書誌レコードに対する著作同定に機械学習を適用する試み：日本古典著作の事例

谷 口 祥 一（慶應義塾大学文学部）
taniguchi@z2.keio.jp

［抄録］書誌レコードに対する効率的かつ網羅的な著作同定を意図して、日本古典著作を事例に機械学習の適用を試みた。実験 1 は、人手により判定された書誌レコード群からタイトルと読み、責任表示と著者標目などの項目から値を抽出し特徴量とし、個別の著作を予測させる多クラス分類問題とした（データ数 22 万件）。複数の機械学習モデルと特徴量選択方式を適用した結果、得られた最高性能値は、ロジスティック回帰によるマクロ平均 F 値 0.732 であった。実験 2 は、2 つのレコードが同一著作を表すかを予測する 2 クラス分類問題とした（データ数 376 万件）。「同一著作」クラスについて、ロジスティック回帰により F 値 0.554 が得られた。

１．はじめに

OPAC の機械的な FRBR 化の実現やカタロガーによる著作典拠コントロールの支援などに向けて、書誌レコードに対する効率的かつ網羅的な著作同定は大きな課題である。これまではルールベース方式等による機械的な著作同定が殆どであり、筆者も国内の書誌レコードを用いてこうした実験を行った [1),2)]。そこでは、個々の書誌レコードから抽出した著作同定用のキー（「著者名＋タイトル」など）が一致したものを同一著作と見なすという、比較的単純な方式が用いられている。

このような著作同定の課題に対して、近年着目されている機械学習を試行することは、そのおおよその性能値の確認、さらには現時点での限界の確認という点で意義がある。機械学習は、重複する書誌レコードの同定、あるいは各種の名寄せという問題に関して、CiNii をはじめとする実運用の場面において既に採用されている。他方、書誌レコードが表す著作の同定という問題への適用は、これまで試みられていない。

本研究では、著作同定のうちでも難度が高いとされる日本古典著作の同定問題に対して、機械学習の適用を試みる。無著者名著作が多い、著作（原典）と多様な派生や関連著作（注解書など）があることなどが、日本古典著作の同定を難しくしている。

なお、教師あり機械学習を適用するには正解データが必要となるが、本研究では FRBR 研究会が人手により判定した書誌レコード群（NDL 作成の明治期～2009.3 収録分レコード；JAPAN/MARC 2006 フォーマット）を使用する [3),4)]。著作の判定基準を変えれば、必然的に判定結果も変更されるため、今回は同研究会が採用した判定基準とその判定結果に依拠することになる。

本研究は、Python の機械学習ライブラリ scikit-learn を用いた実験とし、設定課題を大きく変えた 2 つの実験として実行した。

２．実験１：個別著作の判定実験

２．１　対象データと前処理

実験 1 は、個別の書誌レコードが、事前設定されているいずれの著作に該当するかを機械学習により予測させる多クラス分類問題とした。いずれの著作にも該当しない場合は、「非該当」クラスに属させる。

FRBR 研究会が人手により判定した書誌レコード群のうち、当該著作に属する書誌レコードが 10 件以上ある著作 89（著作データ数 4,714）と、それぞれの著作において非該当と判定されたレコードを採用した。該当レコード数 10 件以上のものに限定したのは、交差検証を適正に行うためであり、10 件未満の場合にはそのすべてのレコードを非該当データとして使用した。その結果、非該当データは、17,345 となった。なお、単一のレコードが複数の著作に該当すると判定されている場合も複数ある。また、内容細目の記載事項が著作に該当すると判定されているものは、その扱いの難度が高いため、今回はすべて「非該当」として扱った。

上記のデータに NDC 9 類のレコード 20 万件を、上記のいずれにも含まれないものからランダムに抽出し、「非該当」クラスのデータとして追加した。この結果、実験 1 では、「非該当」クラスを含めて分類対象クラス数 90、総データ数 222,059 を実験対象データとした。なお、著作ごとに該当するデータ数には幅があり、非該当クラスを除けば、最大は源氏物語 984 件、次点は平家物語 282 件である。

個々の書誌レコードから以下のデータ項目値を抽出し、記号の除去などの表記の正規化を加え、機械学習実験用の特徴量（属性値）とした。漢字の異体字の統合などについては未処理である。特徴量の抽出方式については複数設定

し、著作同定性能への影響を検証した。また、いずれの方式においても採用する値は、データセット全体で 2 回以上出現するものに限定した。

a)著者標目:フィールド 751 と 79X(791~799;各巻レベル) のそれぞれにおいて、選択肢①最初に出現した個人または団体の標目のみを採用するか、あるいは②出現したすべての個人・団体を採用する。いずれの場合も、サブフィールド$B「漢字形」(ただし、生没年など付記事項を除去) および$3「典拠番号」を採用した。

b)責任表示:フィールド 25X と 29X (各巻レベル) のサブフィールド$F「責任表示」について、選択肢①最初に出現したもののみ (同一役割の 1 つまたは複数の個人・団体) を採用するか、あるいは②出現したすべてのものを採用する。いずれの場合も、役割表示は除去している。

採用する値は著者標目と責任表示では区別せず、いずれも「au_○○○」とした。

c)タイトル:フィールド 25X と 29X のサブフィールド$A「本タイトル」と$B「タイトル関連情報」を採用した。選択肢①は、$A と$B のそれぞれを独立したタイトルとして抽出したものに加えて、「$A+$B」、「$B+$A」という結合形を生成し、それぞれをタイトルとして採用した。他の選択肢として、②$A と$B の文字列に対して形態素解析を実行し、形態素単位のユニグラム、バイグラムをそれぞれタイトルとする方式とした。

d)タイトル標目:フィールド 55X と 59X (各巻レベル) のサブフィールド$A「カタカナ形」に対して、選択肢①そのままタイトルとして採用する、あるいは②分かち書きの単位でそれぞれの文字列をタイトルとして採用する。

採用する値は、記述のタイトルとタイトル標目とは区別せず、いずれも「ti_○○○」とした。

e)分類標目:フィールド 677 から NDC の版次 ($V) の区別を付けて、$A「分類記号」を採用した (「ndc6_○○○」、「ndc8_○○○」、「ndc9_○○○」)。フィールド 685 から NDLC 分類記号 ($A) を採用した (「ndlc_○○○」)。

f)件名標目:フィールド 650(個人名件名標目) と 658 (一般件名標目) から、サブフィールド $B「漢字形」と$3「典拠番号」を抽出した。$B において細目が付いている場合には、選択肢①「主標目＋細目」という全体を採用する、あるいは②主標目と細目はそれぞれ独立した値とし、さらに複数の細目が結合しているときには、それぞれの細目を独立した値とする (「ndlsh_○○○」)。

g)請求記号:フィールド 905 のサブフィールド $A「請求記号」から分類記号に該当する部分のみ採用した (「clno_○○○」)。

２．２　実験結果

機械学習モデルとして比較的計算量が少なく、かつ性能が安定しているといわれるロジスティック回帰、線形 SVM、ランダムフォレストを採用した。

いずれの機械学習モデルにおいても、訓練データは全体の 70%のデータを使用し、残り 30%のデータ (66,618 件) を評価に使用した。訓練データには、層化 k 分割交差検証法を、k=5 として実行した。また、訓練データによる学習においては各クラスサイズが不均衡であるため、バランスが取れるようにパラメータを設定した (class_weight="balanced")。さらに、グリッドサーチにより、F 値のマクロ平均を基準にして機械学習モデルごとに最適なハイパーパラメータを探索した。「非該当」クラスが極めて大きいため、マイクロ平均は適正ではなく、そのためマクロ平均を採用している。

主要な実験結果を、表 1 に示した。3 つの機械学習モデルごとの、全著作 (「非該当」を含む) に対するマクロ平均の性能値と、4 つの個別著作に対する性能値を示してある。

特徴量選択方式 A:選択肢を設けた特徴量の採用において、著者標目の選択肢①、責任表示の選択肢①、タイトルの選択肢①、タイトル標目の選択肢①、件名標目の選択肢①を採用した。採用した特徴量 (2 回以上出現) の異なり数は 127,858 であった。

3 つの機械学習モデルは、大まかには同程度の性能を示す結果となった。F 値については、ロジスティック回帰 (ハイパーパラメータ C=2000) による 0.681 が最高値であった。

なお、著者標目と責任表示の選択肢②は、いずれの場合にも、選択肢①に比べて、低い性能値であった。方式 B および C においてもこの点は同様の結果であった。

方式 B:著者標目の選択肢①、責任表示の選択肢①、タイトルの選択肢② (ただし、ユニグラムのみ)、タイトル標目の選択肢②、件名標目の選択肢①と②を採用した。採用した特徴量の異なり数は 124,304 であった。

ロジスティック回帰と線形 SVM の F 値は、方式 A に比べて上昇したが、ランダムフォレストは僅かながら低下を見せた。

方式 C:著者標目の選択肢①、責任表示の選択肢①、タイトルの選択肢①と② (ユニグラムと

バイグラムの両者を採用)、タイトル標目の選択肢①と②、件名標目の選択肢①と②という、網羅的な採用方式とした。採用した特徴量の異なり数は 294,555 であった。

3 つの機械学習モデルとも、方式 A および B に比べて性能を上昇させた。たとえば、ロジスティック回帰（ハイパーパラメータ C=100）は、F 値 0.732 を示した。

３．実験２：書誌レコードペアの著作同一性の判定実験

３．１　対象データと前処理

実験 1 の機械学習適用方式は、正解データが存在する事前設定の著作についてのみ予測が行われる。つまり、事前に設定されていない著作については、「非該当」との予測が行えるのみであり、新たな（未知の）個別著作を同定する機能はない。

そこで、実験 2 では 2 つの書誌レコードの組み合わせ（ペア）が同一著作を表すのか、それとも異なる著作を表すのかを機械学習により予測する 2 クラス分類問題とした。同一著作と予測されたレコードペアがいずれの著作に属するかはそのままでは不明であるが、それらをレコード番号によりグループ化することによって、該当する著作が判明する。

書誌レコード N 個において、その組み合わせは N(N-1)/2 という膨大なペアとなるため、今回の実験では、人手により判定された源氏物語、徒然草、伊勢物語、宇治拾遺物語という 4 著作のデータセットのみを選択し、実験データとした。具体的には、各著作の判定済みレコード群において、当該著作に該当すると判定されたレコード同士のペアは「同一著作」を表し、該当するレコードと該当しない（非該当）レコードのペアは「異なる著作」とした。なお、該当しないレコード同士は、当該著作とは別の著作として同一の著作を表す可能性があるため、レコードペアを生成していない。また、著作ごとの判定済みデータセットをまたがってペアを生成することもしていない。この結果、上記の 4 著作の判定済みのデータセット（著作に該当するレコード 1,285）から、同一著作を表すレコードペア 501,346、これに異なる著作を表すレコードペアを含めて総数 376 万強の実験データとした。なお、単一のレコードが 2 つの著作に該当すると判定されているもの 1 件、ある著作で該当し他の著作で非該当とされたもの 1 件、そして複数の著作において非該当とされたものが多数あった。実験 1 と同様、内容細目が著作に該当すると判定されているものは、今回の実験では「非該当」として扱った。

これらレコードペアにおいて、タイトルなど、13 項目についてその値の一致（「1」）・不一致（「0」）を表したものを特徴量として採用した。それら 13 項目とは、タイトル、タイトル標目、責任表示、著者標目、責任表示と著者標目のクロス照合、NDC6, 8, 9 版の各分類記号、NDC 版次ごとの分類記号のクロス照合、NDLC 分類記号、NDLSH 件名標目、その主標目のみ、そして請求記号のうち分類記号部分である。タイトルと責任表示については記号等の除去、役割表示の除去などを加えた上で照合している。

責任表示と著者標目は、選択肢①最初に出現したもののみ採用する、あるいは②出現したすべてのものを採用するを設けている。

なお、同一著作のレコードペアの場合であっても、すべての特徴量の値が不一致となる事例もあった。

３．２　実験結果

機械学習モデルとして、ロジスティック回帰とランダムフォレストを採用した。いずれの機械学習モデルにおいても、訓練データは全体の 70%のデータを使用し、残り 30%のデータ（1,128,497 件）を評価に使用した。訓練データには、層化 k 分割交差検証法を、k=5 として実行した。また、訓練データによる学習においては各クラスサイズが不均衡であるため、バランスが取れるようにパラメータを設定した（class_weight="balanced"）。さらに、グリッドサーチにより、精度のマクロ平均を基準にして機械学習モデルごとに最適なハイパーパラメータを探索した。ただし、F 値のマクロ平均を基準に採用しても、選択されるハイパーパラメータの値は同じであった。

実験の結果、「同一著作」クラスについては、ロジスティック回帰（ハイパーパラメータ C=0.1）が F 値 0.554、ランダムフォレスト（n_estimators=10）が F 値 0.550 であった（表 2）。なお、責任表示と著者標目は選択肢①を採用している。

得られた性能値はあくまでもレコードペアが同一著作か否かの予測に対する性能を示している。そこで、個々の評価用データに対する機械学習による予測結果を付けた形式で出力し、レコードペアを形成しているレコード ID からレコード単位でのグループ化と集計を別途行い、その性能値を 4 つの著作についてのマイクロ平均で求めた。いずれも精度が低いのは、

複数の著作のレコードが混合して大きなレコードグループを形成する結果となったからである。併せて、4 著作それぞれについて最大の該当レコード数かつ F 値を得たグループにおける性能値を示した。

引用文献

1) 谷口祥一. FRBR OPAC 構築に向けた著作の機械的同定法の検証：JAPAN/MARC 書誌レコードによる実験. *Library and Information Science*. No.61, 2009, p.119-151.

2) 谷口祥一. 総合目録データに対する機械的書誌同定と著作同定の試み：ゆにかねっとレコードによる実験. 日本図書館情報学会誌. Vol. 57, No. 4, 2011, p.124-140.

3) FRBR 研究会. 著作ページ. http://inforg.slis.tsukuba.ac.jp/jworkpage/

4) Takuya Tokita, et al. Identifying Works of Japanese Classics for Construction of FRBRized OPACs. *Cataloging & Classification Quarterly*. Vol.50, No.5-7, 2012, p.670-687.

表 1 実験 1 による機械学習の性能評価結果

	ロジスティック回帰			線形 SVM			ランダムフォレスト		
	精度	再現率	F 値	精度	再現率	F 値	精度	再現率	F 値
方式 A：ハイパーパラメータ	C=2000			C=20			n_estimators=500		
マクロ平均の性能値	0.671	0.719	0.681	0.626	0.685	0.632	0.680	0.607	0.619
宇治拾遺物語	0.750	0.500	0.600	0.333	0.333	0.333	0.600	0.250	0.353
伊勢物語	0.441	0.556	0.492	0.419	0.481	0.448	0.545	0.444	0.490
吉田兼好-徒然草	0.486	0.686	0.569	0.194	0.667	0.301	0.508	0.588	0.545
紫式部-源氏物語	0.926	0.945	0.935	0.766	0.948	0.847	0.878	0.941	0.908
方式 B：ハイパーパラメータ	C=500			C=20			n_estimators=100		
マクロ平均の性能値	0.682	0.777	0.716	0.653	0.718	0.673	0.681	0.601	0.615
宇治拾遺物語	0.667	0.500	0.571	0.600	0.500	0.545	0.714	0.417	0.526
伊勢物語	0.459	0.630	0.531	0.469	0.556	0.508	0.522	0.444	0.480
吉田兼好-徒然草	0.735	0.706	0.720	0.744	0.627	0.681	0.778	0.549	0.644
紫式部-源氏物語	0.945	0.952	0.948	0.936	0.952	0.944	0.975	0.941	0.958
方式 C：ハイパーパラメータ	C=100			C=20			n_estimators=500		
マクロ平均の性能値	0.703	0.790	0.732	0.696	0.728	0.696	0.714	0.638	0.649
宇治拾遺物語	0.750	0.500	0.600	0.667	0.500	0.571	0.714	0.417	0.526
伊勢物語	0.486	0.630	0.548	0.450	0.667	0.537	0.565	0.481	0.520
吉田兼好-徒然草	0.809	0.745	0.776	0.854	0.686	0.761	0.829	0.569	0.674
紫式部-源氏物語	0.959	0.966	0.962	0.961	0.941	0.951	0.979	0.945	0.961

表 2 実験 2 による機械学習の性能評価結果

	ロジスティック回帰			ランダムフォレスト		
	精度	再現率	F 値	精度	再現率	F 値
ハイパーパラメータ	C=0.1			n_estimators=10		
レコードペア単位の予測						
クラス「同一著作」	0.483	0.650	0.554	0.473	0.657	0.550
2 クラスのマクロ平均	0.713	0.771	0.736	0.709	0.772	0.732
レコード単位の集計：						
4 著作のマイクロ平均	0.479	0.834	0.609	0.368	0.870	0.517
宇治拾遺物語	0.661	0.837	0.739	0.580	0.959	0.723
伊勢物語	0.037	0.724	0.071	0.028	0.827	0.055
吉田兼好-徒然草	0.069	0.851	0.128	0.049	0.916	0.094
紫式部-源氏物語	0.493	0.951	0.650	0.339	0.983	0.504