分類木の類似性を利用した商品コメント評価

三原 隆義[1]　塩飽　朝美[2]　小林 伸行[3]　椎名 広光[4]
[1,2] 岡山理科大学大学院 総合情報研究科 情報科学専攻
[2] 山陽学園大学 総合人間学部 生活心理学科
[3] 岡山理科大学 総合情報学部 情報科学科
i14im03mt@ous.jp[1], i15im01sa@ous.ac.jp[2], koba_nob@sguc.ac.jp[3],
shiina@mis.ous.ac.jp[4]

1　はじめに

ショッピングサイトには一億二千万件以上の商品レビューが投稿されており，興味のある商品の評価を知ることができる．これらの商品レビューコメントとその評価には関係があるように考えられ，コメントから評価を予測できるのではと考えられる．

従来の商品コメントの評価を分類には，コメントに対応する単語ベクトルを作成して，その距離を利用して分類することが多く用いられる．しかし，コメントのような単語数が少ない場合は，単語の頻度の揺れが大きくコメントの評価に影響を及ぼしてしまう問題点がある．

そこでコメントのように短い文章を分類するためには，単語だけではなく，単語の背景を表すような分類木のような二次的な情報を利用することで，構造を持った情報量を追加することで，コメントの分類を行うことを目的としている．

コメントに対応する分類木間の類似性は，木の距離を利用し，コメント間の距離のみを利用する．そのためコメント間の位置は，多次元尺度法 [1, 2] によって，コメント間の位置関係を決めることとしている．

また，コメント間の位置関係を決定した後については，従来の機械学習による分類では，SVM[3, 4] のようにカーネル関数によりデータの分布を仮定するものと，k-NN 法 [5] のようなデータの分布を仮定しない分類器が知られている．それに対して，クラスごとのデータのパラメータと似た値を取りやすいと考えられ，パラメータの値の変更が少ないパラメータが同じクラスのデータとして現れる確率が高いと考えられる．そこで，パラメータが取る空間において，現れたデータのパラメータの位置を中心に同じクラスのデータが現れる確率が高いとして，それに正規密度関数を近似する機械学習法を提案する．

2　レビューコメントからの評価手順概要

本研究のシステムでは商品レビューコメントから 5 段階評価の予測にはまず，商品レビューコメントを分類木に対応させる．次に対応付けした木間の類似度を利用し特徴空間上に各コメントをプロットする．実際の分類は特徴空間上にプロットされた座標値を特徴ベクトルとして従来の SVM による分類手法や提案手法にて分類を行う方法をとっている．商品レビューコメントから分類木への変換，分類木間の類似度の計算，類似度から特徴空間上へのプロット方法の概要を説明する．

(1) 複数のコメントをそれぞれ形態素解析器（MeCab）を用いて形態素解析を行い固有名詞のみを抜き出す．固有名詞のみとなったコメントからナイーブベイズにより Wikipedia のカテゴリの階層構造を木とみなした最下層の葉にあたるカテゴリ 73,766 件のいずれかに分類を行う．また，分類の際の教師データは葉にあたるカテゴリ 73,766 件の固有名詞のみを抜き出したものを利用している．この手法により分類された記事から根にあたる主要カテゴリまでを辿ってそのコメントの木とする．

(2) 木に変換されたコメントを木に対する編集距離である Tree Edit Distance[6] を用いて非類似度を計算する．

(3)(2) で求めたコメント間の非類似度を多次元尺度法を用いて特徴空間上の座標値を決定する．求めた座標値を特徴ベクトルとし，従来の分類手法や後記の提案手法によりレビューコメントから 5 段階評価への分類を行う．

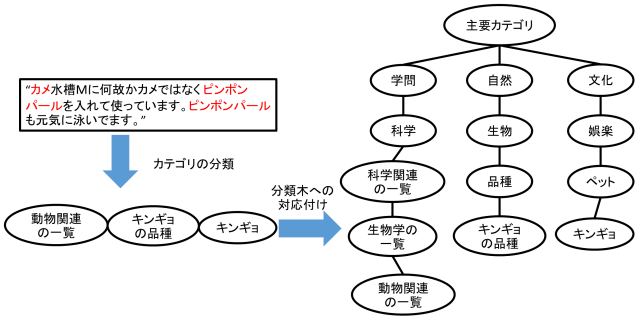

図 1: 商品レビューコメントから分類木への変換例

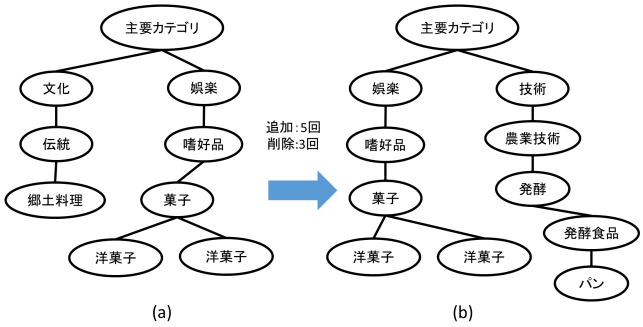

図 2: 分類木間の距離

3　レビューコメントの分類木の作成

Wikipedia のカテゴリー階層を用いて、レビューコメントから分類木を作成する．レビューコメントが Wikipedia のどのカテゴリーに属するのかは，レビューコメントから固有名詞のみを抜き出しナイーブベイズによるカテゴリ分類を用いて上位 3 件を取得し，Wikipedia のルートのカテゴリーから分類されたカテゴリーまでに到達できるリストを併合した木を，レビューコメントに対応した分類木としている．実際の商品レビューコメントから分類木の変換例を図 1 に示す．例のコメントでは“動物関連の一覧”，“キンギョの品種”，“キンギョ”のカテゴリに分類された．

4　分類木間の距離

コメントに対応した分類木から，コメント間の距離を求める．距離については，編集距離や部分木の一致率を求める方法があるが，本研究では，編集距離である Tree Edit Distance[6] を利用している．例として図 2(a) の分類木と図 2(b) の分類木の距離は追加，削除，置換のコストを全て 1 としたとき 5 回の追加，3 回の削除を行うため 8 となる．

5　多次元尺度法の適用

多次元尺度法 [1, 2] は計量多次元尺度法と非計量多次元尺度法に分けられ，計量多次元尺度法は距離データを利用して位置を決める方法で，非計量多次元尺度法は順序尺度のデータの類似度を利用して位置を決める．本研究では，両方の方法を利用して実験評価を行っている．

6　コメントデータに周辺分布の近似

本研究では，クラスごとのデータのパラメータは似た値を取りやすいという仮定から，パラメータが取る空間において，現れたデータのパラメータの位置を中心に同じクラスのデータが現れる確率が高いとして，それに正規分布の密度関数を近似する機械学習法を提案する．また，提案手法は学習データのクラスへの所属確率，分類判定，パラメータ推定の 3 つのステップからなる．学習データのクラスへの所属確率では学習データごとにある幅内の同クラスの学習データからそのクラスが周囲にどの程度生起しやすいかを求める．分類判定ではテストデータのクラスを近傍にある学習データから求める．パラメータ推定では最急降下法を用いて分類判定時に使用する 2 つのパラメータの準最適解を求める．

6.1　学習データの所属クラスへの所属確率

特徴量 $\mathbf{x} = (x^1, x^2, \cdots, x^d)^T$ とクラスラベル $\mathbf{y}$ とし，学習データ $\mathbf{D}$=$(\mathbf{x_1}, y_1)$, $(\mathbf{x_2}, y_2)$, …, $(\mathbf{x_n}, y_n)$ として，提案アルゴリズムを示す．

Step1: 学習データ $\mathbf{x}_i$ と各成分の最大値，最小値と，同じクラスラベル $\mathbf{y}$ の学習データ数 $n_{\mathbf{w}}$ から求めるビン数 $l = \lceil \log_2 n_y + 1 \rceil$ からビン幅 $\mathbf{w}$ を求める．ビン幅 $\mathbf{w}$ の各成分 w_i は $w_i = (\max(d_i) - min(d_i)/l$ で求める．学習データ を中心として $x_i \pm w_i/2$ の範囲内に 1 つでも成分が存在する学習データを近傍データ $V_w(\mathbf{x}_i) = \{\mathbf{x}_1, \cdots, \mathbf{x}_k\}$ とする．近傍データ $V_w(\mathbf{x}_i)$ から正規分布の確率密度関数の平均 $\hat{\mu}(\mathbf{x}_i)$ ，分散共分散行列 $\hat{\mathbf{\Sigma}}(\mathbf{x}_i)$ は，最尤推定法を利用して次式で求める．

$$\hat{\mu}(\mathbf{x}_i) = \frac{1}{k+1}\left(\sum_{j=1}^{k} \mathbf{x}_j + \mathbf{x}_i\right),$$

$$\hat{\mathbf{\Sigma}}(\mathbf{x}_i) = \frac{1}{k+1}\left(\sum_{j=1}^{k} (\mathbf{x}_j - \hat{\mu}(\mathbf{x}))(\mathbf{x}_j - \hat{\mu}(\mathbf{x}))^T + (\mathbf{x}_i - \hat{\mu}(\mathbf{x}_i))(\mathbf{x}_i - \hat{\mu}(\mathbf{x}_i))^T\right)$$

Step 2: Step1 を学習データ集合 $\mathbf{D}$ のすべての学習データ $\mathbf{x}_i$ ごとに繰り返す.

6.2 分類判定

Step1: クラスラベル y ごとにテストデータ $\mathbf{d}$ の所属確率 $P_y(\mathbf{d}) = 0$ とし，閾値 β に対して $\sum_y P_y(\mathbf{d}) < \beta$ である限り，Step2 と 3 を繰り返す.
Step 2: テストデータの特徴量 $\mathbf{d}$ の近傍にある学習データ のクラスラベル y の所属確率をクラスごとに加算する．所属確率は，周囲への影響度を表す変数 を導入して，6.1 の所属確率で求めた分散共分散行列 $\hat{\mathbf{\Sigma}}(\mathbf{x}_i)$ に α を乗じた $\alpha\hat{\mathbf{\Sigma}}(\mathbf{x}_i)$ を新たな分散共分散行列として，正規分布の確率密度関数 $p_y\left(\mathbf{x}, \hat{\mu}(\mathbf{x}_i), \alpha\hat{\mathbf{\Sigma}}(\mathbf{x}_i)\right)$ を次式で求める.

$$p_y\left(\mathbf{x}, \hat{\mu}(\mathbf{x}_i), \alpha\hat{\mathbf{\Sigma}}(\mathbf{x}_i)\right) = \frac{1}{2\pi^{\frac{n}{2}}\left|\alpha\hat{\mathbf{\Sigma}}(\mathbf{x}_i)\right|^{\frac{1}{2}}} \exp\left(-\frac{(\mathbf{x}-\hat{\mu}(\mathbf{x}_i))^T\left(\alpha\hat{\mathbf{\Sigma}}(\mathbf{x}_i)\right)^{-1}(\mathbf{x}-\hat{\mu}(\mathbf{x}_i))}{2}\right)$$

Step 3: テストデータ $\mathbf{d}$ のクラス y の所属確率 $P_y(\mathbf{d})$ に，近傍の学習データ $\mathbf{x}_i$ からみたテストデータ $\mathbf{d}$ の所属確率 $p_y(\mathbf{d}, \hat{\mu}(\mathbf{x}_i), \alpha\hat{\mathbf{\Sigma}}(\mathbf{x}_i)$ を加える.

$$P_y(\mathbf{d}) = P_y(\mathbf{d}) + p_y\left(\mathbf{d}, \hat{\mu}(\mathbf{x}_i), \alpha\hat{\mathbf{\Sigma}}(\mathbf{x}_i)\right)$$

Step 4: 判定は所属確率が最大となったものをそのデータの所属クラスとする.

$$argmax_y P_y(\mathbf{d})$$

6.3 パラメータ推定

テストセットに対して，最急降下法によって，影響度 α と閾値 β の最適値を求め，テストデータについては，ここで求めた α と β を用いた 6.2 の分類判定を行う.

6.4 ベンチマークテスト

多次元尺度法を組み合わせた提案手法を評価する精度実験を行った．データセットは UCI Machine Learning Repository[7] から Pima Indians Diabetes(Pima), Heart，Iris の 3 つを用いた．また，SVM[3, 4] との精度の比較として Machine Learning and Data Mining Group[8] の LIBSVM ライブラリからは線形カーネルと RBF カーネル，同じく LIBLINEAR ライブラリからは線形カーネルを使用した精度も記載する．提案手法については最急降下法により影響度 α と閾値 β の 2 種類のパラメータの推定を行っている．そのため初期値は 5 組の乱数を利用し，一番精度が高かったものを採用した．提案手法と SVM による精度実験の結果を表 1 に示す．また，提案手法は最急降下法によってパラメータ最適化を行っているため選んだ初期値によっては局所解になり良い精度が得られないことがあり，高い精度を実現するためには適切なパラメータを設定する必要がある．影響度 α と閾値 β に対して得られる精度の等高線を図 3 に示す.

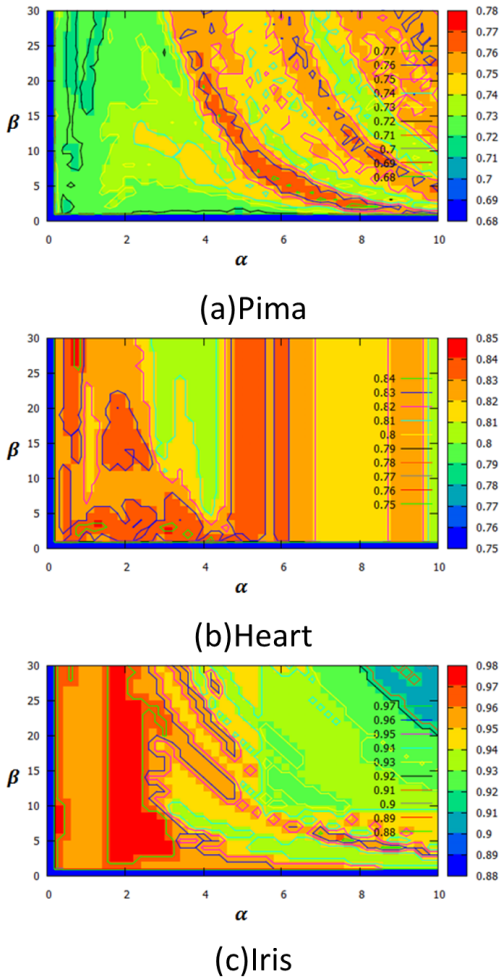

図 3: パラメータ変化による精度等高線 (

表 1: 一般のデータセットによる精度実験

Data	提案手法	LIBSVM		LIBLINEAR
		Linear	RBF	Linear
Pima	77.08%	77.86%	76.56%	68.75%
Heart	84.44%	82.22%	84.44%	82.96%
Iris	97.33%	96.0%	93.33%	89.33%

表 2: 商品レビューコメントの評価分類実験

手法		次元数	20	100	200	300	400	500	600
LIBSVB	Linear	ユークリッド距離	40.48%	45.0%	44.0%	42.6%	43.72%	43.88%	45.28%
		Tree Edit Distance	19.0%	20.56%	20.92%	21.08%	21.28%	21.32%	21.16%
	RBF	ユークリッド距離	40.4%	44.8%	44.52%	44.28%	45.4%	46.4%	46.52%
		Tree Edit Distance	20.16%	21.4%	21.04%	21.2%	21.88%	20.96%	21.24%
LIBLINEAR	Linear	ユークリッド距離	40.56%	44.44%	43.92%	42.16%	42.24%	43.88%	44.04%
		Tree Edit Distance	18.88%	20.84%	21.0%	23.28%	21.48%	21.44%	21.16%
提案手法		ユークリッド距離	33.4%	-	-	-	-	-	-
		Tree Edit Distance	21.68%	-	-	-	-	-	-

7 商品レビューデータを用いた評価分類の精度実験

商品レビューデータを用いて精度を調べる実験を行った．商品レビューは 1〜5 の整数値の五段階評価とその評価のコメントとなっている．評価をそのままクラスとした 5 クラス分類の場合を実験した．教師データとテストデータは各クラス 500 件の計 2500 件ずつとした．また，分類木に変換を行う手法の評価のためにコメントの単語の頻度情報からベクトルを作成する手法も同時に評価した．方法は単語のカイ二乗値が高い上位 2400 単語のみを抜き出し，TF-IDF 値で重み付けを行いユークリッド距離により非類似度を計算する．その後，同様に多次元尺度法により特徴空間上にプロットする．分類器はベンチマークテストと同様に提案手法と SVM の場合で評価をした．多次元尺度法によりプロットする特徴空間の次元数を変化させた場合の結果を表 2 に示す．

8 おわりに

商品レビューコメントと評価には関連があるように考えられることから，レビューコメントからその商品の 5 段階評価を行うシステムの作成を行った．文書分類タスクでは一般に単語の頻度情報から特徴ベクトルを作成するが，本研究ではコメントなど短い文章での単語の頻度の揺れを考慮し，一度分類木に対応付ける手法で分類を行った．分類器については近辺のデータの影響に正規分布の確率密度関数を想定する機械学習法を提案し，精度実験では SVM 以上の精度を実現したが商品レビューコメントの分類では SVM 以下の精度となった．また，本研究で提案した，分類木へ対応付ける手法は従来の単語頻度を利用した手法より低い精度となった．問題として使用する品詞や対応付けに利用する分類木等の考慮が必要である．

参考文献

[1] Edwards,J. and Oman,P: Dimensional Reduction for Data Mapping-A practical guide using R, R News, Vol. 3/3,2-7. 2003.

[2] 田口, 大野, 横山: 非計量多次元尺度構成法への期待と新しい視点, 統計数理, 49(1), 133–153, 2001.

[3] Vapnik,V.N., Statistical Learning Theory, Wiley, 1998.

[4] C. M. Bishop,C.M., Pattern Recognition and Machine Learning, Springer (2006)

[5] Shakhnarovish,G. , Darrell,T., Piotr Lindyk: Nearest-Neighbor Methods in Learning and Vision : Theory and Practice (Neural Information Processing series), The MIT Press, 2006.

[6] Kaizhong Zhang, Dennis Shasha, Simple fast algorithms for the editing distance between trees and related problems, SIAM Journal of Computing, 18:1245-1262, 1989.

[7] UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/

[8] Machine Learning and Data Mining Group: http://www.csie.ntu.edu.tw/ cjlin/mlgroup/