3B2-4

テキスト分類における 重みつき類似度を用いた SVM 判別モデルの説明

Explaining A Discriminant Model Constructured by SVM in Text Categorization

*1板橋 広和 Hirokazu Itabashi　*2松井 藤五朗 Tohgoroh Matsui　*2大和田 勇人 Hayato Ohwada

*1東京理科大学 大学院 理工学研究科 経営工学専攻
Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science

*2東京理科大学 理工学部 経営工学科
Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science

Support Vector Machine is a very good machine learning system and is used in text categorization widely. However, we cannot understand the reason why it works well, because the discriminant model is represented by a weighted sum of the features of documents, even if we use linear SVM. In this paper, we propose a method for explaining a discriminant model constructured by SVM using weighted similarity in text categorization. We also show the experimental results which indicates that our method works well.

1. はじめに

SVM (Support Vector Machine) によるテキスト分類の研究 [3, 5] は、積極的に行われてきたが、SVM が分類結果を作る過程で得られる判別モデルへの説明はあまりなされてこなかった。SVM などの機械学習は、判別のための数式を示すのみで、その学習の中身を人間が分かるように解釈するのは困難である。このように判別モデルの説明をすることは新たに有益な知見を発見できると期待されるものである。

そこで我々は、SVM の判別モデルを説明し、その判別モデルがどのような事例を根拠に分類するかわかれば、誤った判別モデルを得てしまったときに、その誤りの基となった事例を特定すれば、ユーザーに誤った原因を説明することができると考える。

すなわち、SVM が未知の事例のラベル付けを間違えたとき、SVM の判別モデルを作成するのに用いられた学習事例に原因があるといえる。

例えば、図 1 のような SVM が正事例と負事例の境界を決める超平面を考える。ここで、正事例として学習されているが、正事例としてはあまりふさわしくない学習事例があるとしよう。その場合、SVM の超平面はその学習事例を考慮した分、正しい判別モデルである真のモデルから遠ざかってしまう。その結果、テスト事例のラベル付けを間違えることが起こってしまうと考えられ、誤分類の原因はその学習事例にあるといえる。また、この判別モデルが分類の根拠とした学習事例を知るためには、SVM が未知の事例と学習した事例との近さを示す必要があると考える。

文書間の近さを示す指標として、類似度の研究が広くなされている。近年でも [2, 4] らのように、様々な類似度に関する手法が提案されている。しかし、このような類似度は文書間の類似度を測るために様々な工夫がなされたもので、これらの類似度を用いても、SVM の判別モデルを説明することはできない。

そこで本研究では、SVM が学習に用いた事例間の類似度を調べることによって、SVM の判別モデルを説明する方法を提案する。

連絡先: 板橋広和, 東京理科大学 大学院 理工学研究科 経営工学専攻 大和田研究室, 千葉県野田市山崎 2641, j7408604@ed.noda.tus.ac.jp

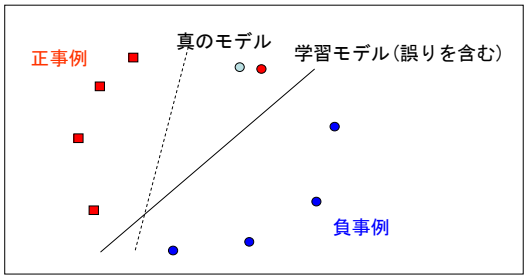

図 1　SVM による超平面学習におけるイメージ

2. 提案手法

本研究では SVM の判別モデルの説明に、分類されたテスト事例に対し、どのような学習事例が近いと示されるか類似度を用いてテスト事例と学習事例の近さを表現する。しかし通常の類似度では SVM の判別モデルを反映できないので、SVM の重みを考慮した重みつき類似度を用いる。本論で述べている SVM の重みとは、SVM の学習から得る判別モデルから取得する重みベクトルのことである。また、本論では特徴語が出現するかしないかに着目し、出現頻度は考慮しない。

2.1 重みつき類似度

本論文では、文書 i, j 間における重みつき類似度 $Sim_{i,j}$ を次のように定義する。

$$Sim_{i,j} = \frac{a_{i,j}}{a_{i,j} + b_{i,j} + c_{i,j}} \quad (1)$$

ここで、$a_{i,j}$ は文書 i, j において共起している特徴語、$b_{i,j}$ は文書 i にのみ出現している特徴語、$c_{i,j}$ は文書 j にのみ出現している特徴語を意味し、それぞれ次のように定義される。n は次元数を示し、α_k は k 番目における SVM の学習によって得られる重み α を示し、$x_{i,k}$ は文書 i における属性番号 k 番目の属性

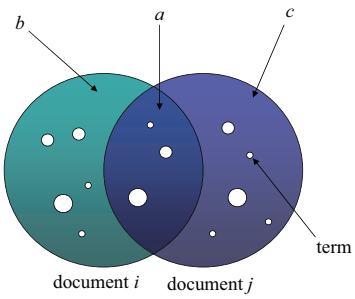

図 2　記事 i, j における a, b, c の関係

値を示す。

$$a_{i,j} = \sum_{k=1}^{n} |\alpha_k| x_{i,k} x_{j,k} \tag{2}$$

$$b_{i,j} = \sum_{k=1}^{n} |\alpha_k| x_{i,k} (1 - x_{j,k}) \tag{3}$$

$$c_{i,j} = \sum_{k=1}^{n} |\alpha_k| (1 - x_{i,k}) x_{j,k} \tag{4}$$

この $Sim_{i,j}$ は重みが全て 1 のとき、つまり $\forall k = 1, 2, \cdots, n$ について $\alpha_k = 1$ のとき、Jaccard 係数の式に一致するという数学的性質を有している。

また、式 (2), (3), (4) の関係を示すと図 2 のようになる。

2.2　アルゴリズム

本提案手法のステップは大きく分けると以下の 5 段階になる。

1. 学習事例を形態素解析して、特徴語リストを生成する。
2. 得られた特徴語リストから文書データをベクトルデータに変換する。
3. SVM による学習を行い、SVM 判別モデルから重みベクトルデータを取得する。
4. テスト事例を形態素解析し、文書データをベクトルデータに変換する。
5. テスト事例と学習事例の重みつき類似度を計算する。

ここでは各ステップを詳細に述べていくことで、本提案手法の計算手順を説明していく。

まずステップ 1 は、用いる文書データのうち学習事例に用いる文書データに、形態素解析ツールにて形態素解析を行う。得られた単語の語幹を特徴語としてソートし、番号付けをして抽出するものである。これにより得られる特徴語数が次元数 n となる。

次にステップ 2 は、ステップ 1 で得られた特徴語とその番号を基にして、文書データを形態素解析する。その際、一致する特徴語があれば同じ特徴語番号をつけ、対応する属性値として 1 を付与する。また一致しなかった特徴語の属性値には 0 を付与し、ベクトルデータ化する。また各文書ベクトルデータの先頭にその文書の正事例、負事例を表すラベルとして $1, -1$ を付与する。

ステップ 3 は、ステップ 2 で得られた文書ベクトルデータに、SVM の学習に用いることで SVM の判別モデルを生成する。その際評価される各特徴語の重み値を特徴語番号の属性値としてベクトルデータ化するものである。これによって重み α が得られる。

ステップ 4 は、テスト事例に用いる文書データを形態素解析し、得られた単語の語幹を特徴語としてソートし、番号付けする。ここでステップ 1 で得られた特徴語とステップ 4 で得られた特徴語を比較し、一致する特徴語のみでテスト事例をベクトルデータ化する。また各文書ベクトルデータの先頭に、その文書のラベルを表すものとして 0 を付与する。

最後にステップ 5 は、得られた学習事例の文書ベクトルデータとテスト事例の文書ベクトルデータ、それに SVM の重みベクトルデータを用いて重みつき類似度を計算する。計算はテスト事例 1 つに対し、それぞれの学習事例との重みつき類似度 $Sim_{i,j}$ を全て計算する。

2.3　重みつき類似度計算の例

ここでは実際に文書から重みつき類似度の計算方法について例を用いて説明する。はじめに次の 1〜3 のような文書があるとする。

1. 明日の天気は晴れでしょう。
2. 今日の天気は晴れのち曇りです。
3. 明日から天気が悪くなるでしょう。

これらを形態素解析によって得られた特徴語を SVM で学習した結果、次のような重みづけがされたとする。

明日 $= 0.3$
天気 $= 0.5$
今日 $= 0.2$
晴れ $= 0.3$
曇り $= 0.1$
悪く $= 0.01$

まず文書 1, 2 の重みつき類似度を考える。このとき、

$$a_{1,2} = \text{天気} + \text{晴れ} = 0.5 + 0.3 = 0.8$$
$$b_{1,2} = \text{明日} = 0.3$$
$$c_{1,2} = \text{今日} + \text{曇り} = 0.2 + 0.1 = 0.3$$

と計算することができる。これより $Sim_{1,2}$ は

$$Sim_{1,2} = \frac{a_{1,2}}{a_{1,2} + b_{1,2} + c_{1,2}} = \frac{4}{7} \tag{5}$$

と求まる。同様にして文書 1, 3 の重みつき類似度を計算してみると、

$$Sim_{1,3} = \frac{0.8}{0.8 + 0.01} = \frac{80}{81} \tag{6}$$

となる。ここで注目すべきことは一見文書 1, 2 のほうが近いと思えるのに、文書 1, 3 のほうが類似度が大きくなっている。これは SVM が学習するときの判断は人間の判断とは異なるということである。例えば文書 3 のように SVM は「悪く」という語の重みを大きく取っていない。つまり、SVM にとっては「悪く」という単語はあまり重要ではないことを意味している。このように、SVM の重みを類似度の計算に利用することで、SVM の学習に基づいた類似度を計算できることが本提案手法の強みである。

また今回図 3 のように、実際にテキストデータを用いて形態素解析を行い、SVM の学習結果から得られる特徴語の重みを用

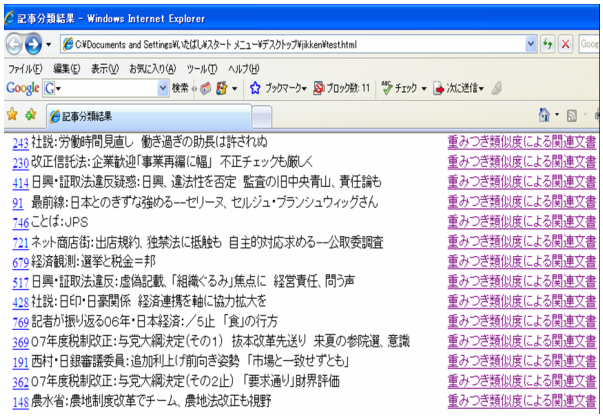

図 3　SVM によって経済と分類されたテスト事例の出力画面

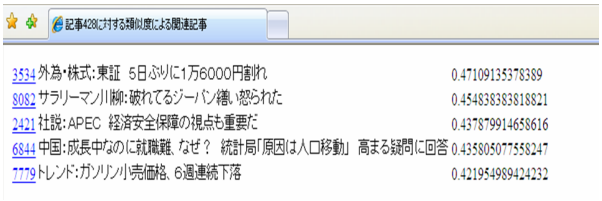

図 4　あるテスト事例に近いと判断された学習事例の出力画面

いて、それぞれのテスト事例に対して SVM が近いと判断した学習事例 5 件を表示した。またその結果を HTML 出力にて表示し、それぞれのリンクをクリックすることで、図 4 の近いと考えられる学習事例を表示するページに遷移するようなシステムを実装した。

3. 実験

3.1 実験データ

本研究では、実験データに 2006 年毎日新聞 [1] のデータを用いた。一般に新聞記事は、堅い内容を多く含み、文語表現が多く、事実だけを伝える記事だと思われがちである。しかし、新聞記事には多くのカテゴリがあり、カテゴリ毎に傾向がある。例えば、経済や国際のカテゴリには文語表現が多く、事実を主に伝える傾向があるが、芸能や家庭のカテゴリでは新聞記事以外の他からのソースの引用が多く、口語表現や意見を多く含む傾向がある。さらに社説やその中にあるみんなの広場といった記者の意見や読者の意見を紹介する記事などもあり、その表現は多彩である。

また新聞記事は公共性が高い刊行物であり、記事の信頼性が高いことも大きなメリットである。掲示板やブログなどの記事は公共性、信頼性は高いとは言えず、データとして用いる際には記事自体を 1 つ 1 つ人手で選別する必要性がある。それに対して、新聞記事をデータとして用いる際には、収集・選別の手間を省くことができる。

今回の実験に用いた新聞記事のカテゴリは「社説」と「国際」と「経済」である。

実験に用いるデータは前処理を行い、ベクトルデータに変換した。またデータの形式はプレゼンス情報のみを扱うものとし、ある特徴語が存在する場合 1, そうでない場合は 0 とした。

表 1　経済と分類された記事

カテゴリ	テスト事例の見出し
社説	社説：労働時間見直し　働き過ぎの助長許されぬ
経済	改正信託法・企業歓迎「事業再編に幅」 不正チェックも厳しく
経済	日興・証取法違反疑惑：日興、違法性を否定 監査の旧中央青山、責任論も
経済	最前線：日本とのきずな強めるーセリーヌ、セルジュ・ブランシュウィッグさん
経済	ことば：JPS
社説	社説：日印・日豪関係　経済連携を軸に協力拡大を
経済	０７年度税制改正：与党大綱決定 (その１) 抜本改革先送り 来夏の参院選、意識
経済	農水省：農地制度改革でチーム、農地改正法も視野

表 2　経済と分類されたある社説記事に近いと判断された学習事例

カテゴリ	学習事例の見出し	*Sim*
経済	外為・株式　５日ぶりに１万６０００円割れ	0.471
経済	サラリーマン川柳：破れてるジーパン縫い怒られた	0.454
社説	社説：APEC 経済安全保障の視点も重要だ	0.437
経済	中国：成長中なのに就職難、なぜ？統計局「原因は人口移動」	0.435
経済	トレンド：ガソリン小売価格、６週連続下落	0.421

3.2 実験方法

提案手法の有効性を確認するため、分類されたテスト事例に近い学習事例を計算するのに、重みを用いない従来の類似度を用いた場合と比較した。SVM のツールには SVM-light を用いた。また SVM-light には学習に用いられた重みを出力するようなオプションがないため、ソースコードを改良し出力できるようにした。実験は、次のように行った。

1. 新聞データから対象とするラベルの記事を抽出。
2. 抽出したデータのうち 11 月分までを学習事例、12 月分をテスト事例とした。
3. 形態素解析後、SVM に学習・分類をさせ、テスト事例を図 3、学習事例を図 4 のように表示させた。
4. テスト事例のうち誤分類されたもので、リンク先に表示される 5 件の学習事例のラベルに注目。
5. 学習事例を K=5 の K 近傍法で、誤分類されたテスト事例を分類し、その精度を求めた。

3.3 実験結果

今回の実験において、一般的なテキスト分類に用いられる重みを用いない従来の類似度と、提案手法の重みつき類似度を比較した。その結果、類似度の高いものを 5 件表示した場合、表示される学習事例の順序関係が変化することがわかった。また結果を HTML 出力することで、それぞれのテスト事例がどういった学習事例を近いと考えているのか表示することができた。実際にどのような文書があったのかを表 1 に示す。また、表 1 の 1 行目のテスト事例に対して近いと表示された学習事例を表 2 に示す。

次に具体例として、社説と経済カテゴリを用いて分類した際に、誤って経済の事例を社説に分類してしまったテスト事例の一部を示す。

三菱自・欠陥隠し：横浜簡裁判決（要旨）　三菱自動車製大型車のタイヤ脱落事故で、横浜簡裁が１３日、同社元幹部らを無罪とした判決理由の要旨は次の通り。　◆認定事実の概要　１　（中略）　２　国土交通省は異常な事故と認識し、同１１日、自動車交通局技術安全部審査課リコール対策室の係長らが、三菱ふそう品質統括部に、発生原因、再発防止策などの報告を求めた (中略)
◆無罪の理由　◇法の規定　改正前の道路運送車両法は「同法が規定する虚偽の報告をした者」を「２０万円以下の罰金に処する」と規定している。　また（報告の前提となる）報告要求については国交相がするものとし、質問検査についてはその職員にさせるものとして、明確に書き分けていることから、報告要求が国交相によるものとしてなされなければならないことは当然である。
従って、虚偽報告で被告らを罰するには、その前提として、国交相からの同法に基づく特定の報告要求が存在しなければならない。　(中略)　同室長らが国交相の代理として報告要求を行った（授権行為）事実も認められない。さらに、事実上行政官庁の補助機関が、その名において行う専決があったとも証拠上認められない。(一部抜粋)

上記の文書は社説カテゴリのものである。次にこの記事に対して提案手法の重みつき類似度によって計算した結果、近いと示された学習事例の一部を下に示す。また、二つの間の *Sim* の値は 0.4 であった。これは今回の実験では類似度の高いものでも 0.5〜0.6 程度であったため、比較的高い方に入る。

社説：パロマ給湯器　安全装置の劣化もあったとは　「製品自体に問題はない」と強調していたメーカーが一転して、製品の劣化による事故も発生していたことを認め、謝罪した。ガス瞬間湯沸かし器による相次ぐ一酸化炭素中毒事故が発覚した「パロマ工業」と親会社「パロマ」のことだ。　パロマ側は１８日の記者会見で、１９８５?０５年に起きた事故が、これまで判明していた１７件、死者１５人からその後の調査で増え、２７件、２０人だったと発表した。　このうち、湯沸かし器に取り付けられている安全装置の端子を針金でつなぐなどした「不正改造」によって、安全装置が正常に作動せず事故に至ったとみられるのは１４件。４件は機器そのものの経年劣化によって安全装置が作動しなくなったケースだったという。　その４日前の会見では、パロマ側が把握している事故原因はすべて「不正改造」で、その改造を誰が行ったかは「分からない」と力説し、責任を否定していた。(中略) 事実確認は後手に回った。(中略) 行政が事故情報を防止策に生かせなかった問題も併せ、安全対策をさまざまな角度から検討してほしい。(一部抜粋)

また実験によって得られた精度を図 5 に示す。今回、SVM が事例判断に誤ったときにどんな学習事例が近くにあったのか示すために、SVM が事例判断を誤った際、SVM と同じように事例判断を誤っている割合について調べた。その結果、図 5 のように、提案手法のほうが従来手法に比べて高いことがわかった。また、今回一番間違えた事例数が多かったのは国際と経済カテゴリを分類したときであった。逆に社説と国際カテゴリを用いて分類した場合、分類の間違いは少なかった。

4. 考察

実験の結果、提案手法による重みつき類似度の計算と従来の類似度による計算の間には表示される学習事例の順序関係等が変化することがわかった。このことより、重みつき類似度を導入することに意味があることが確認された。また、今回の結果からは従来の場合と比較して、提案手法による場合のほうが精度が高くなっていることから、重みつき類似度のほうが SVM の判別モデルに近い分類をできていることがわかる。

実験結果に挙げた記事は上が社説に誤分類された経済記事の一つ、下はそのときに類似度の近い 5 件から表示された学習事

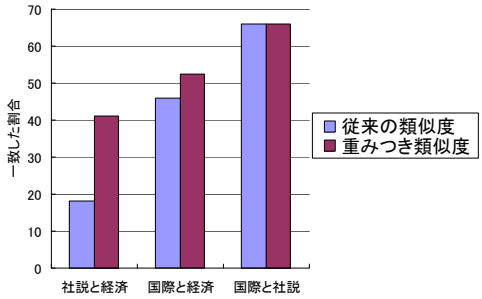

図 5　SVM の求めたラベルと類似度によって求めたラベルの一致率

例の一つで社説のカテゴリである。二つを比較すると、上は不祥事問題における判決文で、下は別件の同問題に関する記事である。内容から両者は共に関係の深い記事であるといえる。このため、SVM は誤ってこの経済記事を社説記事に分類してしまったと解釈できる。このように本手法を用いて、類似した学習事例を示すことで、テスト事例が誤分類される原因を説明できた。

また国際と社説カテゴリの分類では誤分類が少なかったことより、社説カテゴリと国際カテゴリなどの、主観的文書と客観的文書のテキスト分類は、SVM には簡単にできるようである。逆に国際カテゴリと経済カテゴリの分類では多くの特徴語がかぶっているため、SVM の判別モデルによる分類が難しかったものと考えられる。

5. まとめ

実験結果から得られたように、重みつき類似度を導入することで従来の重みを用いない類似度と比較して、社説と経済のカテゴリでは、従来の類似度に対して約 2 倍、国際と経済のカテゴリでは約 13 %重みつき類似度のほうが SVM の判別モデルに基づいた説明ができていることが確認された。さらに、SVM の学習事例とテスト事例を結びつけることで、そのテスト事例の分類の根拠を学習事例によって表現することができた。これより本提案手法の有効性が確認された。今後は、この提案手法が SVM 以外の機械学習においても有効かどうか検証する必要があると考えられる。また、SVM は訓練データのラベル付けを人手で行わなければならず、それが用意できるデータに制限をかけてしまう。そのため、あらかじめラベルが決まっていないデータの場合に適用する方法なども検討する必要があると思われる。

参考文献

[1] 毎日新聞 CD-ROM(2006).
[2] Hung Chim. A new suffix tree similarity measure for document clustering. *IW3C2*, pp. 121–129, 2007.
[3] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. *In Proc. of the 10th European Conference on Machine Learning*, pp. 137–142, 1998.
[4] Li-Wei Lee and Shyi-Ming Chen. New methods for text categorization based on a new feature selection method and a new similarity measure between documents. *IEA/AIE 2006*.
[5] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, Vol. 34, p. 147, 2002.