データ周辺の生起確率推定による機械学習による コメントの評価クラス分類

三原 隆義[1]　小林 伸行[2]　椎名 広光[3]　北川 文夫[4]
[1] 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
[3,4] 岡山理科大学 総合情報学部 情報科学科
i14im03mt@ous.jp[1], koba_nob@sguc.ac.jp[2], shiina@mis.ous.ac.jp[3], kitagawa@mis.ous.ac.jp[4]

1 はじめに

機械学習による分類にはSVM[1, 2]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 [3, 4] 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，それに正規分布の密度関数を近似する機械学習法を提案する．

また，近年，Googleインスタント検索，Amazonのおすすめ商品お知らせサービスなどのユーザーの行動を予測するサービスが登場している．一方，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いて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．そこで本研究では，商品コメントの評価分類に提案する機械学習手法を用いて効果について考察する．

2 レビューコメントからの評価手順概要

はじめに本研究のシステムでは商品レビューコメントの単語頻度などから特徴ベクトルを作成する方法をとっている．特徴ベクトルとは形態素解析などを利用し単語の出現順は考えずに単語の出現頻度などによって文章をベクトルで表現したものである．特徴ベクトルを利用した本研究のシステムの概要を説明する．

(1) 複数のコメントをそれぞれ単語の出現頻度やTF-IDF値 [5] を使用して特徴ベクトルに変換する．ここでは形態素解析器（MeCab）や係り受け解析器（CaboCha）の結果を利用して単語や単語の共起を素性とした．また，精度の向上のために情報利得，相互情報量やカイ二乗値の数値を使用して素性選択を行う．

(2) 特徴ベクトルを行列に変換するが処理速度の向上やノイズ除去のために潜在的意味インデキシング（LSI[6]）による次元圧縮を行う．

(3) 行列に変換したコメントを教師データとしてSVMや提案するデータの周辺における生起確率の推定による分類器を使用して機械学習を行う．

(4) 評価を予測させたいコメントを入力する．

3 データの周辺における生起確率の推定による機械学習

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，データが密集している場合は，少ない数の相違するクラスのデータの影響も大きくなり，近傍の同じクラスのデータが多い場合は，その付近により多く同じクラスのデータが現れやすいとして，現れやすくなる増加分を近傍する k 個のデータを利用して補完分を計算し，精度の改善も行う．最後に提案手法では，同じクラスのデータの分布の分散，データ間の

影響を補完する近傍の個数，正規分布による近辺の影響と補完される重みの 3 つのパラメータがあり，パラメータ最適化に最急降下法を用いて準最適解によるパラメータ推定を行っている．

3.1 学習データの生起確率を正規分布で近似する

特徴ベクトル $\mathbf{x}$，特徴ベクトルに対するクラス分類 (教師信号) を s として，提案手法のアルゴリズムを次に述べる．

Step1: 学習データ集合 $\mathbf{D}$ からデータ $\mathbf{d}$ を取り出し，クラス分類 s ごとに，その特徴ベクトルが所属するセルを中心に正規分布の確率を加算する．本研究では，学習データ $\mathbf{d}$ に対する特徴ベクトル $\mathbf{x}$ の生起確率に $p(\mathbf{x}, \mathbf{d})$ を分散共分散行列 Σ の共分散を 0 で，分散 σ^2 が等しいとして次式で定義する．

$$p(\mathbf{x}, \mathbf{d}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} exp\Big(-\frac{(\mathbf{x}-\mathbf{d})^{\mathrm{T}}(\mathbf{x}-\mathbf{d})}{2\sigma^2}\Big) \quad (1)$$

セルの中心の座標を代表とする特徴ベクトルを $\mathbf{c_x}$ として，クラス分類 s ごとの特徴空間のセル全体の生起確率 $N(s, \mathbf{c_x})$ とすると，次式で表す加算をセル全体に行う．

$$N(s, \mathbf{c_x}) = N(s, \mathbf{c_x}) + p(\mathbf{c_x}, \mathbf{d}) \quad (2)$$

Step2: Step1 を学習データ集合 $\mathbf{D}$ のすべてのデータ $\mathbf{d}$ ごとに繰り返す．

Step3: 学習データのクラス分類 s ごとの特徴空間のセル全体の生起確率の合計が 1.0 となるように正規化を行う．

$$N(s, \mathbf{c_x}) = \frac{N(s, \mathbf{c_x})}{\sum_{\mathbf{c_y}} N(s, \mathbf{c_y})} \quad (3)$$

3.2 学習データ間の影響を補完

学習データの生起確率を正規分布で近似による方法のみではデータが密集していても他のクラス分類のデータが存在する近傍はその影響を受けてしまう．そこで近傍にあるデータ間の影響を正規分布で補完するとした．方法としては，ある学習データから近傍にある同じクラス分類の学習データとの中点を正規分布の平均の位置とした．学習アルゴリズムは学習データ $\mathbf{d}$ に対する同じクラス分類の k 個の近傍点集合 $\mathbf{V_d}$ ，

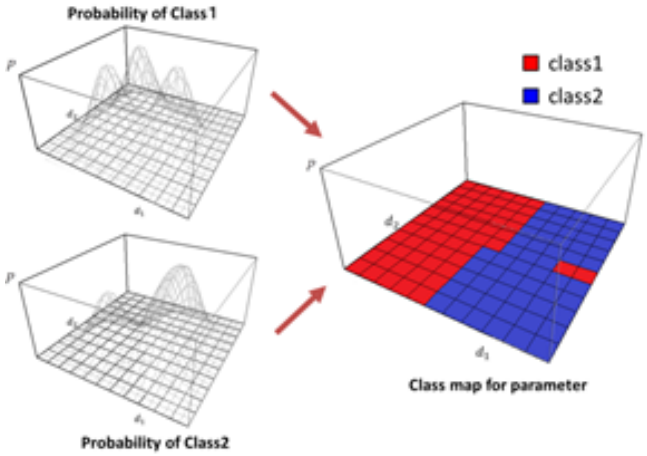

図 1: クラスマップ

重み w が与えられたとすると，セル全体の生起確率 $N(s, \mathbf{c_x})$ の計算式を次式のように変更する．

$$\begin{aligned} N(s, \mathbf{c_x}) =& N(s, \mathbf{c_x}) + (1-w) \cdot p(\mathbf{c_x}, \mathbf{d}) \\ &+ w \cdot \sum_{v \in \mathbf{V_d}} p(\mathbf{c_x}, \frac{\mathbf{d}+\mathbf{v}}{2}) \end{aligned} \quad (4)$$

また,3.1 の Step3 式 (3) と同様に正規化を行う．

3.3 分類判定

学習データのクラス分類 s ごとに，特徴空間の全体にデータが生起する確率 $N(s, \mathbf{c_x})$ を学習アルゴリズムで求めているので，未知のデータ $\mathbf{d}$ のクラス $C(\mathbf{d})$ の判定は，データ $\mathbf{d}$ の特徴ベクトルの所属するセル $C(\mathbf{d})$ での生起確率が最も高いクラス分類とする．例えば，2 パラメータの 2 クラスの判定をセルごとに色分けした場合，図 1 のようにマップが作ることができる．

$$C(\mathbf{d}) = argmax_s(N(s, \mathbf{c_d})) \quad (5)$$

3.4 パラメータ最適化

高い精度でクラス分類するためには適切なパラメータを設定する必要がある．最急降下法により精度が最大となるパラメータ探索を行った．最急降下法により求めるパラメータは，分散，近傍数，補完される正規分布の重みの 3 種類とした．初期値については分散を 1.0，近傍数を 0，補完される正規分布の重み 0.6 とした．評価実験としては，UCI Machine Learning Repository[7] の Pima Indians Diabetes 他のデータセットを用いた (本稿では Pima Indians Diabates の結果のみを提示)．また，提案手法の比較対象として

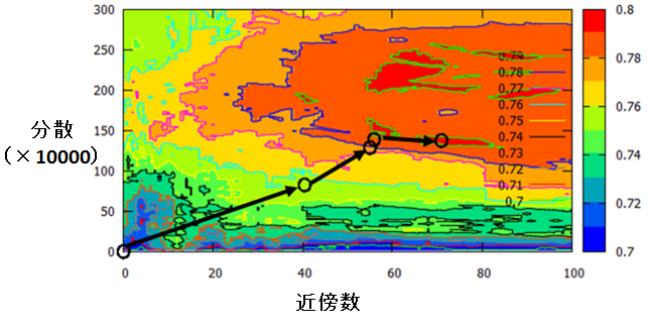

図 2: 分散と近傍数のパラメータごとの精度等高線 (Pima Indians Diabetes)

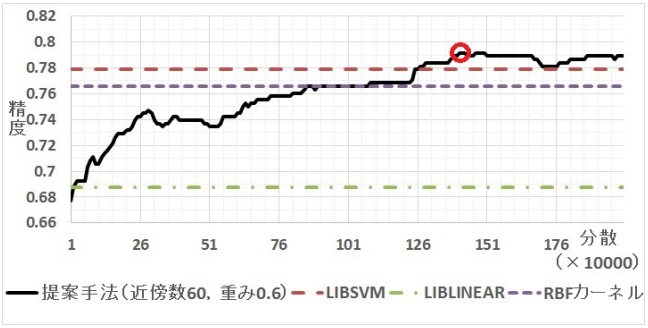

図 3: 分散を変化させたときの精度比較 (Pima Indians Diabetes)

表 1: 全単語ベクトルによる精度

	頻度	TF-IDF
線形カーネル	42.01%	40.61%
RBF カーネル	21.12%	37.00%

SVM のうち線形カーネルと RBF カーネルを利用した．SVM のシステムとしては，Machine Learning and Data Mining Group[8] の LIBSVM と LIBLINEAR の線形カーネルと，LIBSVM の RBF カーネルを利用した． なお，特徴空間の各軸は 100 等分割している．

最急降下法によるパラメータ推定については，最急降下法の反復は，4 回で収束した．パラメータに対する精度の等高線表示上に更新されるパラメータをプロットしたものを図 2 に示す．また，その時の精度は 79.16 %で SVM 以上の精度となっている．また，精度に大きく影響しない補完される正規分布の重みを 0.6 に固定し，分散を変化させた場合おけるパラメータの位置と精度を図 3 に示す．

4 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

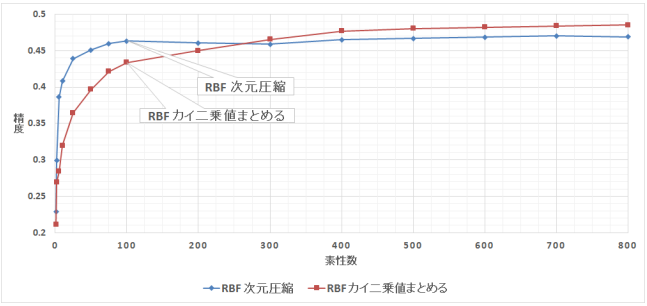

図 4: 素性数の変化による精度

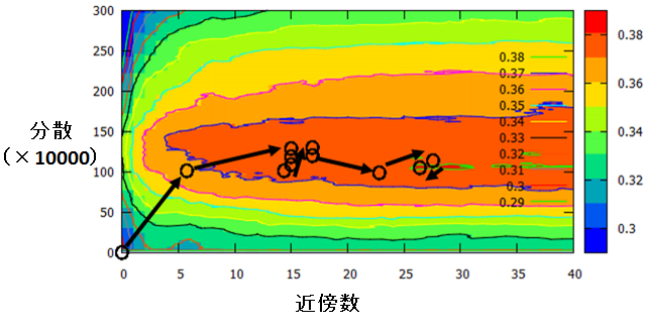

図 5: 分散と近傍数のパラメータごとの精度等高線 (コメント)

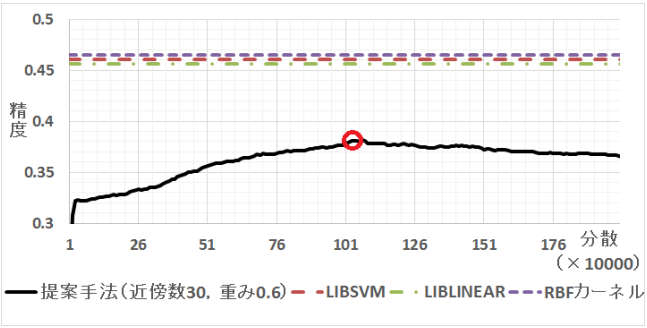

図 6: 分散を変化させたときの精度比較 (コメント)

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．また，教師データとテストデータは各クラス 3000 件の計 15000 件のデータを各クラス半分に分けて計 7500 件ずつとした．

(1) 全素性を利用した場合

出現する単語すべてを利用して作成した．素性は頻度を利用したものと単語の重要度を示す TF-IDF 値を利用したものの二種類で実験した．また，分類器には SVM を利用し，LIBSVM の線形カーネルと RBF カーネルを利用した．その実験の結果を表 1 に示す．単語の頻度より TF-IDF 値，RBF カーネルより線形カーネルの方が良い結果が得られた．

(2) 素性数を削減した場合

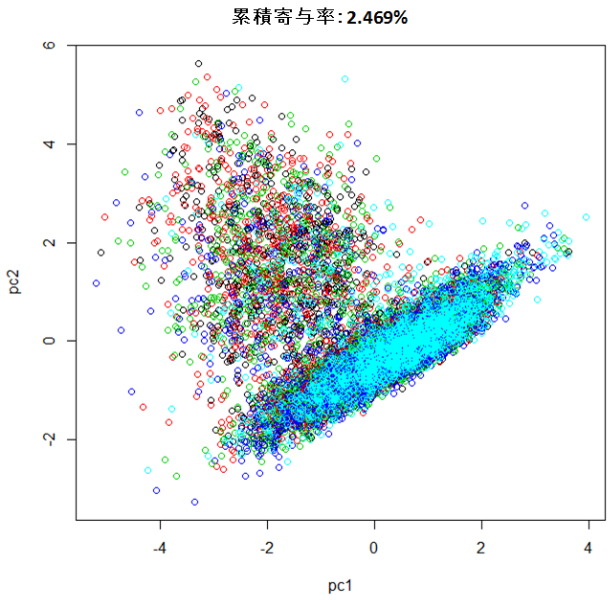

図 7: コメントデータの分布

少ない素性数で良い精度を実現するため LSI による次元圧縮を行った場合を実験した．素性は精度の良い類義語をまとめたカイ二乗値の数値の高い上位 2100 個の単語を使用した．また，分類器には SVM の場合で RBF カーネルを使いグリッドサーチにより最適なパラメータを設定している．素性数が 0〜800 までの精度を図 4 に示す．この結果から素性数 75 程度から安定した精度が得られることが分かった．また，次元圧縮を行わない場合より精度の改善が見られた．

(3) データの周辺の生起確率による手法

本研究で提案するデータの周辺の生起確率による手法をコメントデータの評価分類に実施した．最急降下法によるパラメータの推定は，10 回の反復で収束した．パラメータに対する精度の等高線表示上に更新されるパラメータをプロットしたものを図 5 に示す．また，分散を変化させた場合おけるパラメータの位置と精度を図 6 に示す．最適なパラメータを得ることはできたが，SVM の精度には及ばなかった．原因として提案手法ではセルで分割した場合に各クラスが入り乱れて存在するセルがあるためと考えられる．コメントデータに対して主成分分析を行い第一主成分軸と第二主成分軸のみを使い 2 次元に縮約した場合のデータ分布を図 7 に示す．

5 おわりに

本研究では商品レビューコメントと 5 段階評価は関係があると考えられることから，機械学習により評価の予測を行うシステムを作成した．また，新たに同じクラスの特徴ベクトルは似たパラメータを取りやすいという仮定から，生起確率をデータを中心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た．精度実験では比較対象に SVM を使用し，一般のデータセットを用いた場合は SVM 以上の精度を実現できた．また，パラメータ最適化の実験も同様に SVM 以上の精度を実現するパラメータを得ることができた．商品レビューコメントを用いた場合の精度実験では SVM には及ばなかったが，パラメータ最適化の実験では最適なパラメータを得ることができた．今後の課題としては, パラメータ最適化では初期値によっては最適解に到達できないため精度の改善に限界があること．また，機械学習のベンチマークで使われるデータに比べ，商品レビューコメントのように素性選択から問題になったり，データが大量にある実際のデータ場合には，今のところ良い精度が得られておらず機械学習での生起確率の計算に工夫をする必要がある．

参考文献

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

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

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

[4] Duda, R. O., Hart, P. E., Stork, D.G.: Pattern Classification, Wiley Inter-Science, New York, 2000.

[5] 北, 津田, 獅々堀: “情報検索アルゴリズム”, 共立出版, 2002.

[6] S.T.Dumais, T.K.Landauer, G.W.Furnas and R.A.Harshman: Indexing by latent semantic analysis, Journal of the Society for Information Science, 41(6), 391–401, 1990.

[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/