語の出現予測を用いたテキスト分類

岡嶋 穣† 松尾 豊‡ 石塚 満†

東京大学大学院情報理工学系研究科† 東京大学大学院工学系研究科‡

1. はじめに

テキスト分類における最も基本的な素性は、文書中に出現する個々の語である。ひとつの素性がひとつの語を表わし、個々の語がどれだけ出現しているかを手がかりに、文書の属するカテゴリを予測する。さらに、単独の語を素性にするよりも良い素性を得るために、複数の語の情報を組み合わせて新しい素性を作る手法が様々に提案されてきた。

複数の語を組み合わせ素性にする手法の多くは、語の出現の間の共起や相関を手がかりとする。n-gram は連続して出現する複数の語を素性とする。独立成分分析[3]は互いに独立になるように語を組み合わせ素性を生成する。Latent Semantic Indexing[2]は語同士の相関関係を元に素性を生成する。

これらの相関・共起を用いる手法にはいくつかの問題点がある。第一の問題点は、複合する語の選び方である。互いに強く相関する語の組み合わせだけが、カテゴリの予測に役立つ素性であるとは限らない。第二に、語の組み合わせ方が予測の精度を上げるために適切なものとは限らないことである。

これらの従来手法に対し、本論文では、語の出現の予測を手がかりとした複合素性の生成手法を提案する。文書中の各語の出現の有無を正例と負例として、語の出現を予測する二値の分類問題を作成する。これを分類器を用いて学習することにより、ある語の出現を予測するような、それ以外の語からなる予測式を得る。この予測式の正負を新たな素性とみなし、カテゴリの分類の素性として用いる。この手法は、共起や相関関係を前提とせずに素性を複合させることができ、また素性の複合の仕方は語の予測を基準に適切なものとなる。

新聞記事コーパス Reuters-21578 を用いて文書分類実験を行い、語の出現予測に基づく素性が分類精度に寄与することを確かめた。

Text Classification based on Term Occurrence Prediction
† Graduate School of Information Science and Technology, The University of Tokyo
‡School of Engineering, The University of Tokyo

2. 提案手法

基本的なアイディアは、「語の出現を予測することができる素性は、カテゴリの予測にも役に立つ」というものである。本論文では、語の出現の予測を分類器で学習し、その出力式の正負を素性とすることを考える。ある語 x_i が出現する文書を正例とし、出現しない文書を負例とする。この分類を他の語を素性として線形分類器で学習するとき、x_i の有無を予測する次の形の式が得られる。

$$w_{i1} \cdot x_1 + ... w_{i(i-1)} \cdot x_{i-1} + w_{i(i+1)} \cdot x_{i+1} + ... + w_{in} \cdot x_n + b_i > 0$$

このとき x_j は語、w_{ij} はその重み、b_i は閾値である。この式が正となるとき１を取り、負となるとき０となる２値の素性を考える。この素性は、複数の語の出現に基づく複合素性となる。

この素性はその文書に「語 x_i が実際に出現している文書か」ではなく「語 x_i が出現していそうな文書であるか」を表わす。カテゴリ分類をする上では、前者よりも後者のほうがよりカテゴリという概念に接近した重要な情報となることが期待できる。ある文書があるカテゴリに属するとき、必ず同じ語が出るとは限らないが、読者はそのカテゴリに属する語が出現することを予測しながら読み進めるだろうからである。

従来の共起や相関に基づく複合素性と異なり、この素性は共起や相関関係ではなく、語の出現する状況を表わすのに最適な重みと閾値が設定される。語の出現しそうな状況がカテゴリと類似したものならば、この素性はカテゴリ予測においても有用な素性になる。

提案手法の手順は以下の通りになる。

1. 予測する対象となる語 x_i （**被予測語**）を決める。
2. 全ての被予測語 x_i について、

2-1. x_i を予測するための素性となる語（**予測語**）を選択する。

2-2. 学習用の文書を、x_iの有無によって正例と負例に分ける。
2-2. 予測語を素性として各文書を表わし、xの予測を線形学習器で学習する。
2-3. 学習器の出力として、予測語の線形結合による判別式を得る。この式の正負をx_iの複合素性とする

ここで語の予測の学習の評価基準について考える。SVM などの学習器は、予測ができるだけ実測と一致するように、つまりエラー率を最小化するように学習する。しかし、本手法では、必ずしも予測と実測がよく一致するような式を得たいわけではない。実際には語が出現していない文書であっても、語が出現しそうな同じカテゴリに属する文書であれば正になるような素性が、実際の語の観測値よりもカテゴリ予測に役立つと考えられる。そこで本手法では Recall を優先して最大化するように語の予測を学習する。すなわち、語x_iが実際に出現する文書では必ず正になり、x_iは出ていないが類似した文書でも正になるような判別式を学習し素性とする。

3. 実験

Reuters-21578 コーパスを用いてカテゴリ分類実験を行った。ステミングとストップワード処理を行い、３文書以上に登場する語を選び、8127 語を得た。これを基本素性とする。最も用例数の多いトピック１０個を選び、予測するカテゴリとした。

複合素性を表わす判別式を得るのには、Joachims[1]が提案した Recall を評価基準とする SVM を用いた。ランダムに選んだ 4000 文をこの学習の用例とする。被予測語として高頻度語 1000 語を選び、複合素性を 1000 個生成した。このとき予測語としては、それぞれの語と相互情報量が高い語 100 語を用いた。

カテゴリの分類学習には、エラー率を評価基準とする一般的なＳＶＭを用いた。ランダムに選んだ 3000 文について８分割交差検定を行った。単独の語を表わす 8127 個の基本素性と、提案手法による素性を追加した 9127 個の複合素性から、それぞれカテゴリとの相互情報量が高い素性を 100 個選択して学習を行った。この 100 個のうち複合素性を含める数の上限を 10、20、30、上限なし、で変化させて、複合素性の効果を調べた。

表１に実験結果を示す。複合素性数を多くすることで精度が上昇していることが分かる。その一方で、複合素性数の上限を取り払うと、精度が低下する。このことは、複合素性に冗長な素性が多く含まれることを示している。精度改善をより安定なものにするために、冗長な素性の削除や統合を行うことが考えられる。

表１．カテゴリ分類精度と複合素性数

	基本素性	複合素性数			
上限数		10	20	30	上限なし
earn	96.8	96.2	96.3	96.4	95.5
acq	91.0	91.0	91.3	90.8	89.3
Money-fx	64.4	66.1	65.8	67.1	61.7
grain	85.6	85.8	85.8	85.8	90.7
crude	87.9	89.0	89.7	90.2	61.5
trade	67.9	66.4	71.2	69.1	62.0
interest	61.4	59.4	59.4	65.8	61.5
ship	76.5	75.7	75.7	78.0	79.5
wheat	86.9	91.0	91.0	86.5	85.3
corn	78.6	78.6	77.5	78.6	77.1
マクロ平均	79.7	79.9	80.2	80.8	78.6

4. まとめ

語の出現の予測を行うことで、語を複合し新たな素性を生成する手法を提案した。この複合素性が新聞記事のトピック予測に寄与することを確かめた。

今後の課題としては、冗長な素性を削除したり、あるいは冗長な素性同士を複合して素性同士の独立性を高めることなどが考えられる。

参考文献

[1] T. Joachims, “A Support Vector Method for Multivariate Performance Measures”, Proceedings of the International Conference on Machine Learning (ICML), 2005.
[3] Tao Liu, Zheng Chen, Benyu Zhang, Wei-ying Ma, Gongyi Wu, “Improving text classification using local latent semantic indexing” Fourth IEEE International Conference on Data Mining(ICDM) 2004.
[3] X. Sevillano, F. Alías, J. C. Socoró, “Reliability in ICA-based Text Classification” Fifth International Conference, ICA 2004, Granada, Spain (2004) 1213-1220