E-11

分野別関連語辞書を用いた新聞記事の自動分類
Automated Categorization of Newspaper Articles using Sectorial Dictionary with Relevant Terms

荒木 淳[†] 中村 文隆[‡] 中山 雅哉[‡]
Jun Araki Fumitaka Nakamura Masaya Nakayama

1. はじめに

大量の電子テキストの蓄積や流通に伴い、これらの大量のテキスト情報の中から有用な情報をいかに効率的に抽出するかが重要な課題となっている。情報を予め分類することは重要な情報へのアクセスを支援する一つの方法と考えることができるので、テキストを予め定められたカテゴリに分類することは上述の課題に対処する一つの手段である。

近年では、Support Vector Machine (SVM) [5] や AdaBoost 等の最先端の機械学習アルゴリズムが次々とテキスト分類に適用された結果、様々な学習理論の実用性を共通のベンチマークに基づいて比較検討することが可能となった [7]。

本稿では、テキスト分類の基本的な問題設定について説明し、次に近年分類精度の高さが注目されている SVM について紹介する。最後に、分野別関連語辞書を導入した手法を提案する。

2. テキスト分類

2.1 テキスト分類の定義

テキスト分類というタスクを予め設定された 2 つ以上のカテゴリに文書を自動的に分類することと定義する。カテゴリが未知のテキストに対して、そのテキストが所属するべきカテゴリをできるだけ正しく予測することがテキスト分類の主目的である。

一般に、テキスト分類では文書を多次元のベクトル空間上に落とし込むことを考える。例えば、(”愛”, ”逆転”, ”国会”, ”ホームラン”) という 4 つの単語の出現の有無を素性 (feature) として、次の 2 つの文書

- 文書 1:「最終回に逆転満塁ホームランが飛び出した。」
- 文書 2 :「国会で与野党の勢力が逆転した。」

をそれぞれベクトル表現すると、文書ベクトル $\mathbf{x}_1 = (0,1,0,1)$, $\mathbf{x}_2 = (0,1,1,0)$ が得られる。文書ベクトルの各要素は、ある単語がその文書に出現するか否かという 2 値の場合もあれば、統計的な手法によって重み付けした数値の場合もある。

多様な文書を高精度で分類するためには、できるだけ多く (数万以上) の素性を使用することが望ましいとされる。しかし、学習を行う際の過学習や計算時間の問題から、多くの分類器は数百から数千程度に素性を削減する必要がある。そこで、単語出現頻度、文書頻度、相互情報量などの様々な評価基準を用いた素性選択法が提案されている [6]。

[†]東京大学大学院工学系研究科
[‡]東京大学情報基盤センター

2.2 Support Vector Machine

Support Vector Machine (SVM) は、近年テキスト分類において機械学習アルゴリズムの中で非常に高い分類能力を持つことが示されている [3, 4]。

図 1 に SVM の概念図を示す。最も負例寄りの正例側の境界面と最も正例寄りの負例側の境界面の間の距離をマージンと呼ぶ。この場合求める超平面は、$\mathbf{w}\cdot\mathbf{x}+b=0$ である。また、$\mathbf{w}\cdot\mathbf{x}+b=\pm 1$ 上の訓練データをサポートベクタ (support vector) と呼ぶ。

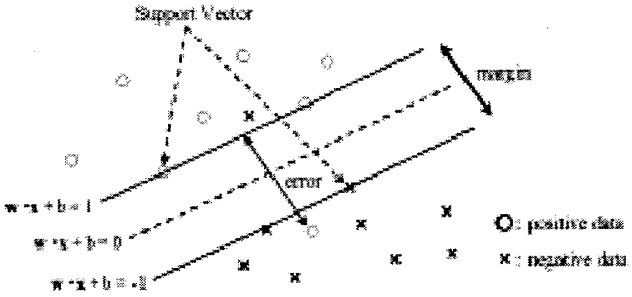

図 1: SVM の概念図

詳しい計算は省略するが、マージンを最大化する問題は 2 次計画問題に帰着でき、これを数値計算で解くことで $\mathbf{w}$ と b を求め、分離境界面を決定する。

3. 分野別関連語辞書の導入

本章では、まず SVM を用いたテキスト分類の実験結果について述べる。次に、テキスト分類における素性選択法に関して問題点を示し、その解決案として分野別関連語辞書を提案する。

3.1 SVM を用いたテキスト分類

分類対象として、Reuters-21578[1] を使用した。このコーパスには、9603 個の訓練記事と 3299 個のテスト記事が含まれ、カテゴリは 135 個ある。この中から、”acq”、”corn”、”crude”、”earn”、”grain”、”interest”、”money-fx”、”trade”、”wheat” の 10 個のカテゴリに含まれる記事について実験を行った。素性選択としてカテゴリとの相互情報量が高い上位単語を用いる選択を行った。

図 2 に素性数を増やしていったときのカテゴリごとの分類精度 (F 値) を示す。F 値は、適合率 P と再現率 R より、

$$F_\beta = \frac{1+\beta^2}{\frac{1}{P}+\beta^2\frac{1}{R}} \quad (1)$$

[1]ロイター通信社による新聞記事コレクション。http://www.research.att.com/~lewis/reuters21578.html から取得できる。

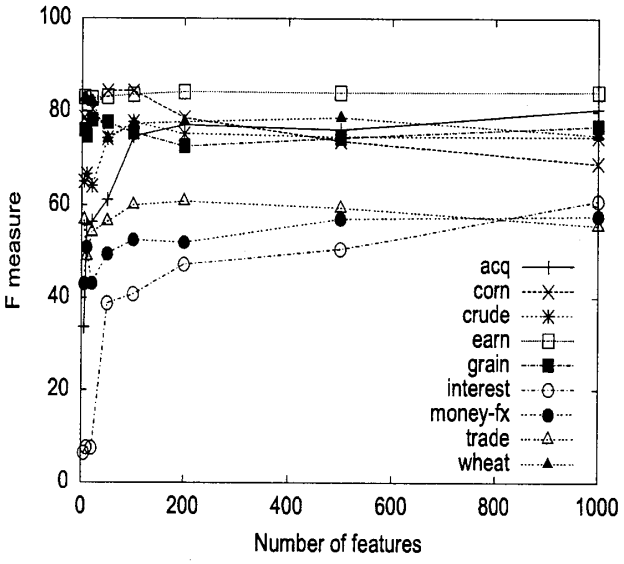

図 2: 素性数と F 値の関係

と表される。βは重み付けのパラメータであるが、今回の実験では$\beta = 1$とした。

この結果から、カテゴリごとに素性数が増加したときの振る舞いが異なるが、カテゴリ平均で見ると素性数が多いほど精度が高くなることが分かる。また、カテゴリによって精度の高さ自体にばらつきはあるが、平均して60〜80の F 値である。このように F 値が低くなる原因として、素性選択時に複数のカテゴリに対して高い相互情報量を示すような単語を素性として選択してしまうために、結果的に異なるカテゴリに属する文書ベクトル間の距離が近いので文書ベクトルをカテゴリ別に明確に分割するような超平面を求めるのは困難になってしまい、こういった状況下で生成した分類器ではテストデータに対する分類精度が下がってしまうのではないかと考えた。

こういった単語としては、例えば"crude"と"grain"の両方のカテゴリに対してともに高い相互情報量を持つ"plant"という単語が考えられる。この場合は、各々のカテゴリに対して情報量の高い「植物」と「工場」という意味を持つ"plant"の両義性が要因である。

relevant words about "fashion"			
accessories	shoes	wristwatch	necklace
Levis	**suit**	bag	sneakers
shirt	trousers	earrings	Nike
...	...	...	...

(a) about the category ,"fashion"

relevant words about "crime"			
police	murder	terrorism	gang
larceny	fraud	attempted	drug
robbery	smuggle	**suit**	victim
...	...	...	...

(b) about the category ,"crime"

図 3: 分野別関連語辞書の例

3.2 分野別関連語辞書とその導入法

分野別関連語辞書は、カテゴリごとに関連性が深いと思われる単語を登録した辞書である。例として「ファッション」と「犯罪」というカテゴリに関する辞書を図 3 に示す。この 2 つのカテゴリにともに関連性が深い単語として"suit[2]"がある。

提案手法では、素性選択においてカテゴリに関連性の深い単語を収集することによって分野別関連語辞書を作成する。カテゴリとの関連性の深さは、カテゴリと単語の相互情報量の高さを指標として想定している。分野別関連語辞書によって、複数のカテゴリに対して関連性の深い単語（複数のカテゴリに対してある閾値以上の相互情報量を有する単語）をフィルタリングし、残った単語を新しい素性として選択する。この新しい素性を用いてベクトル化した文書に SVM を適用する。以上述べてきた提案手法の処理の全体的な流れを図 4 に示す。

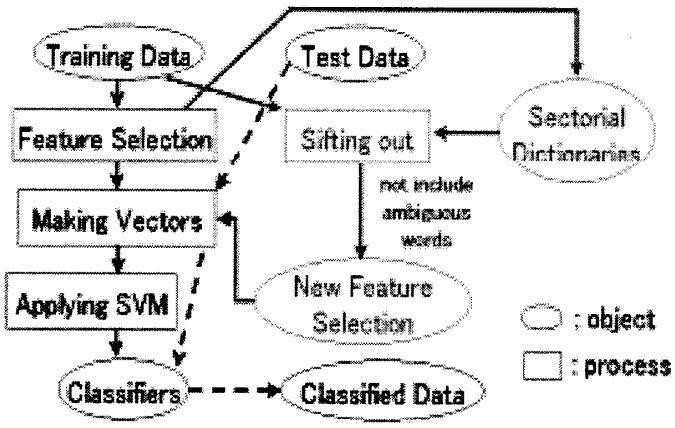

図 4: 提案手法の作業フロー

4. まとめ

本稿では、まずテキスト分類研究の背景、基礎的な問題設定について述べた。次に近年の高い分類精度を誇る SVM を簡単に紹介した。最後に分野別関連語辞書を導入したテキスト分類法を提案した。現在、SVM を用いた標準的なテキスト分類法と分野別関連語辞書を用いた分類法の比較を行い、分野別関連語辞書の効果について実証実験を行っている。

参考文献

[1] Kjersti Aas, Line Eikvil, "Text Categorisation: A Survey", Report No.941, ISBN 82-539-042-B, 1999

[2] Corinna Cortes, Vladimir Vapnik, "Support-Vector Networks", *Machine Learning*, Vol.20, No.3, pp.273-297, 1995

[3] Susan Dumais, John Platt, David Heckerman, "Inductive Learning Algorithm and Representations for Text Categorization", *Proc. 7th International Conference on Information and Knowledge Management*, 1998

[4] Thorsten Joachims, "Text Categorization with Support Vector Machine: Learning with Many Relevant Features", *Proc. 10th European Conference on Machine Learning*, pp.137-142, 1998

[5] Vladmir N. Vapnik, "The Nature of Statistical Learning Theory – Second Edition", Springler-Verlag New York, Inc, 2000

[6] Yiming Yang, Jan O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization", *Proc. 14th International Conference on Machine Learning*, pp.412-420, 1997

[7] 永田昌明, 平博順, "テキスト分類 — 学習理論の「見本市」(特集 情報論的学習理論とその応用) ", 情報処理, Vol.42, No.1, pp.32-37, 2001

[8] 平博順, 向内隆文, 春野雅彦, "Support Vector Machine によるテキスト分類", 情報処理学会研究報告, NL-128, pp.173-180, 1998

[2]ファッションに関連するところでは「スーツ」、犯罪に関連するところでは「訴訟」の意である。