F-004

重みつき類似度を用いたテキスト分類における判別モデルの改善

Improving a Discriminant Model Using Weighted Similarity in Text Categorization

板橋 広和† 松井 藤五郎‡ 大和田 勇人‡
Hirokazu Itabashi Tohgoroh Matsui Hayato Ohwada

1 はじめに

近年のブログ利用者は 860 万人を超えている。ブログには最新の話題や個人の主観的な意見が多く含まれることから、これらの情報を収集・分析できれば意思決定支援や企業リスク管理に利用できると考えられるため、我々は自動的にブログ記事を分類するシステム BREVIS を開発している [4]。BREVIS は、EM アルゴリズムを用いて、ごく少数のラベルつきのブログ記事からラベルが未知の新しく収集されたブログ記事のラベルを推定し、そのラベル付き事例から次に収集されるブログ記事を分類するための分類器を学習し直す。しかしながら、これらの学習事例に付けられた推定ラベルには誤りが含まれてしまう点が大きな問題となっている。

このような問題に対して、神嶌らは、飼いならしと呼ばれる新しい学習問題を提案した [3]。これは、学習の訓練事例に飼育データと野生データと呼ばれる 2 種類の事例が混在している状況における学習問題である。飼育データは目標概念と無矛盾なラベルが注意深く選ばれているものである。野生データのラベルは、厳密には管理されておらず、目標概念に一致してるものもあればそうでないものもふくまれていて、完全には信用できない。

飼いならし問題の解決法として、神嶌らは、BaggTaming と呼ばれる手法を提案した。BaggTaming は、バギング [1] のアイデアを応用し、分類器を飼育データから学習するかわりに、野生データからサンプリングを行って複数の弱分類器を学習し、それらの投票によって分類を行う。しかし、学習に用いられたデータのラベルが不正確であるため、学習される弱分類器の判別モデルには誤りが含まれている。そこで、BaggTaming は、学習した弱分類器を飼育データを用いて評価し、優良な弱分類器を選定することによって分類器の精度を向上させている。

ここで、学習事例の誤ったラベルを修正するか、誤った学習事例を削除することができれば、個々の弱分類器の精度を向上させ、分類器全体の精度を改善することができると考えられる。我々は、判別モデルに基づいた重みつき類似度によってラベルが誤っている学習事例を特定する手法を提案している [2]。

そこで、本論文では、重みつき類似度を用いてラベルが誤っている学習事例を特定し、その事例を削除あるいはそのラベルを修正することによって判別モデルの精度を改善する手法を提案する。また、ブログ記事を用いた実験によりその有効性を示す。

2 提案手法

本研究では、神嶌らの BaggTaming において、学習事例から得られる個々の学習器を重みつき類似度を用いて改善することを提案する。しかし通常の類似度では機械学習の判別モデルを反映できないので、機械学習の重みを考慮した重みつき類似度 [2] を用いる。

本提案手法のステップは大きく分けると以下の 4 段階になる。

1. 学習事例を用いて学習を行い、判別モデル D を得る
2. 得られた判別モデル D を用いて、テスト事例のラベル予測を行う
3. 予測した結果、ラベルが誤っていた場合、重みつき類似度を用いて、ラベルの異なるテスト事例周辺の学習事例を調べる
4. 重みつき類似度によって特定された誤った学習事例を削除あるいはそのラベルを修正し、再学習する

ここでは各ステップを詳細に述べていくことで、本提案手法の流れを説明していく。

まずステップ 1 にて、学習事例から機械学習を用いて学習し、判別モデル D を得る。また、このとき機械学習の重み α も得られる。

次にステップ 2 は、得られた D を用いて、機械学習にてテスト事例のラベルを予測させる。

ステップ 3 では、ステップ 2 で予測した結果、誤ったラベルのついたテスト事例に近いと考えられる学習事例を見つけるため、ステップ 1 で得られた重み α を用いて重みつき類似度 Sim を計算する。

そしてステップ 4 で、重みつき類似度 Sim によって特定された誤ったラベルをつけていると考えられる学習事例の削除あるいはラベルの修正を行い、もう一度学習して新たな判別

†東京理科大学大学院理工学研究科経営工学専攻
‡東京理科大学理工学部経営工学科

モデル D' を得る。

3 実験

3.1 実験データ

実験にはブログ記事を用いた。今回、個人が書くブログを正事例とし、それ以外を負事例として分類を行った。また全ブログのうち 1 日をテスト事例として正しいラベルがわかっているものとし、学習事例として 1 日分のデータを用意した。

3.2 実験方法

提案手法の有効性を確認するために行った実験について述べる。今回機械学習には SVM(サポートベクターマシン) と NB(ナイーブベイズ) を用いた。SVM のツールには SVM-light を、また NB のツールとして Weka 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. 2 日間のブログのうち半分を学習データとし、残りをテスト事例とする
2. 学習事例を機械学習を用いて学習し、判別モデルを得る
3. 得られた判別モデルを用いて、テスト事例のラベル予測を行う
4. 予測した結果誤ったラベルが得られた場合、重みつき類似度を用いてラベルを誤ったテスト事例周辺の学習事例を調べる
5. 重みつき類似度によって特定された誤った学習事例を削除あるいはそのラベルを修正する
6. 修正された学習事例を用いて再学習を行い、もう一度予測をする
7. 以上のステップを学習事例とテスト事例を入れ替えてもう一度行う

また、今回図 1, 2 のように重みつき類似度は、事例を 1 つずつ変化させていったときの変化を見られるように閾値で調整した。

3.3 実験結果および考察

図 1, 2 は学習事例とテスト事例を入れ替えて 2 回行ったときの平均値をとったグラフである。図からわかるように、それぞれある程度事例を削除あるいはそのラベルを修正すると改善が見られ、やり過ぎると逆に悪くなってしまうことがわかる。

4 まとめ

本論文では、重みつき類似度を用いてラベルが誤っている学習事例を特定し、その事例を削除あるいはそのラベルを修正することによって判別モデルの精度を改善する手法を提案した。またブログ記事を用いた実験により SVM については 10 %前後の改善が見られた。これにより本提案手法の有効性が確認された。

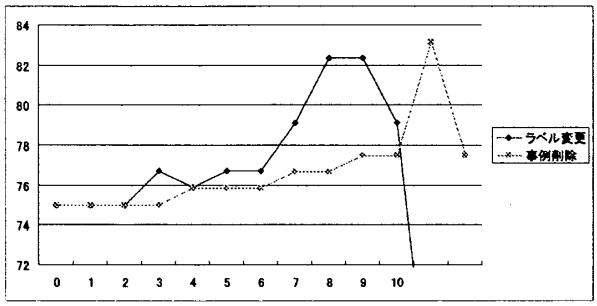

図 1 SVM 学習によって得られる判別モデルを改善した場合

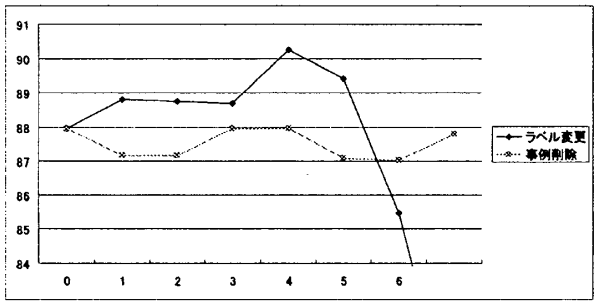

図 2 NB 学習によって得られる判別モデルを改善した場合

参考文献

[1] L. Breiman. Bagging predictors. *Machine Learning*, Vol. 24, pp. 123–140, 1996.

[2] 板橋広和. 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. 第 22 回人工知能学会全国大会.

[3] 神嶌敏弘. 飼いならし-飼育・野生混在データからの学習. 第 22 回人工知能学会全国大会.

[4] 森田悠基. Brevis:ブログにおける評判情報自動収集・検索システム. 第 22 回人工知能学会全国大会.