自動分類　確率モデル　識別学習

テキスト自動分類のための半教師あり学習技術

Webページなどのテキストデータを自動分類するための新たな学習技術を紹介します．本技術では，人手で分類されたデータが少数の場合に，大量のカテゴリ未知のデータを同時に学習に利用することで高精度な自動分類を実現します．実際のテキスト自動分類で，本技術は従来技術より高い効果を得られることを示します．

藤野　昭典（ふじの　あきのり）　/上田　修功（うえだ　なおのり）
斉藤　和巳（さいとう　かずみ）

NTTコミュニケーション科学基礎研究所

半教師あり学習とは

Webページや電子メール，各種文書など，コンピュータ上で扱うことができるテキストデータは飛躍的に増えています．膨大なテキストデータを内容や目的などのカテゴリに区分して管理しておけば，これらのデータを効率的に利用することができます．しかし，膨大なテキストデータを人手で分類することは多大な労力を要します．コンピュータを用いてテキストデータを自動的に分類することができれば大変便利になります．

このような自動分類は，例えば，テキストデータに含まれている単語の情報を用いることで実現できます．**図1**に示すように，ある文書を「科学」「音楽」「スポーツ」のいずれかのカテゴリに自動分類する場合を考えます．「科学」のカテゴリに含まれる文書には「速度」や「元素」といった単語を含むものが多く，「スポーツ」のカテゴリに含まれる文書には「ボール」や「競技」といった単語を含むものが多い，というように，文書に含まれる単語の種類は，カテゴリごとに異なっています．すなわち，各カテゴリの特徴は，図1のグラフに示すような単語の出現頻度（単語頻度）で表せます．単語頻度に基づく文書の自動分類とは，例えば，図1の文書1が与えられたとき，文書1の単語頻度にもっとも近い単語頻度をもつカテゴリ（この場合は「科学」）を見つけることに相当します．ただし，各カテゴリの単語頻度は，カテゴリがすでに判明している分類済みの文書を用いて推定（学習）しておきます．

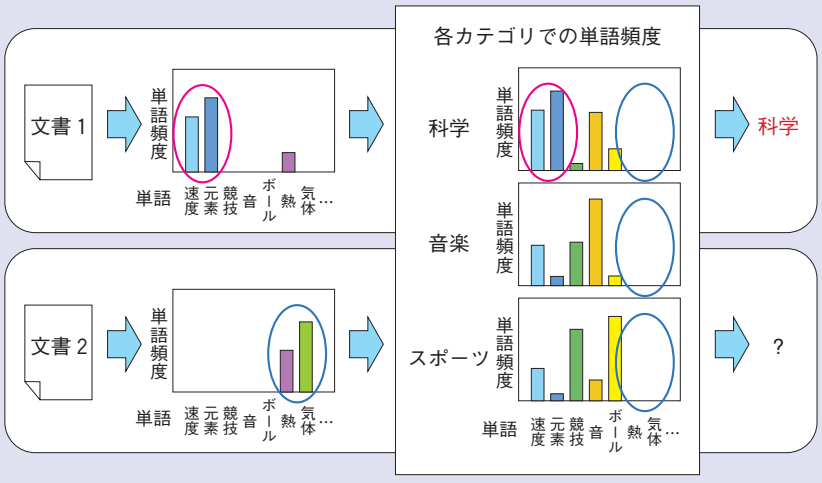


図1　単語頻度に基づく文書の自動分類

一般に，高精度な自動分類を実現するには，多くの分類済みの文書が必要となります．例えば，図1の文書2のように，「熱」や「気体」といった単語を多く含む場合を考えます．分類済みの文書にこれらの単語が含まれないとき，推定されるカテゴリの単語頻度には「熱」や「気体」の情報が含まれないことになります．したがって，文書2のような「熱」や「気体」で特徴づけられる文書を自動分類することが困難となります．この問題を解決するためには，多種多様な単語を含む文書を学習データとして用いる必要があります．

しかし，分類済みの文書の作成は，通常，人手でカテゴリを付与するため時間と労力を要します．一方，カテゴリ未知の文書は，インターネットやデー

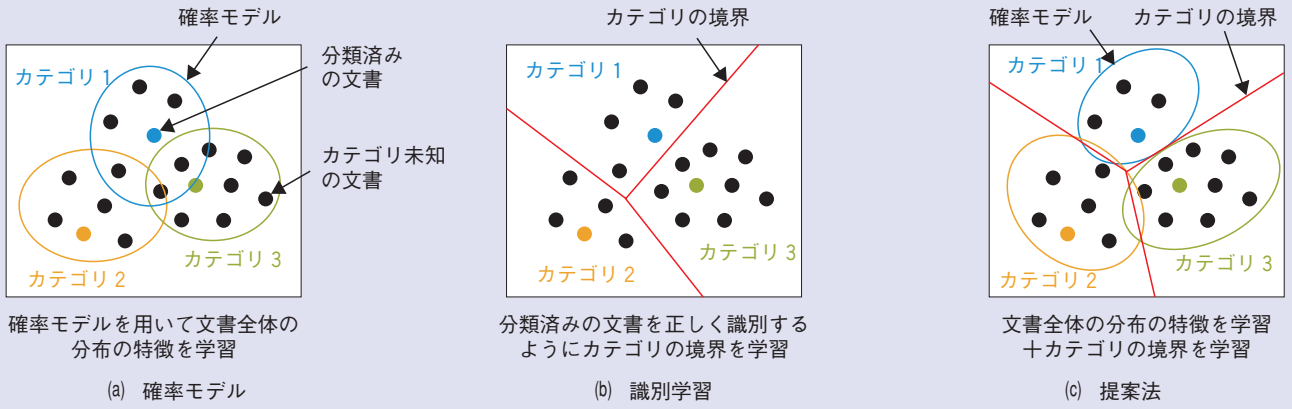


図 2　従来法と提案法の基本アイデア

タベースなどから比較的容易に集めることができます．このような背景から，少数の分類済みの文書に加えて，大量のカテゴリ未知の文書を効果的に用いることにより，自動分類の高精度化を図る「半教師あり学習」という研究が注目されています．

従来法と提案法

従来の半教師あり学習には，確率モデルを用いる方法と，識別学習による方法があります．確率モデルによる方法では，文書の単語頻度分布の確率モデルをカテゴリごとに仮定して学習します．そして，カテゴリ未知の文書に対しては，確率的なカテゴリが割り当てられ，各カテゴリの確率分布の学習に用いられます．

この確率モデルの学習のイメージを**図 2** (a)に示します．図 2 (a)のカラーの点は分類済みの文書，黒の点はカテゴリ未知の文書を表します．楕円は各カテゴリの確率モデルを表し，楕円の中心ほどそのカテゴリへの帰属度が高いことを意味します．図から分かるように，この方法では，カテゴリ未知の文書が与える全体的なデータの分布を反映して各カテゴリの確率モデルが学習されます．しかし，それに伴い，分類済みの文書のカテゴリを誤って識別するような確率モデルを得てしまうという問題が生じます．

一方，識別学習による方法では，カテゴリの境界を直接学習します．図 2 (b)の赤線に示すカテゴリの境界を，分類済みの文書のカテゴリを正確に識別し，かつ，カテゴリ未知の文書をできるだけ分離するように学習します．この方法では，少数の分類済みの文書にカテゴリの境界が過剰に適合し，新規文書の自動分類に悪影響を与える「過学習」の問題が生じます．

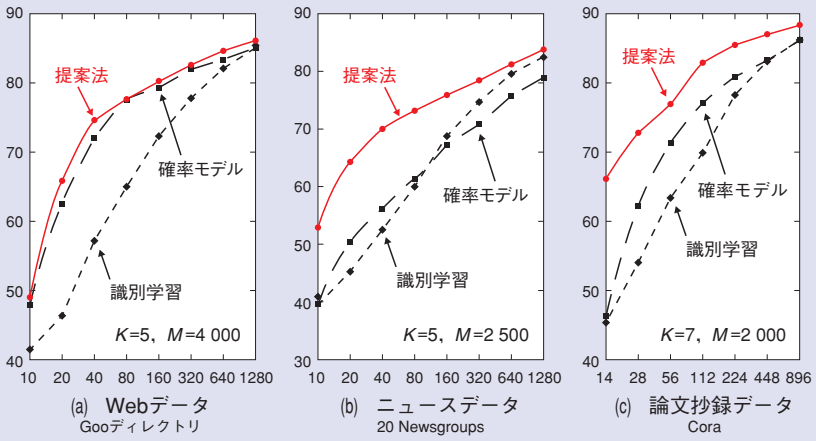
NTTコミュニケーション科学基礎研究所では，従来の確率モデルと識別学習の弱点を相互に補い，分類精度を飛躍的に向上させる新たなハイブリッド法を考案しました[1]．具体的には，図 2 (c)のように，文書全体の分布の特徴を取り込むように学習された確率モデルを利用して識別学習を行うことで，カテゴリの境界が分類済みの文書に過剰に適合するのを抑制します．分類の正確性を与える識別学習と，過学習を抑制する確率モデルの両者を用いる提案法により，新規文書の高精度な自動分類が実現できます．

テキスト分類への応用

提案法による半教師あり学習の効果を確認するために，Webデータ，ニュースデータ，論文抄録データの 3 種類のテキストデータを用いて実験を行いました．Webデータによる実験では，Gooディレクトリ（**http://dir.goo.ne.jp/**）の「ビジネスと経済＞企業＞」配下の 5 つのサブディレクトリに登録されている日本語のポータルサイトを分類対象として用いました．ニュースデータによる実験では 20 Newsgroupsと呼ばれる英文の投稿記事を集めたデータの一部を，論文抄録データによる実験ではCoraと呼ばれる英語論文の抄録と引用情報を集めたデータの一部を分類対象として用いました（使用したカテゴリ数はそれぞれ 5 ，7 ）．

図 3 に，確率モデルと識別学習による従来法と提案法を各々用いて新規データの自動分類を行った場合の精度を比較します．確率モデルによる方法では，文書分類で通常用いられるナイーブベイズモデルを確率モデルとして用い，EMアルゴリズムでモデルの半教師あり学習を行いました．識別学習による方法では，多項ロジスティック回帰モデル[*1]を用いて，最小エントロピー

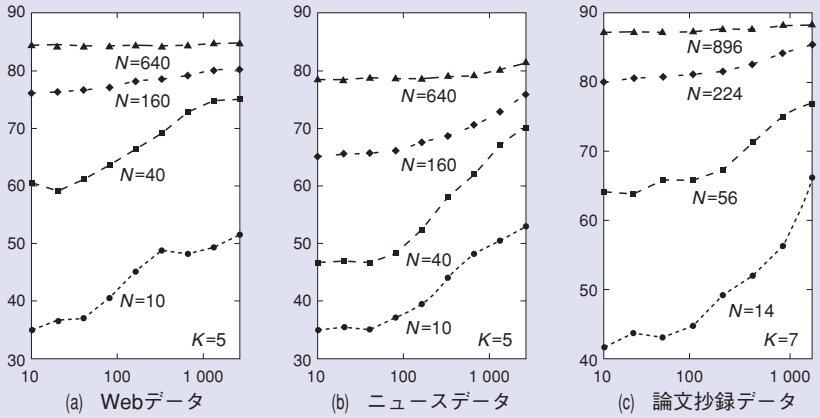
・従来法：
　確率モデル：ナイーブベイズモデル（EMアルゴリズムによる学習）
　識別学習：多項ロジスティック回帰モデル（最小エントロピー正則化による学習）
・グラフの横軸は学習に用いた分類済みのデータの数を，グラフ中のMは学習に用いたカテゴリ未知のデータの数を表す．
・グラフ中のKはカテゴリ数を，グラフの縦軸は新規データの自動分類の精度（%）を表す．



(a) Webデータ（Gooディレクトリ）　(b) ニュースデータ（20 Newsgroups）　(c) 論文抄録データ（Cora）

図3　提案法と従来法による自動分類の精度（%）の比較

・グラフ中のNは学習に用いた分類済みのデータの数を，グラフの横軸は学習に用いたカテゴリ未知のデータの数を表す．
・グラフ中のKはカテゴリ数を，グラフの縦軸は新規データの自動分類の精度（%）を表す．



(a) Webデータ　(b) ニュースデータ　(c) 論文抄録データ

図4　カテゴリ未知のデータを学習に用いる効果

正則化法[*2]により半教師あり学習を行いました．

図3に示した実験結果から，提案法では，従来法より高精度に自動分類できることが確認できます．図3(b)の分類済みデータ数が160の場合のように，確率モデルと識別学習の両手法が類似した性能を示す場合，提案法では両手法より高い精度を得られました．このことから，提案法は両手法の長所を効果的に取り入れた手法であるといえます．

図4は，カテゴリ未知のデータ数を変えたときに提案法で得られる自動分類の精度を調べた結果を表します．図4の結果から，分類済みのデータが少数である場合に特に，大量のカテゴリ未知のデータを学習に用いることで自動分類の精度が向上することが確認できます．

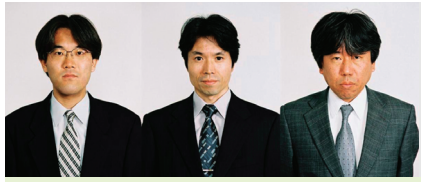
提案法による半教師あり学習技術は，Webページのテキストやリンクなど，データに含まれる異種情報を効果的に用いる自動分類[(2)]や，自然言語処理の固有表現抽出[(3)]などのように自動分類を要素技術として用いるアプリケーションへの応用も可能です．半教師あり学習技術により，さまざまなアプリケーションの高精度化が期待されます．

＊1　多項ロジスティック回帰モデル：3つ以上の値をとる離散変数（テキスト分類ではカテゴリに相当）に対して，多変量データから離散変数の値を予測するのに用いられる回帰モデルの1つ．

＊2　最小エントロピー正則化法：データの分布がカテゴリごとに偏るほど小さくなるエントロピーの特性を利用して，エントロピー最小化によりカテゴリ未知のデータをよく分離させるモデルを獲得する方法．

■参考文献

(1) 藤野・上田・斉藤：“半教師あり学習のための生成・識別ハイブリッド分類器の設計法，”人工知能学会論文誌，Vol.21, No.3, pp.301-309, 2006.
(2) A. Fujino, N. Ueda, and K. Saito: “Semi-supervised learning for multi-component data classification,” Proceedings of the 20th International Joint Conference on Artificial Intelligence, pp. 2754-2759, 2007.
(3) 鈴木・藤野・磯崎：“データの分布特性を利用した半教師有り系列構造学習：言語解析への適用，” 言語処理学会第13回年次大会講演論文集，pp.99-102, 2007.



（左から）藤野　昭典/ 上田　修功/ 斉藤　和巳

半教師あり学習は，次々に生み出される新しいデータに対処するための重要な技術であると考えています．今後は，学習アルゴリズムをさらに高精度化するとともに，さまざまな知識処理タスクへの応用の可能性を探っていきます．

◆問い合わせ先

NTTコミュニケーション科学基礎研究所
協創情報研究部
知識処理研究グループ
TEL　0774-93-5118
FAX　0774-93-5385
E-mail　a.fujino@cslab.kecl.ntt.co.jp