文書分類システムの分類誤りに着目した分類ルール修正法

桂田浩一[†] 小山誠[††,☆] 大原剛三[††]
馬場口登[††] 北橋忠宏[††]

本論文では if-then 形式のルールを分類規範として利用する文書分類システムにおいて，規範を利用者に適応させるためにルールを修正していく手法を提案する．提案手法では，システムが分類を誤った場合に，誤分類される文書（あるルールに対する例外文書）が存在することをユーザがシステムに指摘できる．この指摘に基づいてシステムは誤分類を導くルールを検出し，そのルールをデフォルトルール（例外文書に関してのみ無効となるルール）に変換する．この変換により例外文書に関してルールが適用されなくなるため，当該文書の誤分類が防がれるようになる．さらに例外文書に共通して現れる単語を統計的に推定し，それを基にして例外に関するルールを作成する．これによって新たな例外文書の判別が可能となり，分類精度が向上する．これらの例外処理手法を実装し，電子新聞記事を用いた分類実験によってその性能を検証したところ 1.3%〜6.7%分類精度が向上することが分かり，例外処理の有効性が確認できた．

Modification of Document Classification Rules Based on Misclassification

Kouichi Katsurada,[†] Makoto Koyama,[††,☆] Kouzou Ohara,[††]
Noboru Babaguchi[††] and Tadahiro Kitahashi[††]

This paper provides a method to modify some document classification rules for the purpose of user adaptation. A document classification system sometimes classifies a document into a wrong category. When it occurs, in our method, the user can point out the mis-classification. Given this as a trigger, our method firstly detects the rule which causes this mis-classificaton. Then it converts the detected rule into a default rule which is not applied to the mis-classified document (we call this document an exceptional document). Moreover it creates a new rule to avoid subsequent mis-classifications caused by the same sort of exceptional documents. The experimental results for RWC Text Database show that our method can improve accuracy of the classification.

1. はじめに

近年，文書の自動分類がさかんに検討されている．文書の自動分類とは，単語の出現頻度等に基づいて文書を適切なカテゴリに自動的に振り分ける技術であり，サポートベクトルマシン等で求まる分類境界面を用いた手法[17]や，帰納学習等で生成される if-then 形式のルールを用いた手法[1),3)]等，様々な方法が提案されている[14),15)]．こうした文書分類技術を個人が日常的に利用する場合，分類規範の定め方，および分類過程や規範変更の情報提示が問題となる．一般的に各文書へのカテゴリ付与の基準は個人によって異なるため，すべてのユーザが満足できるような規範を作成するのは不可能である．したがって，たとえ何らかの標準的な規範が準備されたとしても，それを個別に修正していくメカニズムが必要となる．また規範が修正された場合に，修正後の規範が必ずしも各個人が持つ基準と一致するとは限らない．こうした場合に各個人が分類過程や規範の変化を独自にチェックし，修正できると，より個人に適応した分類システムとして利用できるようになる．

分類規範を修正することのできる文書分類システムとして，これまで WINNOW[10)]や WIDROW-HOFF[19)]といったオンライン学習を用いた手法がいくつか提案されている[4),5),11)]．これらの手法では，分類結果の正誤に基づいて分類で用いる重みベクトル

† 豊橋技術科学大学
Toyohashi University of Technology
†† 大阪大学
Osaka University
☆ 現在，株式会社東芝
Presently with Toshiba Corporation

を更新していき，分類精度を向上させている．しかしながら，これらの手法では重みベクトルがパラメータで表されるため，分類の過程や規範変更の様子が個人ユーザにとって直観的に理解しにくく，ユーザ独自の修正も難しい．一方，if-then 形式のルールを分類規範として用いると分類過程およびルールの更新の様子が直観的に理解しやすく，ルールの追加や修正も容易である．しかしながら従来研究[1),3)]では分類ルールの獲得法が主に検討されており，ルールの修正についてはあまり検討されていない．そこで本論文では個人利用を目的とする文書分類システムにおける if-then 形式のルールの修正法を検討する．

提案手法では，システムが分類を誤った場合に，誤分類される文書（あるルールに対する例外文書）が存在することをユーザがシステムに指摘でき，これを契機として分類ルールが格納されている知識ベースを修正する．まずユーザの指摘が分類の否定情報として知識ベースに格納される．この否定情報は，システムがすでに導いている分類結果と相反するため矛盾が生じる．本研究ではこの矛盾を契機として，誤分類を導く**通常ルール**を，例外文書に関してのみ無効化される**デフォルトルール**に変換する（以降，この変換を知識コンバージョン（KC: Knowledge Conversion）[2)]と呼ぶ）．この変換により例外文書に関してルールが適用されなくなるため，当該文書の誤分類が防がれるようになる．さらに例外文書に共通して現れる単語を統計的に推定し，それに基づいて例外に関するルールを作成する．これによって，新たな例外文書が誤分類されることを防ぐ．こうした例外処理手法により誤分類が減少するため，分類精度が向上する．

本論文では以上で述べた手法を提案・実装し，電子新聞記事[12)]を用いた分類実験を通じてその有効性を示す．

2. 文書分類システムの概要

2.1 システムの構成

本研究で用いる文書分類システムは**図 1** のように**特徴抽出部**，**学習部**，**推論部**，**例外処理部**，および**知識ベース**から構成される．学習文書が入力されると，システムはまず特徴抽出部で文書の特徴を抽出し，その特徴を基に学習部において初期ルールを学習する．この初期ルールは知識ベースに格納され，新規文書が入力された場合の分類に用いられる．新規文書の分類が誤りであった場合には，誤りを導いたルール集合がユーザの指摘を契機として推論部で検出され，例外処理部においてルール集合が修正される．本章ではまず

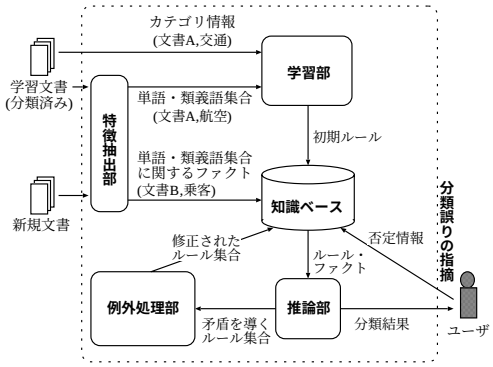

図 1 文書分類システムの構成
Fig. 1 Document classification system.

知識ベースと推論部を説明した後に，例外処理部を除くその他各部を順に説明する．

2.2 知識ベース

各部で得られたルールや特徴は知識ベースに格納される．ここで知識ベースにおけるこれらルール・特徴の格納形式を示す．

通常ルール

$$cat\text{-}c_j(x) \leftarrow word_1(x), \cdots, word_m(x)$$
$$cat\text{-}c_j(x) \leftarrow group_1(x), \cdots, group_m(x)$$

デフォルトルール

$$cat\text{-}c_j(x) \Leftarrow word_1(x), \cdots, word_m(x)$$
$$cat\text{-}c_j(x) \Leftarrow group_1(x), \cdots, group_m(x)$$

制約

$$\bot \leftarrow cat\text{-}c_j(x), not\text{-}cat\text{-}c_j(x)$$

ファクト

$$fact(\mathrm{A}) \qquad \square$$

通常ルール，デフォルトルール，および制約中の x，$cat\text{-}c_j$，$not\text{-}cat\text{-}c_j$，$word_i$，$group_i$，$\bot$ は順に，文書に対応する変数，分類カテゴリ，分類カテゴリの否定，単語，類義語集合，矛盾記号を表す．また $word_i(x)$，$group_i(x)$，$cat\text{-}c_j(x)$，$not\text{-}cat\text{-}c_j(x)$ は**リテラル**であり，それぞれ「文書 x が単語 $word_i$ を含むならば真である」,「文書 x が類義語グループ $group_i$ に含まれる単語を少なくとも 1 つ含むならば真である」,「文書 x はカテゴリ c_j に含まれる」,「文書 x はカテゴリ c_j に含まれない」と解釈する．記号 “$\leftarrow$”，および “$\Leftarrow$” の左側はルールの**ヘッド**，右側はルールの**ボディ**と呼ぶ．各ルール・制約の解釈は以下のとおりである．

通常ルール：「ボディ中のリテラルがすべて真ならば文書 x はカテゴリ c_j の文書である」

デフォルトルール：「ボディ中のリテラルがすべて真であり，もし $cat\text{-}c_j(x)$ を導いても矛盾が発生し

ないならば，文書 x はカテゴリ c_j の文書である」

制約：「ボディ中のリテラルがすべて真ならば**矛盾**が発生する」

ファクトはリテラル中の変数に，文書に対応する定数を代入したものであり，文書の特徴等を表すために用いる．

2.3 推論部での処理

推論部では知識ベースに格納されたルールやファクトの内容に基づき，文書の分類先を決定する．たとえば以下の知識ベース 1 を考える．

[知識ベース 1]

通常ルール

(i) *cat*-刑法 $(x) \leftarrow$ 容疑 (x)

デフォルトルール

(ii) *cat*-交通 $(x) \Leftarrow$ 航空 (x), 空港 (x)

ファクト

(iii) 航空 (A)

(iv) 空港 (A)

(v) 容疑 (A)

(vi) *not-cat*-交通 (A)

制約 (vii) $\perp \leftarrow$ *cat*-交通 (x), *not-cat*-交通 (x) □

知識ベース 1 の通常ルール (i) の x に A を代入した場合，ファクト (v) と通常ルール (i) から**推論結果** "*cat*-刑法 (A)" が導かれる．これは文書 A がカテゴリ "刑法" に分類されることを表す．一方デフォルトルール (ii) の x に同様に A を代入した場合，デフォルトルール (ii) とファクト (iii)，(iv) から推論結果 "*cat*-交通 (A)" を導くことができるが，その推論結果 "*cat*-交通 (A)" とファクト (vi)，制約 (vii) によって矛盾が発生する．このため，デフォルトルールの解釈より，結果としてデフォルトルール (ii) から "*cat*-交通 (A)" は導かれない．

2.4 特徴抽出部での処理

特徴抽出部では，学習部と推論部で利用する文書の特徴を，文書中から抽出する．本研究では文書の特徴として文書中の単語，およびその単語に関する類義語集合を用いる．類義語集合はシソーラスから取得され，たとえば単語 "学生" に関する類義語集合は $G_{\text{学生}} = \{$ 学生，学士，生徒,...$\}$ のように参照する．

得られた特徴のうち新規文書の特徴は，ファクト形式で知識ベースに格納される．一方，学習文書の特徴は学習部に与えられ，次節で述べる学習テーブルの作成に利用される．

2.5 学習部での処理

学習部では各学習文書に 1 つ以上与えられているカテゴリの情報と文書の特徴に基づいて，通常ルールからなる初期ルールを生成する．実質的には，各カテゴリに属する文書中の特徴的な 1 つ以上の単語の組合せに基づいてルールを生成できればよく，本論文では生成されるルールの性能をパラメータによって調整しやすいという理由から，**学習テーブル**[7]に基づいたルールの生成法を利用する．以下では学習テーブルの作成法，およびその学習テーブルに基づいた初期ルールの生成法について順に述べる．

表 1 単語の学習テーブル
Table 1 Learning table of words.

カテゴリ＼単語		試合		受験	
スポーツ	···	134.9	···	0.0	···
教育	···	3.7	···	98.0	···
⋮			⋮		

表 2 類義語集合の学習テーブル
Table 2 Learning table of synonyms.

カテゴリ＼類義語集合		$G_{\text{大会}}$		$G_{\text{学生}}$	
スポーツ	···	161.6	···	3.8	···
教育	···	2.3	···	86.4	···
⋮			⋮		

2.5.1 学習テーブルの作成

学習テーブルはカテゴリごとに単語，および類義語集合の出現頻度の偏りを得点化して表したものであり，本研究では**表 1** および**表 2** のように単語，類義語集合のそれぞれについて作成する．テーブルの各行はカテゴリを，各列は単語（類義語集合）を表し，各欄の得点が高い単語（類義語集合）ほど，当該カテゴリに偏って出現することを表す．単語の学習テーブル中のカテゴリ c_j，単語 $word_i$ の欄の得点 s_{ij} は次式[7]で与えられる☆．

$$s_{ij} = \frac{(\alpha_{ij} - \beta_{ij}) \times |\alpha_{ij} - \beta_{ij}|}{\beta_{ij}}$$

$$\beta_{ij} = \frac{\sum_{i'=1}^{n_w} \alpha_{i'j}}{\sum_{i'=1}^{n_w} \sum_{j'=1}^{n_c} \alpha_{i'j'}} \times \sum_{j'=1}^{n_c} \alpha_{ij'}$$

n_w ：異なる単語の数
n_c ：カテゴリ数
α_{ij} ：カテゴリ c_j における単語 $word_i$ を持つ文書数
β_{ij} ：単語 $word_i$ がすべてのカテゴリに等確率で出現する場合の，カテゴリ c_j における単語 $word_i$ を持つ文書数

式中の i，i' は単語と，j，j' はカテゴリと 1 対 1 に対応する自然数であり，それぞれ $1 \leq i$，$i' \leq n_w$，

☆ 本論文では簡便さから上記の得点付けを選んだが，他の方法（tf-idf や相互情報量を用いた方法）を用いても問題はない．

$1 \leq j, j' \leq n_c$ である．α_{ij} と β_{ij} はそれぞれカテゴリ c_j における単語 $word_i$ を持つ実際の文書数と，確率的な根拠に基づく推定値である．したがって s_{ij} の値は，定義より，α_{ij} と β_{ij} の差が大きいほど，すなわち単語 $word_i$ を持つ文書がカテゴリ c_j へ偏って出現するほど，大きくなる．類義語集合についても同様の式を用いて得点を求める．各式の詳細については文献 7) を参照されたい．

2.5.2 初期ルールの生成

初期ルールを生成する際には，まず学習文書に付与されたカテゴリについて，その文書に含まれる単語の学習テーブルでの得点を調べる．この得点が高い単語は，学習文書に付与されたカテゴリに偏って出現した単語であり，そのカテゴリの識別に重要な単語といえる．そこでそのような単語をボディに持つ通常ルールを生成する．得点の大きさはカテゴリごとに異なり，その閾値をカテゴリごとに定めるのは困難なため，提案手法では，ユーザが初期ルールのボディ中のリテラルの数をパラメータ $b(>0)$ として与える．以下に示す初期ルール生成アルゴリズム中の i は単語，j はカテゴリを表す自然数であり，それぞれに対応する単語，カテゴリを $word_i, c_j$ で表す．$make\text{-}body(W)$ は単語の集合 $W = \{word_1, \cdots, word_b\}$ に対して，ルールのボディ“$word_1(x), \cdots, word_b(x)$” を返す関数である．

アルゴリズム 2.1 （初期ルールの生成）
入力：学習文書の集合 D，学習テーブル内の得点 $\{s_{ij} \mid 1 \leq i \leq n_w, 1 \leq j \leq n_c\}$，初期ルールのボディ中のリテラル数 b
出力：ルール集合 R
1) $R := \emptyset$.
2) 各 $d \in D$ について以下を実行.
2.1) $J := \{d$に付与されたカテゴリを表す自然数 $\}$.
2.2) 各 $j \in J$ について以下を実行.
2.2.1) $I := \{i | word_i$が d に含まれる $\}$
2.2.2) s_{ij} $(i \in I)$ の中から得点の高い上位 b 個を選択．$s_{a_1 j}, \cdots, s_{a_b j}$ とする．
2.2.3) $body := make\text{-}body(\{word_{a_1}, \cdots, word_{a_b}\})$
2.2.4) $R := R \cup \{cat\text{-}c_j(x) \leftarrow body\}$ □

類義語集合に関しても同様の方法を用いて，類義語集合の学習テーブルから通常ルールを生成する．

3. 例外処理機構

本章では，本論文の主題である分類誤りに着目した分類ルールの修正法について述べる．1 章で述べたように，初期ルールはユーザ個人ごとに異なる分類の基準を正確に反映していない場合があるため，個別に修正していく必要がある．そこで例外処理部では，ユーザが指摘した分類誤りを**否定情報**として取得し，取得した否定情報に基づき初期ルールの一部を**デフォルトルール**に変換する．さらに新たな誤分類を防ぐために**例外に関するルール**を生成する．以下ではこれらについて順に説明する．

3.1 否定情報の取得

システムによる誤分類がユーザによって指摘された場合，システムはユーザの指摘を否定情報としてファクトおよび制約の形式で知識ベース内に取り込む．たとえば次の知識ベース 2 を考える．

[**知識ベース 2**]
通常ルール
(1) *cat*-交通 $(x) \leftarrow$ 航空 (x), 空港 (x)
ファクト
(2) 航空 (A)
(3) 空港 (A)
(4) 容疑 (A) □

知識ベース 2 からは，推論結果 “*cat*-交通 (A)” が導かれる．この推論結果がユーザによって誤りと判断され，指摘されたとすると，ファクト “*not-cat*-交通 (A)”，および**制約** “$\perp \leftarrow$ *cat*-交通 (x), *not-cat*-交通 (x)” が知識ベース 2 に追加され，下の知識ベース 3 のようになる．

[**知識ベース 3**]
通常ルール
(1) *cat*-交通 $(x) \leftarrow$ 航空 (x), 空港 (x)
ファクト
(2) 航空 (A)
(3) 空港 (A)
(4) 容疑 (A)
(5) *not-cat*-交通 (A)
制約 (6) $\perp \leftarrow$ *cat*-交通 (x), *not-cat*-交通 (x) □

3.2 知識コンバージョン

知識ベース 3 からは “*not-cat*-交通 (A)” と “*cat*-交通 (A)” がともに導かれ，制約 (6) のボディ中のリテラルがすべて真になるため矛盾が発生する．知識ベースの矛盾が推論部で検出された場合，矛盾を導く際に用いられたルール，ファクトが例外処理部に与えられる．例外処理部では発生した矛盾を解消するために，通常ルールからデフォルトルールへの変換操作（KC）を実行する．

KC を実行する際には次の仮定が置かれる[9]．
a) 矛盾を導く過程で導かれた推論結果は偽である．
b) ファクトは必ず真である．

この仮定の下で，c) **ボディ中のリテラルがすべて真**

かつヘッドが偽となるルールがデフォルトルールに変換される[9]．たとえば知識ベース 3 では，a) に該当する推論結果は “*cat*-交通 (A)” であり，b) に該当するファクトは “航空 (A)”，“空港 (A)”，“容疑 (A)”，“*not-cat*-交通 (A)” であるため c) を満たす通常ルール (1) がデフォルトルールに変換される．この結果，推論結果 “*cat*-交通 (A)” は導かれなくなり，カテゴリ “交通” への誤分類は解消される．

3.3 例外に関するルールの生成

前節で述べたデフォルトルールへの変換によって，文書 A の誤分類は防がれる．しかしながら，この変換だけでは，ユーザからの否定情報をファクトとして持たない新たな例外文書の誤分類は防がれない．たとえば知識ベース 3 にさらに文書 B に関するファクトが追加された知識ベース 4 を考える．

[知識ベース 4]

デフォルトルール

(1) *cat*-交通 $(x) \Leftarrow$ 航空 (x), 空港 (x)

ファクト

(2) 航空 (A)
(3) 空港 (A)
(4) 容疑 (A)
(5) *not-cat*-交通 (A)
(7) 航空 (B)
(8) 空港 (B)
(9) 容疑 (B)

制約 (6) $\bot \leftarrow$ *cat*-交通 (x), *not-cat*-交通 (x) □

もし文書 B がルール (1) の例外文書であるならば，その誤分類を防ぐためにはファクト “*not-cat*-交通 (B)” が必要となり，ユーザは再び先の文書 A のときと同様の誤分類を指摘しなければならない．しかし，このとき，誤分類される例外文書間に共通の特徴が存在すれば，文書 A を基に例外文書の性質を表すルールを生成することにより，同様の新たな例外文書の誤分類を防ぐことが可能になる．そこで知識コンバージョンを実行した後，新たな例外文書の誤分類を防ぐため，例外文書間に共通して出現すると推定される単語の集合 $\{word_1, \ldots, word_m\}$ および誤分類されたカテゴリ c_j の否定について，**例外に関するルール** $not\text{-}cat\text{-}c_j(x) \leftarrow word_1(x), \ldots, word_m(x)$ を生成する．

このために，まず誤分類された例外文書中の単語から，その単語が偏って出現しているカテゴリを学習テーブルから求める．さらに求められたカテゴリに偏って出現するその他の単語集合を，当該カテゴリをヘッドとするルールのボディから求める．最後に求められた単語集合 $\{word_1, \ldots, word_m\}$，誤分類が生じたカテゴリ c_j の否定について上述のルールを生成する．

以下に，文書 A のカテゴリ c_j への分類が取り消されたときの，例外に関するルールの生成アルゴリズムを示す．ただし，i は単語，j はカテゴリを表す自然数であり，それぞれに対応する単語，カテゴリを $word_i, c_j$ とする．t はユーザパラメータであり，学習テーブル内の得点の和が t を超える単語に対応するリテラルの集合をボディとするようなルールのみが生成される．また，r はルール，$body(r)$ はルール r のボディ，$H\_r(j)$ はカテゴリ c_j に対して $cat\text{-}c_j(x)$ をヘッドとするルールの集合とする．

アルゴリズム 3.1 （例外に関するルールの生成）
入力：文書 d，カテゴリ c_j，学習テーブル内の得点 $\{s_{ij} \mid 1 \le i \le n_w, 1 \le j \le n_c\}$，$t$
出力：例外に関するルール R

1) $R := \emptyset$

2) $j' \in \{1, 2, \cdots, n_c\}, j' \neq j$ なる各 j' について以下を実行．

2.1) $S(j', d) := \sum_{i=1}^{n_w} s'_{ij'}(d)$．ただし，

$$s'_{ij'}(d) = \begin{cases} s_{ij'} & (d \text{ 中に } word_i \text{が存在}) \\ 0 & (\text{otherwise}) \end{cases}$$

2.2) $S(j', d) > t$ ならば，
$R := R \cup \{not\text{-}cat\text{-}c_j(x) \leftarrow body(r) \mid r \in H\_r(j')\}$
□

なお，$S(j', d)$ は，文書 d に含まれる単語の，カテゴリ c'_j における学習テーブルでの得点の和であり，文書 d がカテゴリ c'_j に偏って出現した単語を多く含めば含むほど，$S(j', d)$ の値は大きくなる．

ここで知識ベース 3 のルール (1) がデフォルトルール化され，さらにいくつかの通常ルールが加わった知識ベース 5 を考える．

[知識ベース 5]

通常ルール

(10) *cat*-刑法 $(x) \leftarrow$ 容疑 (x)
(11) *cat*-教育 $(x) \leftarrow$ 大学 (x)

デフォルトルール

(1) *cat*-交通 $(x) \Leftarrow$ 航空 (x), 空港 (x)

ファクト

(2) 航空 (A)
(3) 空港 (A)
(4) 容疑 (A)
(5) *not-cat*-交通 (A)

制約 (6) $\bot \leftarrow$ *cat*-交通 (x), *not-cat*-交通 (x) □

たとえば**図 2** のように学習テーブルにおける得点の

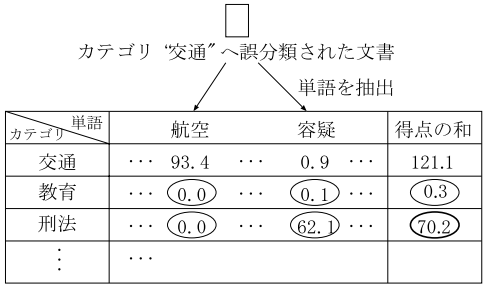

図 2 カテゴリの同定
Fig. 2 Identification of category.

和の閾値 t が 50 であった場合，この文書に含まれる単語 “航空”，“容疑” 等から，閾値 t を超えるカテゴリとしてカテゴリ “刑法” が求められる．知識ベース 5 には，カテゴリ “刑法” に関するルール (10) が存在するため，ルール “*not-cat*-交通 (x) ← 容疑 (x)” が生成される．これにより，以後，文書 A と同様に “航空”，“空港” とともに “容疑” という単語を持つ文書 B が追加されても，上のルールによって “*not-cat*-交通 (B)” が導かれて “*cat*-交通 (B)” が導かれなくなるため，カテゴリ “交通” への誤分類は防がれる．

なお本分類システムでは簡単化のために，例外に関するルールは単語のみをボディに用い，デフォルトルール化しないものとする．もし知識ベース内の正解の分類と例外に関するルールが矛盾した場合は，その例外に関するルールを削除し，正解の分類が導かれなくなるのを防ぐ．

4. 実験と考察

4.1 実験仕様

実装には C 言語と Perl5.0 を用い，PC（CPU 550 MHz，メモリ 320 MB）上で実験を行った．実験に使用した文書データと単語，類義語集合は以下のとおりである．

- 文書

RWC テキストデータベース[18]を使用した．このデータベースは，1994 年版の毎日新聞[12]の約 3 万件の記事に，国際十進分類法に基づく UDC コード[6]を付与したものである．これらの記事の中から，出現頻度の高い 10 種類の分類カテゴリに属する合計 2,000 記事を実験に用いた．

- 単語

各文書から名詞と動詞を形態素解析システム茶筅 2.0b6 [13]を用いて抽出した．

- 類義語集合

表 3 カテゴリごとの記事数
Table 3 Number of articles in each category.

カテゴリ	学習・訓練記事	テスト記事	合計
スポーツ	161	147	308
犯罪	156	148	304
政府	135	142	277
教育システム	110	124	234
交通	112	103	215
軍事	110	118	228
国際関係	96	97	193
コミュニケーション	76	83	159
演劇	86	95	181
農業	78	72	150

シソーラスとして分類語彙表[8]を利用し，その中の最下層の類義語集合 5,134 個を使用した．なお，1 つの類義語集合には平均約 6.3 個の単語が含まれている．

分類精度の評価尺度としては，以下の再現率（recall），適合率（precision）を利用した．

$$\text{再現率} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100 \quad (1)$$

$$\text{適合率} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100 \quad (2)$$

TP = 正しいカテゴリに分類された記事数
FP = 誤ったカテゴリに分類された記事数
FN = 正しいカテゴリに分類されなかった記事数

例外処理によって誤分類が解消されると，**式 (2)** より適合率が向上する．しかしながら誤分類を解消する際に，例外と無関係な記事までもが正しく分類されなくなると**式 (1)** より再現率の低下を招く．したがって再現率を低下させることなく適合率を向上することが本実験の目標となる．

4.2 実験内容

実験で用いた文書中の各カテゴリごとの記事数は**表 3** に示すとおりである．性能比較を容易にするために，表 3 の学習・訓練記事（1,000 記事）およびテスト記事（1,000 記事）は，それぞれ文書分類の従来研究[17]で用いられている学習記事（1,000 記事），テスト記事（1,000 記事）と同じ記事を用いている．学習記事，訓練記事，テスト記事はそれぞれ初期ルールを学習するための記事，例外処理を行うための記事，分類精度を測定するための記事である．訓練記事は，初期ルール学習後のシステムに順に与えて分類させ，もし分類結果があらかじめ訓練記事に付与されているカテゴリと異なった場合に，否定情報を知識ベースに与えるために用いる．記事の中には複数のカテゴリが付与されているものがあるため，合計 2,000 記事に対し，延べ 2,249 個のカテゴリが付与されている．これらの文書を用いて以下の実験を行った．

［実験 1］すべての学習・訓練記事 1,000 記事を学習記事として初期ルールを生成し，分類精度を測定する．

［実験 2］学習・訓練記事を学習記事 750 記事，訓練記事 250 記事に分割した計 4 個のデータセットを用意し，学習記事から初期ルールを生成し，分類精度を測定する．

［実験 3］実験 2 と同じ学習記事から初期ルールを生成した後に，訓練記事 250 記事を用いて例外処理を行う．その後，分類精度を測定する．

［実験 4］実験 3 の例外処理を文献 16) の手法に変更して分類精度を測定する．この手法では，知識ベースに矛盾が発生した際に知識コンバージョンをせず，3.2 節の c) に該当する分類ルールを削除して誤分類を解消する．

上の実験 1 と文献 17) の結果の比較によって，アルゴリズム 2.1 で生成されるルールの基本的な分類精度を評価し，実験 1 と実験 3，実験 2 と実験 3 の結果の比較によって例外処理機構の有効性を検証する．また実験 3 と実験 4 の結果の比較によって従来手法と本手法との相互評価を施す．各実験では初期ルールの違いによる例外処理の効果の違いを調べるため，アルゴリズム 2.1 で生成されるルールの中で，学習文書における例外文書の割合が 0%，10%，20%，30%，40%以下となるルールで構成される 5 通りのルール集合を生成した．

ルールの生成アルゴリズムにおけるユーザパラメータは次のように設定した．まず，アルゴリズム 2.1 におけるパラメータ b（ルールのボディ中のリテラル数）については，予備実験において $b = 1$〜3 を試したところ，上で述べた例外文書の割合（0%〜40%）による再現率・適合率の値域が広く，かつ両者の平均値が高かった $b = 2$ を採用した．またアルゴリズム 3.1 でのパラメータ t（$S(j', d)$ の閾値）については，予備実験において $t = 0$〜400 を試したが結果に変化がほとんど見られなかったため $t = 200$ とした．

4.3 実験結果と考察

表 4 および図 3〜図 6 に実験結果を示す．図に示すグラフの縦軸は適合率，横軸は再現率を表し，右上に位置する結果ほど分類精度が高いことを示す．実験 1 を除く各実験結果は用意した 4 個のデータセットで得られた結果の平均を示している．表 4 より，例外文書の割合が大きいほど再現率が上昇し，適合率が低下していることが分かる．これは，例外文書の割合の上限を大きくすると，より一般化の度合いの高いルールも初期ルールとして許容されるため初期ルールの数が増加し，知識ベースから導かれる結論が多くなるためである．その結果として再現率が上昇し，同時に，誤分類も増えるため適合率が低下する．

表 4 実験結果
Table 4 Experimental results.

	例外文書の割合				
	0 %	10 %	20 %	30 %	40 %
実験 1	50.8	66.2	77.9	83.7	86.2
	85.8	80.6	68.2	62.6	55.5
実験 2	55.8	67.7	77.2	81.3	84.7
	82.3	77.7	69.3	63.4	57.2
実験 3	55.6	67.2	76.4	80.3	83.5
	83.6	79.9	73.1	68.4	63.9
実験 4	49.6	56.6	63.2	65.4	66.7
	87.8	85.1	81.9	79.7	78.5

上段：再現率，下段：適合率，単位（%）

表 5 実験 1 と文献 17) の結果
Table 5 Results of Exp.1 and results in Ref. 17).

	分類精度	特徴数
実験 1・例外文書の割合 10%以下	73.4	293
実験 1・例外文書の割合 20%以下	73.1	346
実験 1・例外文書の割合 30%以下	73.2	374
文献 17)・単語数 300	71.9	300
文献 17)・単語数 500	74.0	500

分類精度：再現率と適合率の平均値，単位（%）

実験に要した時間は実験 3 の初期ルールの生成で 181 分（750 記事：1 記事あたり約 14.5 秒），例外処理で 95 分（250 記事 $\times 5 = 1,250$ 記事：1 記事あたり約 4.6 秒）であった．

4.3.1 実験 1 と文献 17) の結果の比較

表 5 に実験 1 の実験結果と文献 17) で示されている実験結果を示す．文献 17) の結果は，分類研究でよく用いられるサポートベクトルマシンを用い，単語の選択に相互情報量を用いた場合のものである．表に示すように，実験 1 で例外文書の割合が 10%〜30%以下のルールを用いた場合，再現率と適合率の平均値は 73.1%〜73.4%となっている．このとき分類に用いた特徴数（ルールのボディに現れたリテラルの種類）は表 5 に示すとおりであった．一方，文献 17) で同程度の特徴数を用いた場合の実験結果は，単語数 300 の場合で 71.9%，単語数 500 の場合で 74.0%となっている．したがって，例外文書の割合が 10%〜30%以下のルールを用いるよう設定すれば，提案手法の分類精度は従来手法に匹敵する程度になる．実際に分類システムとして利用する場合，例外文書の割合が 10%〜30%以下のルールを用いることは，現実的な設定としても自然であると考える．

4.3.2 実験 2 と実験 3 の結果の比較

図 3 より実験 3 の結果は実験 2 の結果と比べて適合率が向上しており，提案手法の効果が現れていること

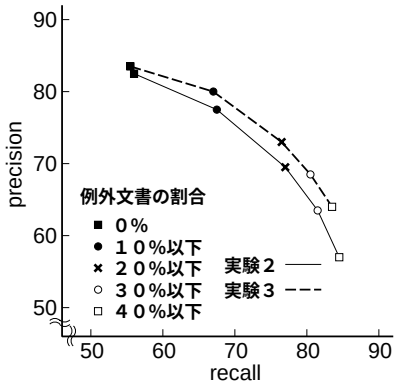

図 3 実験 2，3 の結果（%）
Fig. 3 Results of Exp. 2, 3 (%).

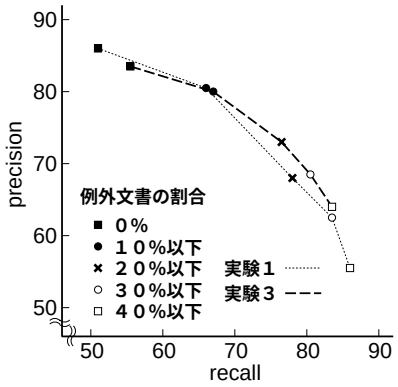

図 4 実験 1，3 の結果（%）
Fig. 4 Results of Exp. 1, 3 (%).

が分かる．例外文書の割合が 0%，10%，20%，30%，40%以下のときの適合率は，実験 3 ではそれぞれ 1.3%，2.2%，3.8%，5.0%，6.7%向上し，適合率が元々低いほどその上昇が大きいことが表 4 より分かる．これは例外文書の割合が大きいほど，例外に関するルールが適用される例外文書が増え，その誤分類が解消されたからである．一方，再現率はそれぞれ 0.2%，0.5%，0.8%，1.0%，1.2%低下しているが，これは提案手法によって生成された例外に関するルールが正解となるべき文書にまで適用された場合があったためである．再現率の低下がわずかにとどまったのは，提案手法では例外に関するルールに対するさらなる例外文書が 1 つでも存在するルールはデフォルトルール化せずに削除しており，例外に関するルールが正解となるべき文書にまで適用されるのを防いでいるためである．このように実験 2 と実験 3 の結果を比較すると，再現率の低下を極力小さくしたうえで適合率を向上することに成功しており，例外処理が有効に働いていることが分かる．

実験 3 では実際に以下のようなデフォルトルール，例外に関するルールが生成された．

デフォルトルール

I) *cat*-教育 $(x) \Leftarrow$ 学校 (x), 高校 (x)

II) *cat*-教育 $(x) \Leftarrow$ 学校 (x), 教師 (x)

例外に関するルール

III) *not-cat*-教育 $(x) \leftarrow$ ドラマ (x), 放送 (x)

IV) *not-cat*-教育 $(x) \leftarrow$ ドラマ (x), 番組 (x)

I)，II) のルールは初期ルールとして生成されたが，訓練記事に「学園ドラマ」に関する，カテゴリ “演劇” に属する例外文書が出現したため，デフォルトルール化された．また同時に例外に関するルールとして III) および IV) が生成された．このように本手法はどの分類ルールがどのように更新・追加されたか直観的に理解しやすいという特徴を持つ．これは本手法が if-then 形式のルールを用いているためであり，重みベクトルの更新等を用いた手法では，分類ルールの変化を直観的に理解することが困難である．

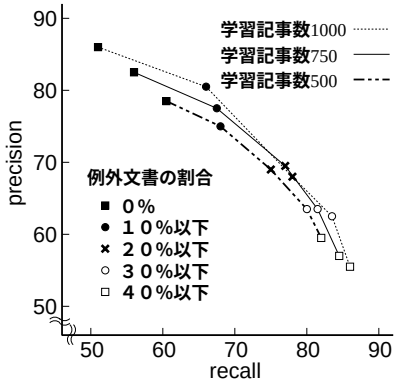

図 5 学習記事数の変化による分類精度の違い（%）
Fig. 5 Difference of classification accuracy (%).

4.3.3 実験 1 と実験 3 の結果の比較

図 4 および表 4 より，実験 3 における例外文書の比率が 20%〜40%以下のとき，実験 3 の結果が実験 1 の結果を上回っていることが分かる．これは実験 3 と比べて実験 1 では学習記事が 250 記事多いにもかかわらず，初期ルールの性能があまり改善されていないためである．ここで，初期ルールの学習記事数を変化させた**図 5** の分類精度を参照されたい．図 5 に示すように，初期ルールの学習による分類精度の向上は，特に例外文書の比率が 20%以下のときには，記事数 750 で限界に達している．これに対して本手法では，例外処理を行うことによって，初期ルールの学習で限界となった精度を向上させている．例外処理に

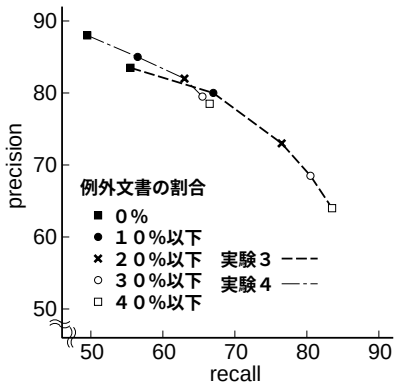

図 6 実験 3，4 の結果（%）
Fig. 6 Results of Exp. 3, 4 (%).

よって否定情報という新たな情報がルールに採り入れられるため，分類精度のさらなる向上が実現できたといえる．これはいい換えるならば，分類規範として重みベクトルを用いている手法において，特徴として用いる単語に正の重みだけを設定する WINNOW でなく，負の重みも設定した BALANCED WINNOW [10] のような規範を用いているととらえることができる．ただし BALANCED WINNOW では各単語（特徴）ごとに重みを設定するため，分類に大きく影響する単語と，あまり影響しない単語を調整できるが，本手法のような if-then 形式のルールを用いた場合，各単語が同等に扱われるため，単語ごとの分類に関する影響が調整できないという問題もある．

4.3.4 実験 3 と実験 4 の結果の比較

図 6 および表 4 に示すように，実験 4 ではいずれの場合も適合率は高いレベルを保っているが，初期ルールにおける例外文書の割合が大きくなるにつれて，修正後の再現率は大幅に下がっている．一方，実験 3 では，実験 4 の結果に比べて，修正後の再現率の低下はわずかである．これらより，誤分類を導いたルールを削除する方法では，適合率に関しては初期ルールのいかんにかかわらず好ましい結果が得られているが，再現率に関しては，初期ルールの再現率が高い場合にはその値は大幅に低下し，多くの正しい分類結果まで得られなくなっていることが分かる．これに対して提案手法では，誤分類を導いたルールをデフォルトルールとして知識ベース内部に保持することにより，再現率が高い場合にもその値をほとんど低下させずに適合率を改善することに成功している．

5. おわりに

文書分類システムにおける誤分類を防ぐための知識ベース修正法として，知識コンバージョン，および例外に関するルールの生成方法を提案し，分類の再現率を低下させることなく適合率を向上させる方法を示した．4 章の実験において示したように，適合率は 1.3〜6.7%向上し，一方で再現率の低下は 0.2〜1.2%にとどまった．このように提案手法は，正しい分類結果が導かれなくなるのを防ぎつつ，誤分類を解消することに成功している．本手法はルールを表現形式とするシステムにおいて一般的に利用できるため，現実世界の問題を対象とするような例外が含まれる様々な知識ベースに対して，知識ベースの拡張にともなう誤りを減らす手法として，その応用が期待できる．

今後の課題としては，1）電子新聞記事以外の文書を用いた場合の例外処理の効果の検証，2）初期ルールの学習方法を変化させた場合の例外処理の効果の検証，3）学習テーブルの更新法の検討，4）例外に関するルールを作成する際の類義語集合の利用や例外に関するルールのデフォルトルールへの変換といった，例外処理法の改良等があげられる．

謝辞 実験で用いた記事をお教えいただいた NTT コミュニケーション科学基礎研究所の平博順氏，毎日新聞 94 年版の記事データの研究利用許諾をいただいた毎日新聞社に感謝いたします．

参考文献

1) Apté, C., Damerau, F. and Weiss, S.M.: Automated Learning of Decision Rules for Text Categorization, *ACM Trans. Inf. Syst.*, Vol.12, No.3, pp.233–251 (1994).
2) 馬場口登：完全/不完全知識を含むデータベースにおける知識獲得，知識科学の展開，大須賀節雄ほか（編），pp.134–142, オーム社 (1996).
3) Cohen, W.W.: Learning to Classify English Text with ILP Methods, *Advances in Inductive Logic Programming*, De Raedt, L. (ed), IOS Press, IOS Frontiers in AI and Applications Series (1995).
4) Cohen, W.W. and Singer, Y.: Context-sensitive learning methods for text categorization, *ACM Trans. Inf. Syst.*, Vol.17, No.2, pp.141–173 (1999).
5) Dagan, I., Karov, Y. and Roth, D.: Mistake-driven learning in text categorization, *Proc. 2nd Conference on Empirical Methods in Natural Language Processing*, pp.55–63 (1997).
6) 情報科学技術協会：国際十進分類表，丸善 (1994).
7) 河合敦夫：意味属性の学習結果に基づく文書自動分類方式，情報処理学会論文誌，Vol.33, No.9,

pp.1112–1122 (1992).
8) 国立国語研究所：分類語彙表，秀英出版 (1964).
9) 小山 誠，桂田浩一，大原剛三，馬場口登，北橋忠宏：文書分類における分類誤りを契機とした例外処理とその実験的評価，人工知能学会研究会資料，SIG-KBS-9903-2, pp.7–12 (2000).
10) Littlestone, N.: Learning Quickly When Irrelevant Attributes Abound: A New Linear-threshold Algorithm, *Machine Learning*, Vol.2, No.4, pp.285–318 (1988).
11) Lewis, D.D., Schapire, R.E., Callan, J.P. and Papka, R.: Training algorithms for linear text classifiers, *ACM SIGIR-96*, pp.298–306 (1996).
12) 毎日新聞社：CD-毎日新聞 94 年版，日外アソシエーツ (1995).
13) 松本裕治，北内 啓，山下達雄，今 一修，今村友明：日本語形態素解析システム『茶筌』version 1.0 使用説明書，奈良先端科学技術大学院大学松本研究室 (1997).
14) 永田昌明，平 博順：テキスト分類 — 学習理論の「見本市」，情報処理，Vol.42, No.1, pp.32–37 (2001).
15) Sebastiani, F.: *Machine Learning in Automated Text Categorization*, Accepted for publication to ACM Computer Survey (2002). http://faure.iei.pi.cnr.it/˜ fabrizio /Publications/ACMCS02.pdf
16) Shapiro, E.Y.: *Algorithmic program debugging*, MIT Press (1983).
17) 平 博順，春野雅彦：Support Vector Machine によるテキスト分類における属性選択，情報処理学会論文誌，Vol.41, No.4, pp.1113–1123 (2000).
18) 豊浦 潤，徳永健伸，井佐原均，岡 隆一：RWC における分類コード付きテキストデータベースの開発，情報処理学会研究報告 NLC96-13, pp.27–32 (1996).
19) Widrow, B. and Stearns, S.D.: *Adaptive Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ (1985).

(平成 13 年 2 月 10 日受付)
(平成 14 年 3 月 14 日採録)

桂田 浩一（正会員）

平成 7 年大阪大学基礎工学部情報工学科卒業．平成 12 年同大学院基礎工学研究科博士後期課程修了．同年豊橋技術科学大学工学研究科助手．博士（工学）．マルチモーダルインタラクション，知識処理に関する研究に従事．人工知能学会，日本音響学会，ヒューマンインタフェース学会各会員．

小山 誠（正会員）

平成 10 年大阪大学基礎工学部情報工学科卒業．平成 12 年同大学院基礎工学研究科博士前期課程修了．同年（株）東芝入社．現在同社研究開発センター，知識メディアラボラトリーに所属．自然言語処理技術，情報検索システムの研究開発に従事．人工知能学会会員．

大原 剛三（正会員）

平成 7 年大阪大学大学院前期課程修了．平成 9 年大阪大学産業科学研究所助手，現在に至る．平成 8 年日本学術振興会特別研究員．博士（工学）．機械学習，例外をともなった知識ベース，非単調推論に関する研究に従事．IEEE, AAAI，電子情報通信学会，人工知能学会各会員．

馬場口 登（正会員）

昭和 54 年大阪大学工学部通信工学科卒業．昭和 56 年同大学院前期課程修了．昭和 57 年愛媛大学工学部助手．大阪大学工学部助手，講師を経て，現在大阪大学産業科学研究所助教授．平成 8〜9 年 UCSD 文部省在外研究員．平成 8 年度人工知能学会全国大会優秀論文賞受賞．工学博士．人工知能，メディア処理の研究に従事．ACM, IEEE，電子情報通信学会，人工知能学会各会員．

北橋 忠宏（正会員）

昭和 37 年大阪大学工学部通信工学科卒業．昭和 43 年同大学院博士課程修了．同年大阪大学基礎工学部助手．同助教授，豊橋技術科学大学助教授，教授を経て，昭和 61 年大阪大学産業科学研究所教授．工学博士．最近は，もっぱらメディア情報処理に関する研究に従事．IEEE, ACM, 電子情報通信学会，人工知能学会，日本認知科学会各会員．