行動計量学 第 41 巻第 1 号（通巻 80 号）
2014 年，35〜46

原　著

統合的分類アルゴリズムを用いた文章の書き手の識別

金　明哲

Using Integrated Classification Algorithm to Identify a Text's Author

Mingzhe Jin

Text classification results often vary depending on the detailed factors in data analysis, including feature data, classification method, and parameter sets adopted in the analysis. The author of an anonymous text can be generally identified by extracting a set of distinctive features of the text, and then using the features to find the most likely author. Numerous efforts have been made to develop the feature extraction technique with more robustness and the classification algorithm, but an important issue is how to select the features datasets and classification method. To address this issue, we propose an integrated classification algorithm that extracts multiple feature datasets from differing viewpoints and aspects of a text and applies multiple strong classifiers to the datasets. Our proposed method achieved 100% accuracy in identifying the authors of literary works and student essays, and identified the author of all but 1 out of 60 diaries which were written by 6 different people.Our proposed method achieved equivalent or better accuracy than the case when any a strong classifier applied to individual feature dataset. Furthermore, the accuracy in identifying the authors of student essays increased by roughly two percentage points.

Key words: integrated classification algorithm, strong classifier, feature datasets, identify the author
キーワード：統合的分類アルゴリズム，強分類器，特徴データセット，書き手の識別

1. ま え が き

データ解析やデータマイニングの結果はデータの質と用いる方法に大きく依存する．匿名の文章の書き手を識別する分野も例外ではない．

文章における書き手の特徴データの抽出方法は多く提案されている．Grieve (2007) は約 40 種類の書き手の特徴データについて，書き手判別の精度の比較分析を行った．その結果，単語および記号と文字列の bigram を組み合わせることによりもっとも高い判別精度が得られたと報告している．

また，判別方法（分類器，classifier）も数多提案されている．Sebastiani (2002) はテキスト分類に主に用いられている分類器精度について，40 数件の研究論文に用いられた分類器の結果について比較を行った．その結果，ニューラルネット法，アダブースト（AdaBoost），SVM 法の正解率が高いことを明らかにした．Sebastiani の研究報告から現在に至るまで，いくつかの新たの分類器が提案されている．例えば，ランダム・フォレスト (Breiman 2001)，ロジット・モデル・ツリー (LMT: Logit Model Tree, Landwehr et al 2005)，高次元判別分析 (HDDA: High-Dimensional Discriminant Analysis, Bouveyron et al 2007)，距離加重判別分析 (DWD: Distance-Weighted Discrimination, Marron et al 2007, Huang et al 2012) などがある．

金・村上 (2007) は，ランダム・フォレスト，SVM，アダブースト，バギング，ニューラルネットワーク，K-NN 法などを用いて書き手の判別を行い，ランダム・フォレスト，SVM，アダブースト，バギングの精度が優れていることを報告した．

このような研究は，よりよい書き手の特徴データの

同志社大学文化情報学部・文化情報学研究科
（Doshisha University）
連絡先：〒 610–0394　京田辺市多々羅都谷 1–3 夢告館 612 号室
E-mail：mjin@mail.doshisha.ac.jp

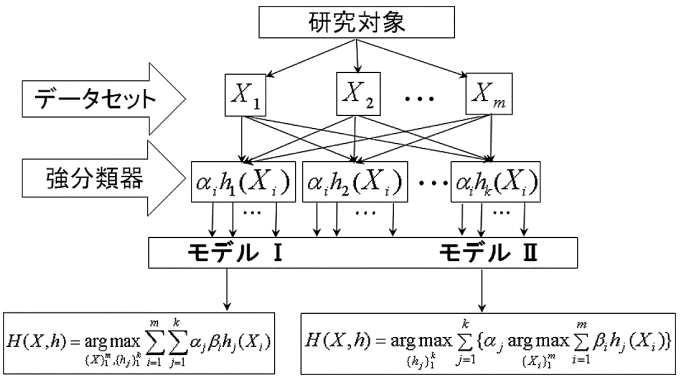

図 1. 統合的判別分析法のイメージ

抽出方法や指標を求めることとデータにマッチする精度が高い判別方法の開発や探索に集中している．しかし，分類器はデータとの適応性があり，どのデータでも常に精度がもっともよいことは保証できない．また，一般的に評価が高い分類器で誤判別されたものが，評価が相対的に低い分類器では正しく判別されるケースも珍しくない．また，研究対象から抽出したデータ A では誤判別されるが，データ B では正しく判別される場合もある．近年情報機器の普及と性能の向上に伴い，研究対象から異なる視点と側面によるデータ抽出が容易になった．

このようなことにより，本研究では，研究対象の異なる側面から抽出した複数のデータセットを，複数の分類器で判別した結果を統合的に用いる判別方法を提案し，匿名の文章の書き手を判別する実証を行った．

2. 提案の方法と用いるコーパス

2.1. 提案の方法

我々の人間社会には多くの委員会があり，重大な意思決定は，しばしば委員会の委員による多数決の方法が用いられている．このような考えを導入した分類器としては，アンサンブル学習（集団学習）がある．その代表的なアルゴリズムは，アダブースト，バギング，ランダム・フォレストである．これらの方法は，一つのデータセットから複数の弱分類器（weakclassifier）を作成し，その結果の多数決による新しい強分類器（strong classifier）を作成する．弱分類器とは決して分類精度が高くない分類器を指し，決定木が多く用いられている．上記のアンサンブル学習法による強分類器は多くの実証が行われ，定評を得ている．

本研究では，このようなアイディアを借り，研究対象の異なる側面から抽出した複数のデータセットと定評がある精度が高い複数の強分類器を用いた統合的分類方法について 2 つのモデルを提案する．そのアルゴリズムを次に示し，そのイメージを図 1 に示す．

アルゴリズム

A) 研究対象の異なる側面から m セットのデータ $U=\{X_i|i=1,2,\cdots,m\}$ を作成する．

B) データセット X_i についてテストを重ね，k 個の強分類器 $h_j(X),(j=1,2,\cdots,k)$ を選出する．

モデル **I**：

$m\times k$ の結果について重みつき多数決をとる．

$$H(X,h)=\underset{\{X_i\}_1^m,\{h_j\}_1^k}{\arg\max}\sum_{i=1}^{m}\sum_{j=1}^{k}\alpha_j\beta_i h_j(X_i)$$

係数 α_j は分類器の重みであり，β_i はデータセットの重みである．ただし，本研究ではすべて 1 にした．

モデル **II**：

(1) m 個のデータセットごとに k 個の分類器の結果について重みつき多数決をとる．

(2) 得られた m 組の結果について，さらに重みつき多数決をとる．

$$H(X,h)=\underset{\{h_j\}_1^k}{\arg\max}\sum_{j=1}^{k}\{\alpha_j \underset{\{X_i\}_1^m}{\arg\max}\sum_{i=1}^{m}\beta_i h_j(X_i)\}$$

表 1. 用いた文豪の文学作品リスト

著者	作品名
芥川龍之介	或阿呆の一生，羅生門，芋粥，枯野抄，地獄変，杜子春，蜘蛛の糸，将軍，春，点鬼簿 （最小値：2726，平均値：8395，最大値：17290）
泉鏡花	化鳥，女客，婦系図 (a)，小春の狐，怨霊借用，木の子説法，絵本の春，縁結び，草迷宮，遺稿 （最小値：2217，平均：15580，最大値：55900）
菊池寛	仇討禁止令，芥川の事ども，勲章を貰う話，三浦右衛門の最後，無名作家の日記，大島が出来る話，恩讐の彼方に，俊寛，勝負事，出世（最小値：3531，平均値：11230，最大値：19090）
森鴎外	かのように，二人の友，余興，堺事件，妄想，寒山拾得，山椒大夫，普請中，最後の一句，百物語 （最小値：2913，平均値：13160，最大値：15180）
夏目漱石	三四郎 (a)，吾輩は猫である (a)，坊っちゃん，幻影の盾，彼岸過迄 (a)，琴のそら音，硝子戸の中，草枕，薤露行，趣味の遺伝（最小値：11580，平均：44220，最大値：92890）
佐々木味津三	なぞの八卦見，千柿の鍔，南蛮幽霊，曲芸三人娘，生首の進物，笛の秘密，耳のない浪人，袈裟切り太夫，身代わり花嫁，青眉の女（最小値：7206，平均：12020，最大値：14750）
島崎藤村	三人，並木，伸び支度，分配，刺繍，岩石の間，桃の雫，海へ (a)，熱海土産，藁草履 （最小値：5220，平均値：25190，最大値：83610）
太宰治	二十世紀旗手，作家の手帖，俗天使，八十八夜諦めよ，散華，断崖の錯覚，春の盗賊，服装に就いて，未帰還の友に，花吹雪（最小値：5744，平均：11260，最大値：23080）
岡本綺堂	ゆず湯，半七捕物帳–石燈籠，寄席と芝居と，影を踏まれた女，心中浪華の春雨，異妖編，穴，箕輪心中，青蛙堂鬼談，鳥辺山心中（最小値：7750，平均：23440，最大値：94680）
海野十三	奇賊悲願，宇宙戦隊, 怪星ガン，恐しき通夜，海底都市，生きている腸，骸骨館，鬼仏洞事件，宇宙の迷子，暗号音盤事件(最小値：4608，平均：22810，最大値：65210）

2.2. 用いるコーパスとデータセット

提案する統合的判別分析法の実証のため，本研究では文豪の作品，学生の作文，一般人の日記，計 3 種類のコーパスを用いた．

（1） 文豪の作品

文豪の作品は青空文庫からダウンロードして用いた．学生の作文や日記とバランスをとるため，10 人の作家による計 100 編（10×10）の文学作品を用いた．そのリストを表 1 に示す．作品の選定には，なるべく同年代であること，新仮名を用いていることなどに配慮した．また長い作品は青空文庫が分割したサイズをそのまま独立した 1 編として扱った．表の中の作品の右にアルファベット a が付いているのは，その作品の一部分である．用いた作品の中，もっとも短いのは約 2200 全角文字，もっとも長いのは 94680 全角文字である．著者別の作品の最小値，平均値，最大値を表 1 に示す．

（2） 学生の作文

学生の作文は，文章の計量分析のため 11 人の大学 3 年生に 10 のテーマ（T1; 住まい，T2; 家族，T3; 友達，T4; 学校，T5; スポーツ，T6; 旅行，T7; 車，T8; アルバイト，T9; 映画は映画館で見るかビデオで見るか，T10; 日本食）について書かせた作文コーパスである（金・宮本 1997, 金 2013）．11 人が書いた作文のサイズをタイトル別に表 2 に示す．表 2 でわかるように，作文のサイズは平均 1124 文字である．もっとも短いのは 978 文字，もっとも長いのは 1761 文字である．

（3） 一般人の日記

用いた日記は，ワープロが普及始まった 1990 年代初期に，NHK が手書きとワープロで書いた日記の文体に違いがみられるかについて比較分析を行うために行った実験データである．用いたのは 6 人が 10 日で書いた日記である．前の 5 日間は手書きで，後の 5 日間はワープロによるものである（金・樺島・村上 1993b, 金 2013）．用いた日記のサイズを表 3 に示す．日記の平均文字数は 529 文字，もっとも短い日記は 268 文字，もっとも長いのは 1244 文字である．

表 2. 分析に用いた 11 人の作文のサイズ（単位は文字数）

書き手	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	平均
WA	1065	1168	1582	1053	1208	1049	1065	1299	1089	1006	1159
WB	1097	1157	978	1270	1374	1295	1167	1126	1235	1054	1175
WC	1068	1761	1155	1414	1114	1017	1242	1292	1229	1102	1240
WD	1102	1035	1129	1032	1007	1089	1046	1054	1051	993	1054
WE	1032	1063	1266	1173	1018	1178	1081	1061	1101	1126	1110
WF	1066	1105	1069	1075	1039	1077	1100	1125	1045	1164	1087
WG	1060	1261	1438	1300	1170	1068	1184	1471	1032	1170	1216
WH	998	1045	1187	1133	1168	1030	1230	993	1238	1194	1122
WI	1046	1060	1047	1113	1109	1111	1044	1042	1101	1090	1076
WJ	1077	1026	1045	1044	1063	1025	1060	1081	1033	1089	1055
WK	1392	1042	1013	1006	1009	1015	1052	1029	1135	1012	1071
平均	1091	1157	1174	1147	1116	1087	1115	1143	1117	1091	1124

表 3. 6 人の書き手の日記のサイズ（単位は文字数）

書き手	0	1	2	3	4	5	6	7	8	9	平均
A	290	268	405	355	325	397	499	577	428	452	400
B	399	412	418	411	402	403	400	608	439	445	434
C	502	549	610	636	581	696	673	697	754	516	622
D	585	567	475	686	546	491	561	567	517	458	546
E	521	555	468	486	623	668	726	693	1244	778	676
F	437	495	503	484	502	470	510	562	492	524	498

2.3. 書き手の特徴データ

文章の書き手を判別する際に，文章からどのように書き手の特徴データを抽出して用いるかに関しては数多くの研究が行われてきた．その代表的なのは，単語の長さ，文の長さ，品詞の使用率，漢字の使用率，語彙の豊富さ指標，語彙や記号の使用率，文字の n-gram，助詞の組み合わせ，文節パターンなどがあげられる．単語の長さ，文の長さ，漢字の使用率，語彙の豊富量の指標は，人によっては書き手特徴となる場合はあるが，一般人が書いた現代文では書き手を判別する有力な情報になれない場合が多い（金 2012, Grieve 2007）．

本研究では記号，形態素，構文の側面から文字・記号の bigram，タグ付きの形態素，タグの bigram，文節パターン，計 4 種類のデータを抽出して用いた．また，用いた文章の長さは均一ではないため，文章から抽出したデータは相対頻度に変換して用いた．

（1） 文字・記号の n-gram

文字・記号の n-gram（以下略して文字の n-gram とよぶ）は，隣接している $n(n = 1, 2, 3, \cdots)$ 個の文字・記号列のパターンをもれなく集計したデータを指す．データ抽出の方法は非常に単純であるが，そのデータには書き手の癖が織り込まれている．文字の n-gram による書き手の判別に関しては，その有効性が示されている（松浦・金田 2000, Grieve 2007, Jin and Huh 2012）．ただし，n を 3 以上にした場合データの次元が高く，データには書き手特徴のみではなく，文章の内容やジャンルに依存する要素がノイズとしてより強く反応することに注意が必要である．読点をどの文字の後に打つかに関するデータを用いて書き手を判別する方法が提案され，その有効性も報告されている (Jin & Murakami 1993, 金・樺島・村上 1993a, 金 1994, Jin and Jiang 2012)．読点をどの文字の後に打つかに関するデータは文字・記号の bigram の一部分にすぎない．ただし，このようなデータの次元は，文字の bigram よりはるかに低いため，教師なしのデータ解析には向いている (Jin and Jiang 2012)．教師ありのデータ解析法を前提とする場合は，文字の n-gram を用いた方がもっと有効である．

文字の n-gram を抽出する際に n をいくらにするかに関しては，文章の長さに依存するので一概には言えない．本研究では，今までの経験を踏まえて bigram を用いることにした．

（2） タグつきの形態素

文字の bigram では，文字と記号を単位としてデータを抽出する．よって，単語や品詞に関する情報は用いていない．単語や品詞に関する情報を用いるためには，形態素解析を行い，形態素の属性に関するタグを付与することが必要である．日本語の形態素解析のツールとしては JUMAN, ChaSen, MeCab などがある．本研究では MeCab により形態素解析を行い，属性タグを付与して用いた．例文「誰が行きますか？」を MeCab で形態素解析した結果を次に示す．

形態素	タグ
誰	名詞, 代名詞, 一般
が	助詞, 格助詞, 一般
行き	動詞, 自立, *
ます	助動詞,*,*
か	助詞, 副助詞／並立助詞／終助詞,*
？	記号, 一般,*

形態素の属性に関してはいくつかの層に分かれている．たとえば，形態素（記号，語）「誰」の属性は，「名詞」,「代名詞」,「一般」になっている．このような「名詞」を第 1 層，「代名詞」を第 2 層，「一般名詞」を第 3 層と呼ぶことにする．タグつき形態素データでは第 2 層までの情報を用いた．タグつきの形態素を用いた書き手の判別に関しては金・村上 (2007) がある．

（3） タグの n-gram

文の構造に関する情報は，形態素解析済みのデータから形態素タグの n-gram データを抽出することが考えられる．形態素タグのほとんどは品詞の属性である．品詞の使用率を用いた文章の統計分析に関する研究は 1950 年代から行われている．品詞の接続関係の情報を用いた早期の研究として Antosch (1969) がある．Antosch は，動詞–形容詞の比率について調査分析を行い，文章のジャンルによってその比率は異なり，民話では動詞–形容詞の比率が高く，科学関連の文章では低いという結論を得た．日本語においては，村上・伊藤 (1991) は日蓮遺文の計量分析に品詞の接続関係の情報などをも用いた．書き手の識別特徴となる品詞の n-gram に関する研究についてはいくつかの試みが行われ，品詞の n-gram は書き手の判別に有効であると報告されている（金 2003, 2004a, 2004b, Jin and Huh 2012）．

タグの n-gram を用いる際にも，n をいくつにするかが一つの問題である．通常は n=2 か 3 が適している（金 2004b）．本研究ではタグの第 1 層の bigram を用いた．

（4） 文節のパターン

文章における構文にも書き手の癖が多く見られる．日本語の構文分析の基本単位となる文節について，文節パターンをモデル化し，そのデータに基づいて書き手の判別を行う方法が提案され，実証が行われている．その結果，文節パターンにも書き手の特徴が比較的に顕著に現れ，匿名文章の書き手判別に有効であることがわかった（金 2013）．

文節の切り分けに関する構文解析ツールとしては JUMAN/KNP, CaBoCha がある．本研究では CaBoCha を用いた．文節に関しては次のようにパターン化した．まず, 例文「誰が行きますか?」を CaBoCha で構文解析した結果を次に示す．この例文は，2 つの文節によって構成されている．本研究では，第 1 文節は「名詞–が」，第 2 文節は「動詞–助動詞–か–記号」にパターン化した．つまり，文節内の助詞は原型を用い，それ以外は形態素のタグを用いた（金 2013）．

```
* 0 1D 0/1 0.00000000
誰      ダレ    誰      名詞–代名詞–一般
が      ガ      が      助詞–格助詞–一般
* 1 -10 3/3 0.00000000
行き    イキ    行く    動詞–自立五段・カ行促音便
ます    マス    ます    助動詞 特殊・マス基本形
か      カ      か      助詞–副助詞/並立助詞/終助詞
？                      記号–一般
EOS
```

2.4. 用いる分類器

分類器は先行研究の結果を踏まえ，最新の分類器を含む 6 種類の分類器の精度について考察を行った．6 種類の分類器は，アダブースト（ADA: AdaBoost），距離加重判別（DWD: Distance Weighted Discrimination），高次元判別分析（HDDA:High DimensionalDiscriminant Analysis），ロジスティク・モデル・ツリー（LMT: Logistic Model Trees），ランダム・フォレスト（RF: Random Forests），サポート・ベクター・マシン（SVM: Support Vector Machine）である．

（1） アダブースト

アダブースト（AdaBoost: Adaptive Boosting）は，Freund と Schapire (1996) によって提案された代表的なアンサンブル機械学習アルゴリズムである．アダブーストは，弱分類器の判別の誤差情報を用いて重みを調整しながら，繰り返し結果を生成し，その結果を組み合わせ，強分類器を構築する．

（2） 距離加重判別 DWD

距離加重判別分析（DWD, Distance Weighted Discrimination）法は，高次元小標本（HDLSS, High Dimension Low Sample Size）データのために提案された分類器である (Marron 2007)．SVM と同じく，クラスを分離するための境界マージンを最大化する方法を用いているが，DWD は平均距離を最大にするアプローチで分類境界の最大マージンを求める．また，DWD はサポート・ベクターのみに頼るのではなく，データの全てのベクターが分類を行う超平面の構築に用いられ，超平面に近いベクターにより強く，遠く離れているベクターに弱い重みを与える．Marron (2007) により提案されたのはバイナリ分類器であるが，多重クラス分類器として拡張されている (Huang 2011)．高次元では SVM より優れた結果が得られることがあると報告されている．

（3） 高次元判別分析 HDDA

HDDA は高次元判別分析（High-Dimensional Discriminant Analysis）方法として，Bouveyron らが提案した新しい分類器である (Bouveyron et al 2007)．HDDA は，データがガウス混合分布であるという仮定に基づいて，決定ルールを構築し分類を行う．HDDA は，モデルパラメータを推定する部分と決定ルールを構築する部分に分けられる．提案者らの比較研究では，用いたデータではサポート・ベクター・マシンより高い正解率を得たと報告している．

（4） ロジスティック・モデル・ツリー LMT

ロジスティック・モデル・ツリー（LMT: Logit model Tree）は，決定木とロジスティックモデルを組み合わせたアルゴリズムである．LMT は決定木の葉の部分のデータを用いてロジスティック判別のモデルを構築する．提案者が用いたベンチマークの分析によると，30 数種類の分類器の中でもっとも正解率が高いと報告されている (Landwehr et al. 2006)．

（5） ランダム・フォレスト

ランダム・フォレスト（RF; Random Forest）は，バギングの提案者 Breiman により今世紀はじめに提案されたアルゴリズムである (Breiman 2001)．RF はバギングと違い，個体だけではなく，変数についてもランダムサンプリングしたサブセットを用いるので，高次元データ解析やテキスト分類に向いている（金・村上 2007）．

（6） サポート・ベクター・マシン

サポート・ベクター・マシン（SVM：Support Vector Machine）は，Vapnik (1998) が提案したアルゴリズムである．分類の問題では，各分類境界のマージンを最大化する超平面を求め，分類を行う分類器である．近年非常に注目され，テキスト分類を含む多くの応用例が報告されている．

本研究では分類器のアルゴリズムは R の関連パッケージを用いた．アダブーストは boosting {adabag}，距離加重判別法は kdwd {DWD}，高次元データ判別法は hdda {HDclassif}，ロジット・モデル・ツリー法は LMT {RWeka}，ランダム・フォレストは randomForest {randomForest}，サポート・ベクター・マシンは ksvm {kernlab} を用いた．{ } で囲んだ文字列はパッケージの名称で，その前の文字列が分類器の関数である．結果の再現性などを考慮し，分類器関数の引数（パラメータ）はすべてデフォルト値を用いた．したがって，分類器によってはパラメータの調整により正解率をさらに高める可能性がある．

2.5. 評価方法

（1） 交差確認

分類器の性能の評価には，学習データで構築した分類器を，テストデータを用いて評価する．伝統的な方法としては LOOCV（*leave-one-out cross-validation*）法がある．データマイニングの分野では，標本が膨大であるため LOOCV を拡張した k-分割交差検証（k-fold cross-validation, $k = 3, 5, 10$) 法が多く用いられている．k-分割交差検証法は分割をランダムに行うため，小標本の場合は，グループの標本のバランスが取れないケースもある．そこで，本研究では各書き手から 1 作品を同時に取り出した LOOCV を拡張した交差確認法を用いる．用いた 3 種類のコーパスの標本サイズでいうと，一人の文章が 10 であるので，10-分割交差確認法に相当する．ただし，各書き手から 1 作品を抽出するので，ランダムに分割を行う一般の 10-分割交差確認法と異なり，グループに偏る分割が起こらないことを強調しておきたい．

表 4. クラス G_i の判別結果の混同行列

クラス G_i		分類法の結果	
		Yes	No
データ	Yes	a_i	c_i
	No	b_i	d_i

（2） 評価指標

著者 $i(i = 1, 2, \cdots g)$ とそれ以外の著者にラベルをつけたグループをクラス G_i とすると，G_i における判別の結果は表 4 に示すクロス表で表すことができる．

分類結果の評価には，再現率（recall）や精度（precision）が多く用いられている．それぞれの定義を次に示す．

$$\text{再現率：} R_i = \frac{a_i}{a_i + c_i}$$
$$\text{精　度：} P_i = \frac{a_i}{a_i + b_i}$$

再現率 R_i は，分類器がどれぐらい「漏れ」なく正しく判別しているかに関する度合であり，精度 P_i は分類器の分離結果に混入された「ゴミ」に対する的中率である．多群分類問題において評価指標としては，再現率と精度のマクロ平均（macro average）がある．その計算式を次に示す．

$$\text{再現率：} \hat{R} = \left(\frac{1}{g} \sum_{i=1}^{g} \frac{a_i}{a_i + c_i} \right) \times 100\%$$
$$\text{精　度：} \hat{P} = \left(\frac{1}{g} \sum_{i=1}^{g} \frac{a_i}{a_i + b_i} \right) \times 100\%$$

本研究では，分類器の評価は次に示す再現率と精度の調和平均 F1（F1-measure）を用いる．F1 値が大きいほど，分類性能がよい（正解率が高い）と評価する．

$$F_1 = \frac{2 \times \hat{P} \times \hat{R}}{\hat{P} + \hat{R}}$$

3. 実 験 結 果

3.1. 文学作品の書き手の判別

10 人の文豪が書いた 100 編の作品について，6 つの分類器を用いた判別結果を表 5 に示す．表の中の網掛けの部分は正解率が高い上位 4 位の分類器であり，太文字はデータセットの中で F1 値が最も高い値である．次に示す表 6，7 でも同様に示す．

文字列の bigram では HDDA と RF の正解率がもっとも高く，F1 値は 99.05 である．これは 100 編の作品の中，1 編の作品の書き手が誤判別された結果である．タグ付き形態素のデータを用いた RF の F1 値は 100 である．その次が DWD, SVM, HDDA であり，その F1 値は 98.9 である．タグの bigram の中，正解率がもっとも高いのは RF であり，その F1 値は 98.05 である．続いて正解率高いのは HDDA, DWD の順である．

文節パターンでは DWD, RF, SVM が同じ正解率でもっとも高い．6 種類の分類器の中，総合的に正解率が高いのは RF である．DWD は 3 種類のデータにおいては SVM 同等であるが，タグの bigram では SVM より高い正解率を得ている．HDDA は「文節パターン」を除くと SVM 同等かそれ以上の正解率を得ている．ADA, LMT は他の分類器より正解率が低い．

3.2. 学生作文の書き手の判別

11 人が書いた 110 編の作文から抽出した 4 種類のデータについて，6 つの分類器で書き手を判別した結果を表 6 に示す．文字の bigram では DWD の正解率がもっとも高く，その F1 値は 98.41 である．続いて高いのは HDDA, SVM である．タグ付き形態素でも DWD の正解率がもっとも高く，その次は SVM, RF の順である．タグの bigram では，わずかでありながら HDDA の正解率が DWD を上回り，もっとも高い．その次が RF である．文節パターンでは DWD,RF の正解率が共にもっとも高く，その次が HDDA である．学生作文では DWD が総合的に高い正解率を得ている．ADA, LMT はその他の分類器には及ばない．

3.3. 日記の書き手の判別

6 人が書いた 60 編の日記から抽出した 4 種類のデータセットについて 6 つの分類器で書き手を判別した結果を表 7 に示す．文字の bigram では，DWD の正解率がもっとも高く，F1 値は 98.41 である．その次が HDDA, SVM である．タグ付き形態素では RF の正解率がもっと高く，F1 値は 96.67 である．その次が SVM, LMT である．ここでは LMT がわずかでありながら DWD, HDDA を上回っている．タグの bigram では，RF の正解率がもっとも高く，次に DWD, LMT の順になっている．文節パターンでの正解率の高いのは HDDA, RF, DWD の順である．日記データでは LMT が上位 3 位に 2 回ランクインされているが，4 位との差はわずかである．

表 5. 文学作品における 6 つの分類器別の書き手判別結果

データ種類とサイズ	評価指標	ADA	DWD	HDDA	LMT	RF	SVM
文字の bigram 100 × 1570	Recall	94.00	98.00	99.00	94.00	99.00	98.00
	Precision	94.49	98.18	99.09	94.52	99.09	98.18
	F1-measure	94.25	98.09	**99.05**	94.26	**99.05**	98.09
タグ付き形態素 100 × 1416	Recall	96.00	98.00	98.00	92.00	100	98.00
	Precision	96.52	98.18	98.18	92.66	100	98.18
	F1-measure	96.26	98.09	98.09	92.33	**100**	98.09
タグの bigram 100 × 423	Recall	84.00	95.00	95.00	93.00	98.00	88.00
	Precision	84.74	95.36	95.45	93.66	98.09	89.82
	F1-measure	84.37	95.18	95.23	93.33	**98.05**	88.90
文節パターン 100 × 1370	Recall	89.00	99.00	83.00	93.00	99.00	99.00
	Precision	89.32	99.09	89.89	93.49	99.09	99.09
	F1-measure	89.16	**99.05**	86.31	93.25	**99.05**	**99.05**

表 6. 学生作文における分類器別の書き手判別結果

データ種類とサイズ	評価指標	ADA	DWD	HDDA	LMT	RF	SVM
文字の bigram 110 × 1856	Recall	84.55	98.18	97.27	80.91	95.45	96.36
	Precision	85.07	98.35	97.52	82.44	95.70	96.83
	F1-measure	84.81	**98.26**	97.40	81.67	95.58	96.60
タグ付き形態素 110 × 1376	Recall	90.00	98.18	92.73	90.00	95.45	96.36
	Precision	90.39	98.35	93.50	91.00	96.01	96.61
	F1-measure	90.20	**98.26**	93.11	90.49	95.73	96.49
タグの bigram 110 × 545	Recall	63.64	90.00	90.91	80.91	87.27	80.00
	Precision	62.80	92.42	93.73	81.71	89.09	89.34
	F1-measure	63.22	91.19	**92.30**	81.31	88.17	84.41
文節パターン 110 × 623	Recall	86.36	96.36	95.45	90.00	96.36	95.45
	Precision	87.75	97.08	96.49	91.96	97.08	96.08
	F1-measure	87.05	**96.72**	95.97	90.97	**96.72**	95.77

3.4. 統合的判別分析

前節の結果で分かるように，すべてのデータセットで常にもっとも高い正解率得る分類器はなかった．また，正解率もデータセットによって異なる．そこで，本節では提案した複数のデータセットと複数の分類器を用いた統合的判別方法について実証を試みる．統合的判別分析に用いる分類器は，前節の分析結果を踏まえて総合的に正解率が高い DWD, HDDA, RF, SVM 計 4 つを用いる．

前の節で分かるように，データセットおよび分類器によって判別率には差がある．提案する統合的判別方法のアルゴリズムは，データセットと分類器に重みを付けることになっているが，問題をシンプルにするためここでは全て 1 にしている．

読者の理解を助けるため，日記から抽出した 4 種類のデータセットについて 2 種類の分類器（RF, HDDA）

表 7. 日記データにおける分類器別の書き手判別結果

データ種類とサイズ	評価指標	ADA	DWD	HDDA	LMT	RF	SVM
文字の bigram 60 × 1242	Recall	70.00	98.33	96.67	68.33	93.33	95.00
	Precision	74.79	98.48	96.97	68.85	93.30	95.71
	F1-measure	72.31	**98.41**	96.82	68.59	93.32	95.35
タグ付き形態素 60 × 473	Recall	83.33	90.00	90.00	90.00	96.67	95.00
	Precision	84.64	90.27	90.27	90.72	96.67	96.15
	F1-measure	83.98	90.13	90.13	90.36	**96.67**	95.57
タグの bigram 60 × 489	Recall	71.67	81.67	78.33	80.00	85.00	75.00
	Precision	74.39	83.36	80.88	80.37	86.93	78.57
	F1-measure	73.00	82.50	79.58	80.18	**85.95**	76.74
文節パターン 60 × 327	Recall	50.00	81.67	81.67	68.33	81.67	75.00
	Precision	60.11	81.45	83.08	69.83	83.02	77.81
	F1-measure	54.59	81.56	**82.37**	69.07	82.34	76.38

表 8. 統合的書き手識別の例（四種類の書き手の特徴データ，2 種類の分類器）

ID	文字 bigram		形態素		文節パターン		タグの bigram		投票結果	書き手
	RF	HDDA	RF	HDDA	RF	HDDA	RF	HDDA		
No.01	A	A	A	A	A	A	A	A	A	A
No.37	D	B	D	B	D	B	B	B	**B**	D
No.38	D	D	D	D	E	E	D	B	D	D
No.60	F	F	F	F	F	B	F	F	F	F

で判別した結果の一部分を例として表 8 に示す．表の中のアルファベットは書き手を示すラベルである．提案している統合的判別分析法は，多数決に基づいているので，各データについて複数の分類器が判別した結果の中からもっとも頻度が高いラベルが統合判別の結果となる．その結果を「投票結果」の縦列に示している．「投票結果」の右の列がその日記を書いた書き手である．No.37 は B が 5 つ，D が 3 つであるので B と判断される．しかし，この日記は D が書いたものであるので誤判別されている．

（1） モデル I

提案したモデル I の実証結果を表 9 に示す．ついでに，分類器の組み合わせの効果を考察するため，4 つの分類器の組み合わせの結果をも示す．用いたデータセットにおいて総合的に正解率が高い分類器の組み合わせは {DWD, RF}，{HDDA, RF}，{DWD, RF, SVM}，{HDDA, RF, SVM}，{DWD, HDDA, RF, SVM} である．これらの組み合わせでは，文学作品，学生作文の書き手判別の F1 値は 100 である．日記の書き手判別の F1 は 98.41 であり，1 編が誤判別されている．これらの分類器の組み合わせには全て RF を含んでいることから，分類器 RF の有効性（金・村上 2007）が再度実証された．

（2） モデル II

モデル II は，複数の分類器で各データセットについて判別した結果を統合し，その結果をさらに統合する．4 つの分類器によるデータセット別の統合結果およびそれをさらに統合した総合的統合結果を表 10 に示す．

表 9. 統合的判別の結果

分類器	文学作品			学生の作文			日記		
	Rec	Pre	F1	Rec	Pre	F1	Rec	Pre	F1
DWD	100	100	100	100	100	100	96.67	96.97	96.82
HDDA	99.00	99.09	99.05	99.09	99.17	99.13	96.67	96.97	96.82
RF	100	100	100	99.09	99.17	99.13	96.67	97.22	96.94
SVM	99.00	99.09	99.05	99.09	99.17	99.13	96.67	97.22	96.94
DWD, SVM	99.00	99.09	99.05	99.09	99.17	99.13	98.33	98.49	98.41
DWD, RF	100	100	100	100	100	100	98.33	98.49	98.41
DWD, HDDA	99.00	99.09	99.05	99.09	99.17	99.13	96.67	96.97	96.82
RF, SVM	100	100	100	100	100	100	96.67	97.22	96.94
HDDA, SVM	99.00	99.09	99.05	99.09	99.17	99.13	98.33	98.49	98.41
HDDA, RF	100	100	100	100	100	100	98.33	98.49	98.41
DWD, RF, SVM	100	100	100	100	100	100	98.33	98.49	98.41
DWD, HDD, SVM	99.00	99.09	99.05	99.09	99.17	99.13	98.33	98.49	98.41
DWD, HDDA, RF	100	100	100	99.09	99.17	99.13	98.33	98.49	98.41
HDDA, RF, SVM	100	100	100	100	100	100	98.33	98.49	98.41
DWD, HDDA, RF, SVM	100	100	100	100	100	100	98.33	98.49	98.41

表 10. データセット別の統合および総合的統合結果

データ種類とサイズ	評価指標	文学作品	学生の作文	日記
文字の bigram	Recall	99.00	98.18	100
	Precision	99.09	98.35	100
	F1-measure	99.05	98.26	100
タグ付き形態素	Recall	98.00	98.18	93.33
	Precision	98.18	98.35	94.17
	F1-measure	98.09	98.26	93.75
タグの bigram	Recall	96.00	96.36	90.00
	Precision	96.27	97.08	91.71
	F1-measure	96.14	96.72	90.84
文節パターン	Recall	100	91.82	91.67
	Precision	100	94.75	92.00
	F1-measure	100	93.26	91.87
総合的統合	Recall	100	100	98.33
	Precision	100	100	98.48
	F1-measure	100	100	98.41

個別のデータセットで正解率が 100%であるのは日記における「文字列の bigram」と文学作品における「文節のパターン」である．各データセットについて 4 つの分類器で統合し結果をさらに統合した結果を表の最下部に示している．この結果はモデル I の結果と同じである．

4. 終 わ り に

本研究では，研究対象から抽出した複数のデータセットと複数の判別方法を用いた統合的判別方法の 2 つのモデルを提案し，匿名文章の書き手の判別を例としてその有効性を実証した．

文章から抽出した書き手の特徴データとしては，文字の bigram，タグ付きの形態素，タグの bigram，文節パターン計 4 種類を用いた．判別方法としては，ADA, DWD, HDDA, LMT, RF, SVM の 6 つの方法から上位 4 種類（DWD, HDDA, RF, SVM）を選出し，統合的判別分析を試みた．

その結果，提案した両モデルには差がなく，文学作品，学生作品では 100 の正解率を得た．日記においては 1 編の日記が誤判別されるのみであった．この判別結果はいずれも，単独分類器の正解率の最大値と同等，あるいはそれ以上である．学生作文においては統合的判別方法により正解率が約 2 ポイント増加した．

提案した統合的判別方法のメリットは，一つの研究対象について異なる側面から抽出したデータを用いて総合的に結論を出すことである．また複数の分類器を用いるので，分類器が個別データへ適合性が欠けているような短所を補うことができる．例えば，RF は全般的には分類性能がよいが，日記の文字の bigram では F1 値が DWD より約 5 ポイント低く，タグ付き形態素データでは F1 値が DWD より約 6 ポイント高い．このようなことから，分類器を組み合わせて統合的に判別する方法でそれぞれの短所を補い，相対的に安定した信頼性が高い結果を得ることが期待できる．

本研究の統合的判別には 4 つの分類器を用いた．それは上位 4 位までとその他には精度に明らかな差が見られなかったからである．4 つの分類器の結果を用いて投票をするとき，4 種類のデータであるので，偶数の結果を用いることになる．幸い，本研究に用いたコーパスでは判定不能なことが起こらなかったが，問題によっては最多投票数が 1 人に決まらないことが起こる可能性は否定できない．その際には，候補となる分類器を追加する方法や，データセットと分類器に重みを付ける方法などの工夫が必要となる．

本研究ではデータセットや分類器は全て平等に扱っている．つまり重みは全て 1 としている．異なる重みを付けることにより，上記の問題は緩和されることが考えられる．どのように重みを付けるかなどは今後の課題にし，別紙に譲りたい．

昨今研究対象からデータを抽出しやすくなった．書き手の同定問題だけではなく，提案する方法はその他の分野における判別・分類問題にも適応できる．

DWD, HDDA は比較的新しい分類器であり，その性能に関する比較分析や応用研究はあまり見られない．本研究に用いたデータでは両方法ともよい結果を示している．特に DWD は高次元小標本では SVM 同等，あるいはそれ以上の正解率を得ている．これらのさらなる実証と応用研究を期待する．

本研究では，ほとんどの特徴データでは文豪の作品に比べ，作文や日記の正解率が低かった．これは用いた文豪の作品における書き手の癖（文体特徴）が作文や日記より顕著であること、用いた文豪の作品は作文や日記より長いためより安定した統計データが得られたことが主な原因であると推測する。これに関する定量的比較研究は別紙に譲る．

また，データによって分類器間の正解率に差が見られるが，これはデータの構造の影響であると考えられる．その誤りに何らかのパターンがあるかに関しては，本研究の結果からは突き止めることができなかった．今後の課題にしたい．

参 考 文 献

Antosch F. (1969). The diagnosis of Literary Style with the Verb-Adjective Ratio. In *Statistics and Style*. Eds. L. Doleszel and R. w. Bailey. New York: American Elsevier.

Bouveyron C., Girard S. & Schmid C. (2007). High-Dimensional Discriminant Analysis. Communications in Statistics: *Theory and Methods*, 36(14), 2607–2623.

Breiman L. (2001). Random Forests. *Machine Learning*, 45, 5–32.

Freund Y., Schapire R. E. (1996). Experiments with a New Boosting Algorithm. In *International Conference on Machine Learning*, pp.148–156.

Grieve J. (2007). Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing*, Vol.22, No.3, 251–270.

Huang H., Y. Liu, Du Y, Perou C., Hayesce D. N., Michael J., Toddg & Marronac J. S. (2012).

Multiclass Distance Weighted Discrimination. *Journal of Computational and Graphical Statistics*. Accepted (29 Jun 2012, http://people.orie.cornell.edu/miketodd/multidwd.pdf)

Jin M & Huh M-H (2012). Author Identification of Korean Texts by Minimum Distance and Machine Learning. *Survey Research* (The koreanAssocia), Vol.13, No.3, 175–190 (The korean Associa)

Jin Mingzhe & Jiang Minghu (2013). Text Clustering on Authorship Attribution Based on the Features of Punctuations Usage. *INFORMATION* (An International Interdisciplinary Journal), Vol.16, No.7(B), 4983–4990.

Jin M. & Murakami M. (1993). Author's characteristic writing styles as seen through their use of commas. *Behaviormetrika*, 20(1), 63–76.

Landwehr N., Hall M. & Frank E. (2005). Logistic Model Trees. *Machine Learning*, Vol.59, Issue 1-2, pp.161–205.

Marron J.S., Todd M.J. & Ahn J. (2007). Distance-Weighted Discrimination. *Journal of the American Statistical Association*, Vol.102, No.480, pp.1267–1271.

Sebastiani F. (2002). Machine Learning in Automated Text Categorisation. *ACM Computing Surveys*, Vol.34, No.1, 1–47.

Vapnik V. N. (1998). *Statistical Learning Theory*, John Wiley & Sons.

金 明哲 (1994). 読点の打ち方と文章の分類. 計量国語学, 19(7), 317–330.

金 明哲 (2003). 中国文章における書き手の識別. 第二届中国社会語言学国際学術検討会中国社会語言学会成立大会要旨集, 31.

金 明哲 (2004a). 社会科学における統計学の応用研究. 国際学術シンポジューム論文集, 人民大学（北京）, 17–25.

金 明哲 (2004b). 品詞のマルコフ遷移の情報を用いた書き手の同定. 日本行動計量学会第 32 回大会抄録集, 384–385.

金 明哲 (2012). 文章の書き手特徴情報と書き手の識別. 『コーパスとテキストマイニング』（石田 基広・金 明哲 編著, 共立出版）, 55–68.

金 明哲 (2013). 文節パターンに基づいた書き手の同定. 行動計量学, Vol.40, No.1, 17–28.

金 明哲, 樺島忠夫, 村上征勝 (1993a). 読点と書き手の個性. 計量国語学, 18(8), 382–391.

金 明哲, 樺島忠夫, 村上征勝 (1993b). 手書きとワープロとによる文章の計量分析. 計量国語学, 19 巻 3 号, 133–145.

金 明哲, 村上征勝 (2007). ランダムフォレスト法による文章の書き手の同定. 数理統計, 第 55 巻, 第 2 号, 255–268.

金 明哲・宮本加奈子 (1999). ラフな意味情報に基づいた文章の自動分類. 言語処理学会第 5 回年次大会発表論文集, 235–238（於 電気通信大学）

松浦 司, 金田康正 (2000). n-gram の分布を利用した近代日本文の著者推定. 計量国語学, 22(6), 225–238.

村上征勝, 伊藤瑞叡 (1991). 日蓮遺文の数理研究. 東洋思想と宗教, 8, 27–35.

（2013 年 9 月 23 日受付，2014 年 2 月 14 日最終修正）