F-013

# クラス情報を用いない相関ルールによるクラス分類手法

# Classification Method based on Association Rules without Class Attributes

岡田 恵理香† 杉村 博‡ 佐賀 亮介† 松本 一教†‡
Erika Okada Hiroshi Sugimura Ryosuke Saga Kazunori Matsumoto

## 1. はじめに

クラス分類知識の獲得には様々な応用があり，以前から多くの研究がなされてきている．決定木やニューラルネットワーク[1]を用いる手法等は広く利用されている．

本論文では，クラス分類の中でも特に文書の著者同定[2]に焦点を絞って考察する．著者同定とは，著者不明な文書の著者を推定する問題である．これを通常のクラス分類として扱うためには，文書の著者をクラス値とした学習データとして使って分類知識を抽出することが一般的である．この方法で獲得される知識は，文書を学習データ中で与えられるクラスのいずれかに分類するためのものであるため，未知の著者の同定には不向きである．

本研究では，相関ルール[1]を用いた新たな著者同定方法を開発する．相関ルールとは，データの同時生起に関する規則性をルール形式で抽出したものであり，そのままではここでの目的に用いることはできない．そこで，ある著者の文書に対する相関ルールを著者不明の文書に適用したときの適合度により著者同定を行う方法を開発する．この方法は，知識獲得時に著者情報（クラス値）を与える必要がないため，従来手法より幅広く適用できる．また相関ルールは，対象となるデータの特徴を表現するものであるため，それを使って行った著者同定結果は人間が解釈しやすいものであることも特徴となる．本論では，提案する手法を文学作品を用いた実験によりその有効性の検証も行う．

## 2. 相関ルール

相関ルールおよびそのマイニング手法の詳細な説明は文献[1]に譲り，ここでは概要だけを説明する．相関ルールはアイテム集合 X，Y に対して，ただし X∩Y= φ，X⇒Y (s,c)の形式で表現される．0≦s, c≦1 は各々支持度(support)と確信度(confidence)であり以下の式で定義される．ここで，アイテム集合 A に対して， A を含むトランザクション集合を T[A]で表し，その要素数を#T[A]と表す．N はデータベース中の全トランザクション数である．

$$s = \frac{\#T[X \cup Y]}{N}, c = \frac{\#T[X \cup Y]}{\#T[X]} \quad (1)$$

与えられた最小支持度 $s_{min}$ と最小確信度 $c_{min}$ をに対して，s≧$s_{min}$ かつ c≧$c_{min}$ となる全ての相関ルールX⇒Y (s,c)を求めることが相関ルールマイニングである．

†神奈川工科大学情報学部情報工学科
Kanagawa Institute of Technology,Faculty of Information Technology,Department of Information and Computer Sciences.
‡神奈川工科大学工学研究科情報工学専攻
Course of Information and Computer Sciences,Graduate School of Engineering, Kanagawa Institute of Technology

## 3. 著者同定手法の開発

### 3.1 従来方法

著者同定には様々な手法が開発されている[2,3,4]．クラス分類に相関ルールを利用する方法が開発されている[5,6]ので，これを著者同定に利用することも可能である．この方法は，相関ルールの右辺（帰結部）がクラス値に限るという制限の元でルールを獲得するものである．このような相関ルールをクラス相関ルールとよぶ．これらの従来手法は，図 1 に示すように，事前に与えられた著者達の文書を用いて文書をそれら著者のいずれかに分類する知識を獲得する方法である．これらの手法では，ルール獲得時の学習データに含まれる著者達の分類にのみ有効な知識が獲得される．

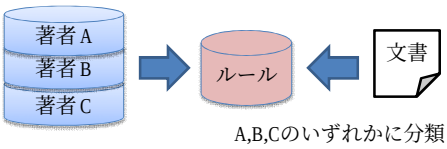


図 1. 従来方式の著者同定

### 3.2 開発した方式

本研究での開発は，学習データで固定される著者達に対する識別ルールを得ることではなく，未知の著者に対する判定能力も持つルールを獲得することである．この概念図を図 2 に示す．

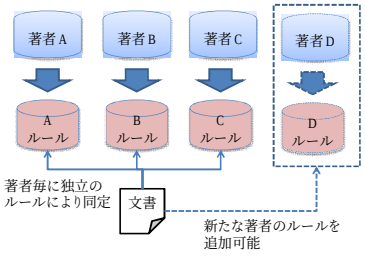


図 2. 開発する著者同定方式

データベース $D_1$ から抽出された相関ルールX⇒Y (s1,c1)に対して，このルールの他のデータベース $D_2$ への適用を考える．データベースが異なるため，当然ながら支持度，確信度ともに変化する．その値を $s_2$，$c_2$ とする．このルールの $D_2$ における反適合度を以下で定義する．いずれのデータベースに対しても同一の支持度，確信度であれば，この値は 0 となる．また，ua の最大値は$\sqrt{2}$である．

$$ua = \sqrt{(s_1 - s_2)^2 + (c_1 - c_2)^2} \quad (2)$$

この反適合度を相関ルール集合に拡張する．$D_x$ からの相関ルール集合$R_x = \{r_1, r_2, \ldots, r_n\}$に対して，その $D_y$ への反適合

度は以下のように，$R_x$ の各ルールに対する反適合度の平均として定義する．ここに，$d_i$は$r_i$の $D_y$ への反適合度である．

$$ua(D_x, D_y, R_x) = \frac{\sum_{i=1}^{n} d_i}{n} \quad (3)$$

反適合度を用いた著者同定は次のようにして行う．

1. 著者 i の文書からなるデータベース $D_i$に対して，最小支持度および最小確信度を設定して，相関ルールマイニングを行いルール集合 $R_i$を得る．
2. 著者が未知の文書 d に対し，全てのデータベース $D_i$に対して反適合度$ua(D_i, d, R_i)$を求める．
3. 最小の反適合度となる著者のデータベースに対して，d はそれと同一の著者であると同定する．
4. 新たな著者 x のデータベース $D_x$が得られた時は，上記の 2 と同様にして反適合度の判定を行い，それまでよりも小さな値となれば著者同定を x に変更する．

このような手順で著者同定を行うことができるが，上記の最終ステップで明らかなように，新たな著者データベースが与えられた場合には，そのデータベースに対する反適合度の計算だけを行うだけで十分である．

## 4. 文学作品による実験

提案手法の有効性を実際の文学作品を用いて検証した．本実験では，青空文庫[5]より著者 10 人をランダムに選択し，一人につき 3 作品ずつランダムに選び出し，この著者のデータベースとする．各作品の長さは 5000 文字に揃えた．文書からトランザクションデータベースへの変換は以下のように行う．句点で区切られた 1 文をトランザクションとし，単語をアイテムとして扱う．予備実験により，作品への依存性が高い人名や固有名詞が悪影響を与えることが確認できたので，以降の実験では名詞や特殊記号は除去して扱うことにした．この結果，平均的に 1 作品は 383.6 トランザクションとなり，アイテム数の平均は 5114 個となった． このようにして準備した 10 人著者分のデータベースの各々に対して相関ルールマイニングを行った．最小支持度，最小確信度の変化に伴って，獲得される相関ルール数は表 1 のようになった．最小支持度が小さくなるに伴って，獲得されるルール数は指数関数的に増大していく．本手法による著者同定の正答率を表 2 に示す．相関ルールの反適合度を用いるため，最小支持度と最小確信度の変化に伴って正答率も変化していることが確認できる．図 2 には相関ルール数の変化に伴う正答率の変化をプロットしたものである．

抽出される相関ルール数は著者データベースの性質により著しく変動する．最小信頼度 0.1 で最小確信度が 0.9 の場合には，表中で示すように平均的には 132.7 個のルールが得られる．しかし，最も少ない場合には 1 個であり，最大では 499 個と大きな変動がある．標準偏差は 177.5 であった．この変動は正答率に影響を与える．

表 1 獲得される相関ルール平均数の変化

| | | 最小確信度 | | |
|---|---|---|---|---|
| | | 0.90 | 0.95 | 0.99 |
| 最小支持度 | 0.05 | 1536.2 | 675.8 | 380.2 |
| | 0.10 | 132.7 | 41.1 | 9.3 |

表 2　正答率の変化

| | | 最小確信度 | | |
|---|---|---|---|---|
| | | 0.90 | 0.95 | 0.99 |
| 最小支持度 | 0.05 | 93.3% | 90.0% | 83.3% |
| | 0.10 | 56.7% | 10.0% | 10.0% |

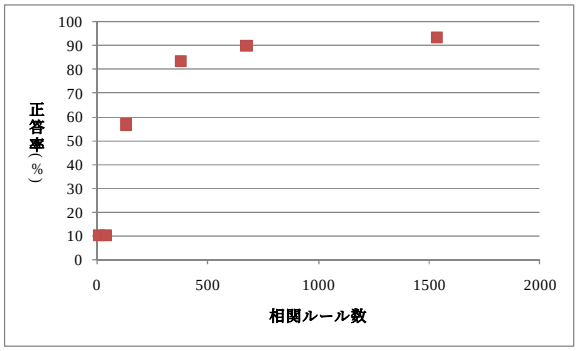


図 3. ルール数と正答率の関係

### 4.1 他研究との比較および今後の課題

本手法の有効性を他の研究と比較する．文献[4]では本論文と同じ青空文庫のデータを用いて，テキストから抽出する N-gram に対して構成する決定木を用いた著者同定を行っている．本方式とほぼ同程度の正答率を得るためには 20000 文字程度のテキスト長を必要としている．また文献[3]では，テキストから抽出する N-gram に対してその確率う分布を著者毎に求めておき，未知の文書の N-gram 確率分布と比較する方式を提案している．さらに分布比較のために複数の方式を検討している．ここでも正答率が 90%に達するのは 20000 文字程度を必要としている．これら他の研究との比較の結果，本方式は 5000 文字程度で高い正答率に到達できる点で優れているといえる．

表 1 で示したように，相関ルール数は最小支持度と最小確信度により変化するが，文書の特徴によっても大きく変化することが分かった．

## 5. おわりに

本論文では，相関ルール抽出時にクラス情報を用いない，従来型の相関ルールマイニングにもとづいてクラス分類知識を獲得する方法を開発した．分類対象となるクラスが増加しても，新たに学習をやり直す必要がないことが特徴であり，文書の著者同定などへの応用が期待できる．

### 参考文献

[1] 元田 浩 ほか，データマイニングの基礎，IT Text 情報処理学会編集，オーム社，(2006)
[2] 金明哲 ほか，言語の心理の統計，岩波書店 (2003)
[3] 松浦司，金田康正，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，Vol.22， No.6，(2000).
[4] 谷口裕太 ほか，N-gram と決定木による著者識別，第 24 回人工知能学会全国大会 , (2010)
[5] 青空文庫, http://www.aozora.gr.jp/
[6] Coene, F. , Leng, P. and Zhang, L., Threshold Tuning for Improved Classification Association Rule Mining, Proc. Pacific-Asia Conf. In Knowledege Discovery and Data Mining (PAKDD05), (2005).
[7] Yin, X. and Han, J., Classification based on Predictive Association Rules, *Proc. SIAM Int. Conf. on Data Mining* (SDM03), (2003)