性のみによる正解クラスの予測性能，すなわち主標目の代表分類記号と一致したときには単純に正例と予測し，それ以外は負例と予測したときの性能値と比較することが適切であろう。完全一致および前方一致を代表分類記号と一致として扱ったときには，正解率 0.836，F 値 0.840 となり，前方一致をすべて適切（正例）と見なしたことになり，再現率を重視した結果となる。これらの性能値を実験のベースラインに設定し，それに比べて個々の機械学習法による性能が上回ったケースには，表 2 において「*」を付した。この結果，上記の単一属性による予測性能を上回ったケースも多いとはいえ，性能上昇の幅は限られており，他方ではそれを下回るケースもいくつか存在した。

先頭出現ペアにかかわる属性群に限定した属性集合 4 と 5 の比較では，機械学習法によっていずれの性能値が高いかが分かれたが，RandomForest，AdaBoostM1，そして J48 の性能が相対的に高い結果となった。属性集合 1～3 の場合の性能値と比べたときには，属性集合 4・5 の性能値が劣る場合（Logistic など）もあれば，反対に性能値が上回る場合（J48 など）もあり，一定の傾向はない。先頭出現のペアに焦点を当てた属性集合 4･5 に変更しても，性能値を大きく押し上げる結果は得られなかったといえよう。

(2)実験 2：学習データ方式 2

殆どが細目のない件名からなり，代表分類記号をもつデータを，学習用データに加えた学習データ方式 2 の実験では，全体的にはやはり AdaBoostM1 や RandomForest などの性能値が高い（表 3）。

属性集合 1・3 について，学習データ方式 2 で得られた性能値（表 3 の⑥・⑦）と学習データ方式 1 での性能値（表 2 の①・③）とを比べてみると，学習データ方式 2 が性能上昇を見せたケースと逆に性能低下を見せたケースの両方がある。正解率の場合，属性集合 1（⑥と①）では性能値の上昇と低下はおおむね半々で

表３．実験２（学習データ方式 2）の結果

	Ada BoostM1	Bayes Net	Decision Table	J48	Logistic	Random Forest	SMO	多数決方式
⑥属性集合 1								
正解率	0.842*	0.817	0.835	0.838*	0.848*	0.876*	0.850*	0.858*
F 値	0.831	0.821	0.825	0.825	0.843*	0.871*	0.837	0.849*
⑦属性集合 3								
正解率	0.865*	0.851*	0.846*	0.835	0.830	0.864*	0.827	0.866*
F 値	0.859*	0.858*	0.842*	0.825	0.829	0.858*	0.823	0.864*
⑧属性集合 4								
正解率	0.836	0.808	0.837*	0.850*	0.820	0.856*	0.828	0.853*
F 値	0.835	0.815	0.825	0.839	0.815	0.846*	0.819	0.848*
⑨属性集合 5								
正解率	0.862	0.843	0.861	0.853	0.830	0.855	0.828	0.866*
F 値	0.860*	0.852*	0.856*	0.885*	0.827	0.887*	0.841*	0.865*

*： 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３（学習データ方式 3 かつ属性集合 1）の結果

	Ada BoostM1	Bayes Net	Decision Table	J48	Logistic	Random Forest	SMO	多数決方式
⑩学習用データ：7,653 ペア								
正解率	0.862*	0.822	0.834	0.845*	0.834	0.856*	0.818	0.870*
⑪学習用データ：4,443 ペア								
正解率	0.861*	0.824	0.807	0.842*	0.810	0.870*	0.814	0.874*
⑫学習用データ：2,159 ペア								
正解率	0.855*	0.822	0.762	0.846*	0.790	0.866*	0.821	0.859*
⑬学習用データ：1,329 ペア								
正解率	0.840*	0.801	0.801	0.771	0.734	0.843*	0.814	0.856*

*： 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

であり，属性集合 3（⑦と③）では学習データ方式 2（⑦）が性能上昇を見せたケースが多数を占めた。属性集合 1・3 の両者において学習データ方式 2 が性能上昇をもたらした機械学習法には BayesNet，Logistic，SMO があり，両者とも性能低下を見せたのは J48 であった。次に F 値で見ると，属性集合 1 では学習データ方式 2 の性能値が低下したケースが増え，属性集合 3 では性能上昇が増えたといえよう。属性集合 1・3 の両者において学習データ方式 2 が性能上昇をもたらした機械学習法には BayesNet と SMO があり，逆に両者とも低下を見せたのは DecisionTable と J48 であった。

以上により，主標目のみからなる件名とそれらに付与された分類記号を示した大量のデータは，細目付き件名の分類記号を同定する際には有効となるケースも多いが，大きく貢献を示すとはいいがたい。

(3)実験 3：学習データ方式 3

学習用データの件数が機械学習の性能に及ぼす影響を検証する目的で，学習用データの件数を順次削減して実験を行った。学習データ方式 1 において 3 交差検証法用に作成した 3 つの学習用データ集合のそれぞれに対して，主標目単位で系統抽出により半減させ，学習用データ⑩（7,653 ペア）を設けた。次に，さらに学習用データを主標目単位で半減させ，新たな学習用データ⑪（4,443 ペア），同様に半減を繰り返し，学習用データ⑫（2,159 ペア），そして⑬（1,329 ペア）を準備した。

これらを用いた実験の結果，徐々に性能値の低下を示したが，学習用データの減少幅に比べるとその低下は緩慢である（表 4）。性能値の大きな低下を見せたのは，Logistic と J48 であった。学習用データの削減が直接，性能値の低下を招かないという結果は多少とも予想外であった。しかし，対象としたデータ自体が主標目の代表分類記号との一致区分という単一属性による予測によって相当程度の性能値が得られるという特徴を有していたことを考えれば，学習用データが限られた件数であったとしても，このような性能値が得られたことは了解されよう。

注 1）谷口祥一・木村麻衣子「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」『第 64 回日本図書館情報学会研究大会発表論文集』2016, p.3-6.