分類スコアに基づいたクラス事後確率の推定

高橋 和子[†]　　高村 大也[††]　　奥村 学[††]
[†] 敬愛大学 国際学部
[†] 〒 285-8567 佐倉市山王 1-9
[††] 東京工業大学 精密工学研究所
[††] 〒 226-8503 横浜市緑区長津田町 4259
[†]takak@u-keiai.ac.jp, [††]{takamura,oku}@pi.titech.ac.jp

本稿では, クラス事後確率の推定を行うために, 分類器の出力するスコア（分類スコア）を複数利用することおよび, 正解率の平滑化に移動平均法などを利用することを提案する. 分類器は複数のクラスに対する分類スコアを出力するが, 提案手法では, 予測されたクラスの分類スコアだけでなくそれ以外のクラスの分類スコアも利用する. 職業データによる実験の結果, すべての手法において, 分類スコアを 2 位まで利用したときに最もよい結果が得られた. また, 手法別では, 利用する分類スコアの数に関係なく, 分類スコアを等間隔に区切り区間ごとに求めた正解率を, 移動平均法により平滑化を行って作成した正解率表を利用して, 間接的にクラス事後確率を推定する方法が最もよかった. 2 番目には, 分類スコアを独立変数とするロジスティック回帰分析により直接推定する方法がよく, ラプラス法による平滑化の方法が最も悪かった.

Tranforming Scores from a Classifier into Posterior Probability Estimates

Kazuko Takahashi[†]　　Hiroya Takamura[††]　　Manabu Okumura[††]
[†] Faculty of International Studies, Keiai University
[†] 1-9　Sanno Sakura, JAPAN, 285-8567
[††] Precision and Intelligence Laboratory, Tokyo Institute of Technology
[††] 4259 Nagatsuta Midori-ku Yokohama, JAPAN, 226-8503

We propose methods for estimating the posterior probability of the predicted class, using classification scores not only for the predicted class, but also for other classes. In the proposed methods, we first make an accuracy table by counting the number of correctly classified training examples in each range or cell of classification scores. We then apply several different smoothing methods to the accuracy table, such as moving average method, Laplace method. We empirically showed that the use of multiple classification scores is effective in the estimation of posterior probability, and that the smoothing methods for the accuracy table including the moving average method work quite well in this task.

1　はじめに

本研究の目的は, ある事例が分類器によりクラス（分類カテゴリ）を決定されるとき, それがどの程度確からしいかというクラス事後確率を推定することである.

与えられた事例に対して分類器が出力するクラスの事後確率を推定することは, さまざまな意思決定の場において非常に有用である (Platt, 1999).

例えば, ある事例のクラスを人手で決定する際, 自動分類システムからの情報を参考にすることができる場合がある. このとき, 単にその事例の候補となるクラスが提示されるだけでなく, どの程度そのクラスらしいかという確率も付与されれば, 判断を行いやすくなる. 実際, われわれは, 社会調査において必須の作業である職業データの分類（職業コーディング[1] ）を担当するコーダを支援するために, 回答（職業データ）の分類クラスの候補を自動的に提示するシステムを開発したが (高橋他, 2005a), システムを利用したコーダ達から出された要望のうち最も強かったものは, 候補となるクラスに対するシステムの確信度を付与するこ

[1]職業コーディングとは, 自由回答で収集される「仕事の内容」を中心とする職業データ（通常は, 「従業先事業の種類」（自由回答）も含まれる）を総合的に判断して, 国勢調査で用いられる職業小分類を簡略化した約 200 種類の職業に分類しコードを付ける作業をいう (1995 年 SSM 調査研究会, 1995) .

とであった (高橋他, 2005b).

実は, 職業データに対してより正確な分類を行うには, コーダを支援するシステムも有効であるが, データの質の問題として, 分類を誤ったり決定しにくいような回答でないこと, すなわち分類に有効な情報をもつ回答が収集されていることも重要である[2]. 従って, 回答が得られた点で, 分類に必要な情報が不足している可能性の高い回答を自動的に選別し, そのような回答に対しては, その場で回答者自身により情報を追加してもらうことができればよい. 「情報が不足する回答は予測されたクラスに属する確率が他の回答と比較すると低い」と仮定すると, 回答のクラス事後確率を選別の基準として用いることができる.

この例では, クラス事後確率の正確な推定が, 次の処理（回答者から情報を追加してもらう）を行うかどうかを決定する基準となるが, 例えば手書き文字の認識や音声認識のように, 分類結果が他の高レベルなシステムの入力となる場合においても, クラス事後確率の推定は重要である (Zadrozny and Elkan, 2002). さらに, データマイニングにおいて最近注目されているコストに敏感な学習のためにも必要であることが報告されている (Zadrozny and Elkan, 2001).

また, EM アルゴリズムを適用した準教師付き学習などにおいても, クラス事後確率が用いられる (Nigam et al., 2000). ここでも, クラス事後確率をより正確に推定することにより, 分類器の性能を向上させることができる (Tsuruoka and Tsujii, 2003).

以上に述べたように, クラス事後確率は, 人間だけでなくコンピュータまで含めた, 広い意味での意思決定が必要な多くの領域で有用であり, この値をより正確に推定する方法がこれまでにいくつか提案されている (Platt, 1999; Zadrozny and Elkan, 2001; Zadrozny and Elkan, 2002; Tsuruoka and Tsujii, 2003). これらはいずれも, 分類器から出力されるスコア（例えば, 分類器が SVM であれば分離平面からの距離）を用いてクラス事後確率の推定を行っている. 分類器の違いにより出力されるスコアはさまざまであるが, 本稿ではこれらをまとめて「分類スコア」と呼ぶ. ここで, 本稿においても, 事例を多クラスに分類する問題を扱い, 分類スコアは各クラスごとに出力されるものを対象とする.

先行研究における問題点は 2 つある. まず, 分類スコアがすべて 1 位のクラスに対するものに限定されていることである. 実際に分類器により決定されるクラスを観察すると, 1 位のクラスの分類スコアが高くても, 2 位の分類スコアも同程度に高ければ（1 位と 2 位の差が小さければ）, 正解ではない場合がある. 逆に, 1 位の分類スコアが低くても, 2 位の分類スコアが非常に低ければ（1 位と 2 位の差が大きければ）, 正解である場合もある. このように, クラス事後確率は, 1 位の分類スコアに関連して 2 位の分類スコアにも依存すると考えられるため, クラス事後確率を正確に推定するためには, 1 位の分類スコアを用いるだけでは不充分で, 2 位の分類スコアも必要であると考えられる.

次に, 提案されている推定の方法が限られているという問題がある. 特に, Zadrozny ら (2001, 2002) においては, 分類スコアにより降順に並べた事例を同数ずつのグループに分け, 各グループごとにそのグループに属する事例の正解率を計算するが, 平滑化を行う際に当該グループしかみていない. 事例を分類スコアの大きさによりグループ化した場合, 隣接するグループにおける正解率の間には何らかの関係（例えば, あるグループの正解率は前のグループのものより小さく, 後のグループのものより大きいなど）があると考えられるため, この視点を取り入れた方法も考える必要がある.

本稿では, 分類スコアに基づくクラスの事後確率の推定において, 先行研究を参考にしながらも, 新たな視点に基づく方法を提案し比較を行う. その際, 分類スコアを 1 位のものだけでなく 2 位や 3 位のものにも拡張し, 最もよい結果が得られる場合を調査する. また, 分類スコアが何位まで用いられるかということと関係なく, 安定してよい結果を示す方法を調査する.

以下, 次節で関連研究について述べた後, 3 節で本稿で提案する各手法について説明する. 4 節で実験と考察を行い, 最後にまとめと今後の課題について述べる.

2 関連研究

クラス事後確率の推定には, 分類器の出力する分類スコアが入力として用いられる. ここでは, クラス事後確率の値を, シグモイド関数により直接計算して推定する方法と, binning により間接的に推定する方法に分けて述べる.

まず, シグモイド関数を利用する方法としては, Platt(1999) および Tsuruoka(2003) がある. これらの方法では, 分類スコア f を, 単調増加で $[0, 1]$ の値をとるシグモイド関数

$$P(f) = \frac{1}{1 + \exp(Af + B)}$$

に代入して, 直接確率値を求める.

Platt(1999) は, 分類器の出力は次の処理を可能にするために calibrate（検量）された事後確率 $P(class|input)$ であるべきであるとするが, 標準的なサポートベクターマシン（SVM）はこのような確率を提供しないために, SVM による処理に続けて, 分離平面からの関数距離を分類スコア f とし, シグモイド関数により確率値を計算する方法を提案している. Reuters など 5 種類のデータセットにより実験した結果, この方法を, SVM を用いずにロジスティックリン

[2] 現実には, 特に職業データのように, 分類されるクラスの数が多く, クラスにより分類に必要な情報がわかりにくい上に, 回答の表現も回答者により多様であるもの (高橋他, 2005b) については, 調査員に対してこのような要求をすることは非常に困難である.

ク関数により直接クラスに分類する方法と比較すると, 対数尤度において劣ることがなく, 通常の SVM と比較すると, 誤分類された事例の数は少ないという結果であった. われわれもこの方法と同様に, SVM の結果を用いてシグモイド関数による推定も行うが, 分類スコアを 1 位のものだけでなく, 2 位や 3 位のものまで拡張する点が異なる.

Tsuruoka ら (2003) は, EM アルゴリズムが望ましくない状態に収束しないための方法として, クラス分布制約の利用を提案している. クラス分布制約は, ラベルなしデータのクラス分布はラベル付きデータから推定される分布と矛盾しないとする制約で, E ステップと M ステップの間でシグモイド関数により計算される. 語の曖昧性解消タスクにおいて, 大規模なコーパスから集められたラベルなし事例に適用され, 有効性が示されている.

次に, Zadrozny ら（2001, 2002）により提案された binning の方法について述べる. binning とは離散型のノンパラメトリックな方法で, ここでは基本的に次のような手続きをいう. まず, 事例を分類スコアの大きさにより並べ替え, 各区間（ビン）に落ちる事例の数が一定の個数になるように間隔を決める. 次に, ビンごとに事例の正解率を算出し, 平滑化を行って改善した値をそのビンの正解率として定める. その後, 評価を行う新たな事例に対して, 分類スコアの値により該当するビンを見つけ, そこでの正解率をその事例の正解率であるとする.

Zadrozny ら (2001) は, ナイーブベイズ法による確率値を binning により改善する方法を提案している. 他にも決定木による確率値を m-estimation と呼ぶ方法[3] による平滑化や, curtailment と呼ぶ新しい枝刈り法により改善する方法も提案している. Zadrozny ら (2002) は, データセットによっては Platt の方法ではうまく適合しない場合があることを示した上で, binning において問題となるビンの個数の決定を, isotonic 回帰問題に対して最も広範に研究されている PAV（Pool Adjacent Violators）アルゴリズムにより自動的に行う方法を提案している.

Zadrozny らによる方法はわれわれの提案手法と類似しているが, 分類スコアを 1 位しか利用しない点でわれわれの方法と異なる. われわれは, ここでも分類スコアを 1 位だけでなく 2 位や 3 位のものまで用いて比較を行う. なお, われわれは, 確率値の平滑化においても, ラプラス推定だけでなく, より多彩な方法を提案する.

[3] $P = (k + b * m)/(n + m)$ により計算される P を正解率とする. ここで, k, n は各セルにおける正解の事例数と訓練事例数, b は正解率の基本比率（base rate）と呼ぶもので, m はスコアが b に対する近づき方をコントロールするパラメタである. $b = 0.5$, $m = 2$ の場合がラプラス推定である.

3 方法

評価事例のクラス事後確率の推定を, 表（「正解率表」と呼ぶ）を利用して間接的に求める方法と, シグモイド関数（本稿においてはロジスティック回帰分析を用いる）により直接計算する方法に分けて述べる.

3.1 正解率表を利用する方法

正解率表を利用する方法は, 次の通りである.

STEP1　訓練データにおける分類スコアと正誤状況から正解率表を作成する

STEP2　評価事例の分類スコアにより, 正解率表において該当する場所（セル）を探す

STEP3　評価事例が属するセルの正解率をこの事例のクラス事後確率の推定値とする

ここで, 正解率表を作成する方法（STEP1）について述べる.

まず, 分類スコアを軸として, 等間隔の区間に分ける. これは, 分類スコアを複数利用する場合は, それぞれの分類スコアに対して行う. 例えば分類スコアを 1 位のみ利用する場合は, 線分をいくつかの区間に分割することであり, 分類スコアを 2 位まで利用する場合には, 1 位の分類スコアと 2 位の分類スコアを別々に等間隔に区切る. この場合の正解率表の形態は, 2 つの分類スコアをそれぞれ軸とする平面と考えることができるが, 各区間はこれを分割した長方形である. 同様に, 分類スコアを 3 位まで利用する場合の正解率表は三次元空間と考えることができ, 各区間はこれを細分化した直方体である. 本稿では, これらの区間をセルと呼ぶ. 区切る間隔を小さくすればするほど, セルの個数は増える.

次に, 訓練事例が分類スコアに従ってどのセルに属するかを決める. 最後に, 各セルごとに訓練事例の正誤状況を調べ, そのセルにおける正解率を計算する. このような手続きにより, 各セルに対して正解率が 1 つ決定されるが, これが正解率表である.

ここでの問題点は次の 2 つである. 1 つは, 正解率表の分割をどのように行うのが最適かという問題で, これに対しては, 現段階では試行錯誤的に行ってその結果により妥当なセルの大きさを設定している. もう 1 つの問題は, 分割の仕方によっては（例えば, セルを小さくし過ぎた場合）, セル内に訓練事例が出現しないために正解率が欠損値となったり, 出現率が非常に小さいために正解率に対する信頼性の問題が生じる可能性があることである. 従って, 平滑化を行う必要がある.

平滑化の方法

本稿では, 次の 4 つの方法による平滑化を行って比較する.

- ラプラス推定 (北, 1999) による平滑化（以下, ラプラス法と略す）.
- リッドストーン法 (北, 1999) による平滑化（以下, リッドストーン法と略す）.

- 移動平均法による平滑化（以下, 移動平均法と略す）.
- メディアンを用いた平滑化（以下, メディアン法と略す）.

ラプラス法とは, すべての事象について, 生起回数に擬似的に 1 を加える方法である. 本稿においては, 注目するセル $c(f)$（f は分類スコア）に出現する訓練事例数を $N(c(f))$, そのうちの正解事例数を $N_p(c(f))$ とすると,

$$P_{Lap}(f) = \frac{N_p(c(f))+1}{N(c(f))+2} \quad (1)$$

により計算した値をそのセルの正解率とする.

リッドストーン法は, ラプラス法と同様に加算法の一種であるが, 次の式により計算を行う. ただし, δ は生起回数の補正値である:

$$P_{Lid}(f) = \frac{N_p(c(f))+\delta}{N(c(f))+2\delta}. \quad (2)$$

ラプラス推定やリッドストーン法は各セルを独立に扱う. 正解率表全体の状況を観察すると, 大まかには, 近くにあるセル同士は互いに正解率が類似しており, セルの位置が移動するにつれてこの値が単調増加（または減少）する傾向がある. 従って, 平滑化を行う場合に, 対象とするセルの近くにあるセルの正解率を利用することは有効であると考えられる.

移動平均法およびメディアン法 (安居院, 1991) は, この考えに基づいて, われわれが新たに提案した手法である. ここで, どの範囲まで他のセルの情報を利用するのが最適かという問題があるが, 今回は, 単純に平滑化の対象とするセルに隣接するものに限定した. 隣接するセルの数は, 分類スコアを 1 位のみ用いる場合は 2 個[4] , 2 位まで利用する場合は 8 個, 3 位までの場合は 26 個である.

移動平均法は, 平滑化の対象とするセルおよび隣接するセルにおける正解率の平均を計算し, この値を対象とするセルの正解率とする方法である. すなわち, $P_{MA}(f)$ は次式により計算される:

$$P_{MA}(f) = \frac{\frac{N_p(c(f))}{N(c(f))} + \sum_{s\in Nb(c(f))} \frac{N_p(s)}{N(s)}}{n}. \quad (3)$$

ただし, s は正解率表における任意のセル, $Nb(c(f))$ は注目するセル $c(f)$ に隣接するセルの集合を表す. また, n は, 平滑化の対象とするセルと隣接するセルのうち正解率が存在するセルの数を表す[5] .

この移動平均法においては, 平滑化を行うセルは, 隣接するセルのいずれとも関連の程度が等しいと仮定している. しかし, もし分類スコア間で, 隣接するセル同士の関連の程度に違いがあると仮定する場合には, セルの正解率に重み付けを行う必要がある. 例えば, 分類スコアを 2 位まで利用する場合, 平滑化を行うセルの正解率と垂直方向で隣接するセル（1 位の分類スコアで並ぶ）の正解率と, 水平方向で隣接するセル（2 位の分類スコアで並ぶ）の正解率とで, 関連する程度が異なると仮定するなら, 両者に対して異なる重み付けを行う必要がある.

メディアン法は, 平滑化の対象とするセルと隣接するセルにおけるメディアンを求め, この値を対象とするセルの正解率とする方法である. すなわち, $P_{Med}(f)$ は次式により計算される:

$$P_{Med}(f) = MEDIAN\Big(\frac{N_p(c(f))}{N(c(f))}, \{\frac{N_p(s)}{N(s)}\}_{s\in Nb(c(f))}\Big). \quad (4)$$

3.2 ロジスティック回帰分析を用いる方法

ロジスティック回帰分析を用いる方法では, 評価事例の分類スコアをロジスティック回帰式の独立変数として, 直接クラス事後確率を計算する. ここで, 分類スコアを 1 位のみ利用する場合（f）, 2 位まで利用する場合（f_1, f_2）, 3 位まで利用する場合（f_1, f_2, f_3）のロジスティック回帰式はそれぞれ,

$$P_{Log}(f) = \frac{1}{1+\exp(Af+B)}, \quad (5)$$

$$P_{Log}(f_1,f_2) = \frac{1}{1+\exp(\sum_{i=1}^{2} A_i f_i + B)}, \quad (6)$$

$$P_{Log}(f_1,f_2,f_3) = \frac{1}{1+\exp(\sum_{i=1}^{3} A_i f_i + B)} \quad (7)$$

で表される. ただし, (5) 式から (7) 式におけるパラメタは, 次に示すように最尤法により推定する.

ここでは, (5) 式におけるパラメタの推定を示す.

$$P_{Log}(f;A,B) = \frac{1}{1+\exp(Af+B)} \quad (8)$$

と置く. 与えられた事例の分類スコアを f^i とすると, 正解（$Y^i=1$）である確率は $P_{Log}(f^i;A,B)$, 不正解（$Y^i=0$）である確率は $1-P_{Log}(f^i;A,B)$ であるため, $Y^1, \cdots, Y^n$ を得る同時確率を A, B の関数と考えれば, 次の尤度関数が得られる:

$$L(A,B) = \prod_{Y^i=1} P_{Log}(f^i;A,B) \times \prod_{Y^i=0} [1-P_{Log}(f^i;A,B)]. \quad (9)$$

最尤推定量 A, B を求めるためには, 両辺の対数をとり, これを最大にする推定量を求めればよい.

(6) 式や (7) 式におけるパラメタの推定方法も同様である.

[4]対象とするセルが端点である場合には隣接するセルがないためにこれより少ない. 以下同様である.

[5]正解率が欠損値であるセルは加えない. 以下同様である.

4 実験

4.1 実験方法

データセット

実験に用いたデータセットは, JGSS（日本版 General Social Survey）[6] により 2000 年から 2003 年まで毎年実施された調査（JGSS-2000, ⋯ , JGSS-2003）(大阪商業大学・東京大学, 2005) のうちの有職者 23,838 サンプルである. このうち, JGSS-2000, JGSS-2001, JGSS-2002（20,066 サンプル）を訓練データとし, JGSS-2003（3,772 サンプル）を評価データとした. 用いたデータは, 仕事の内容（自由回答）, 従業先事業の種類（自由回答）, 従業上の地位（選択回答）から構成される職業データである. 職業データは, すでに調査終了後に行われた職業コーディングにより, 職業コード（1 個）が付与されており, われわれはこの職業コードを正解として扱う.

分類スコア

今回の実験では, 分類器はサポートベクターマシン（SVM）を用いた. SVM は二値分類器であるために, one-versus-rest 法を用いて多値分類器へと拡張した. また, 高橋他（2005b）により, SVM のカーネル関数は線型カーネルを用い, 事例に与える重みの上限であるソフトマージンパラメタは $C = 0.6$ に設定した.

分類スコアとしては, SVM により出力される分離平面からの距離を用いた. 分類スコア間には次の関係がある：

1 位の分類スコア >2 位の分類スコア >3 位の分類スコア.

正解率表の作成

今回, 正解率表を作成するためのデータセットとしては, JGSS-2000, JGSS-2001, JGSS-2002（20,066 サンプル）を用いた. まず, これらのデータにより分類器を生成し, 入力データとしてこれらのデータを分類器に与えた. 次に, 分類器より出力された分類スコアの分布状況を概観した結果, どの分類スコアも 0.25 刻みの等間隔に分割することにした. これにより, 1 位の分類スコア f_1 は, （ $f_1 \leq -0.75$, $-0.75 < f_1 \leq -0.5$, ⋯ , $1.75 < f_1 \leq 2$, $2 < f_1$ ）のように 13 個に分割された. 以下, 2 位の分類スコア f_2 は, （ $f_2 \leq -1$, $-1 < f_2 \leq -0.75$, ⋯ , $-0.25 < f_2 \leq 0$, $0 < f_2$ ） の 6 個, 3 位の分類スコア f_3 は, （ $f_3 \leq -1$, $-1 < f_3 \leq -0.75$, $-0.75 < f_3$ ） の 3 個に分割された. 従って, セルの数は分類スコアを 1 位のみ利用する場合は 13 個, 2 位までの場合は 78 個, 3 位までの場合は 234 個となる[7] .

以上により生成されたセルごとに正解率を算出したものが基本となる正解率表である. 前述したように, この正解率表には欠損値などの問題が生じる可能性が高いために, そのまま用いることができない. 従って, 本稿で提案する手法によりそれぞれ平滑化を行い, 新たな正解率表を作成した.

各手法の評価は負の対数尤度により行い, この値が小さいものほどよい手法であると判断する.

ここで, リッドストーン法においては, 事前に補正値 δ の最適値を決定しておく必要がある. このため, 訓練データをさらに JGSS-2000, JGSS-2001（13,296 サンプル）と JGSS-2002（6,770 サンプル）の 2 つに分割し, まず, JGSS-2000, JGSS-2001 を用いて基本となる正解率表を作成した. 次に, δ の値を, 0.01〜350（分類スコアを 1 位のみ利用する場合）, 0.01 〜100（分類スコアを 2 位まで利用する場合）, 0.01 〜350（分類スコアを 3 位まで利用する場合）の範囲で変化させてそれぞれ平滑化を行った正解率表を作成した. この正解率表を用いて JGSS-2002 の対数尤度を計算し, その値が最も大きいときの δ の値を使用した.

なお, 移動平均法において重み付けを行う場合も, 事前に最適な重み付けの方法を調査しておく必要がある. 本稿においては, 平滑化の対象とするセルに隣接するセルに対して, 1 位の分類スコアにより隣接するセルとそれ以外のセルに対する重みを変えて実験を行ったが[8] , 重み付けを変えない場合に最もよい結果を示した. 従って, 移動平均法においては重み付けを行わない.

4.2 各手法における対数尤度の比較

正解率表を利用する方法

まず, リッドストーン法における δ の最適値は, 分類スコアを 1 位のみ利用する場合は 300, 2 位まで利用する場合は 50, 3 位まで利用する場合は 35 であった.

表 1 に, 正解率表を利用する場合の負の対数尤度と順位を分類スコアの利用別に示す. 順位はすべての手法の中での順位である. ベースラインはすべての事例においてクラス事後確率値が 0.5 の場合である. ただし, リッドストーン法については, 最適な δ に対する値のみを示す. また, *を付けた値は, 欠損値のためにクラス事後確率が 0 となった事例に対して, 仮に非常に小さい値（今回は $P = 1/2^8$ とした）を用いて対数尤度を計算したことを表す. ただし, この方法が適切であるかどうかについては議論が必要である.

実験の結果, 次の 4 つが明らかになった.

まず, 平滑化を行う手法はいずれも平滑化を行わない場合を上回る. また, 分類スコアを 1 位のみまたは 2 位まで利用する場合は, 平滑化を行わなくてもベースラインを上回る.

さらに, すべての手法において, 分類スコアを 2 位まで利用する場合が最もよい結果が得られ, 次には 3 位まで利用する場合で, 最も結果が悪いのは 1 位のみ利用する場合である. ただし, 利用する分類スコア数が増えるに従い 1 つのセルに含まれる事例数が減るた

[6] http://jgss.daishodai.ac.jp/

[7] ただし, 前述した分類スコア間の関係より, 複数の分類スコアを利用する場合には無効なセルが存在する.

[8] 用いたデータセットは, リッドストーン法の場合と同様である.

表 1: 各手法における負の対数尤度と順位（利用したスコア別）

利用したスコア	ベースライン	平滑化なし	ラプラス法	リッドストーン法	移動平均法	メディアン法	ロジスティック回帰分析
1 位のみ	3772.0 18	3213.0 16	3200.5 15	2641.7 8	3110.8 13	3242.6 17	3172.4 14
2 位まで	3772.0 18	2832.1* 12	2664.7 10	2612.3 7	2476.9 1	2532.6 3	2597.3 5
3 位まで	3772.0 18	4329.9* 19	2728.0 11	2658.1 9	2492.5* 2	2610.7* 6	2545.4 4

めに, 移動平均法やメディアン法においては欠損値の可能性がある. 実際に, 分類スコアを 3 位まで利用する場合にはこの問題が生じている.

利用する分類スコアの数に関係なく, 最も安定してよい結果を示したのは移動平均法である. 特に, 分類スコアを 2 位または 3 位まで利用する場合に最もよい. このとき, 隣接するセルの情報に重み付けを行わず, 均等に用いる方がよい.

最後に, ラプラス法やリッドストーン法は欠損値の問題は生じないが, 利用する分類スコアの数に関係なく, 他の手法に劣る場合が多い. 特に, ラプラス法の結果はよくない. ただし, 分類スコアを 1 位のみ利用する場合には, リッドストーン法が最もよい.

以上より, 正解率表を利用する方法においては, 分類スコアを 2 位まで利用し, 重み付けを行わない移動平均法による平滑化の手法が最もよい.

ロジスティック回帰分析を用いる方法

ロジスティック回帰分析 (5)〜(7) におけるパラメタの推定結果は次の通りであった.

$A = 1.873$, $B = -1.837$,

$A_1 = 2.022$, $A_2 = -3.114$, $B = -0.756$,

$A_1 = 1.991$, $A_2 = -3.052$, $A_3 = -3.503$, $B = -4.194$

この値を用いた回帰式に分類スコアを入力として対数尤度を計算した結果を, 表 1 の右端列に示す.

表から明らかなように, ロジスティック回帰分析を用いる場合には, 分類スコアを利用する数が多いほどよい結果が得られる. しかし, 2 位まで利用する場合と 3 位まで利用する場合の差は 1 位のみの場合と 2 位までの場合ほど差がない.

ロジスティック回帰分析を用いる方法を正解率表を利用する方法と比較すると, 分類スコアを 3 位まで利用する場合は 2 番目, それ以外の場合は 3 番目で比較的よい. また, 欠損値の問題が生じない点もよい[9] .

4.3 セルのサイズを変化させた場合の対数尤度

セルのサイズを 0.25 にして正解率表を作成した場合は, 表 1 に示すように, 分類スコアを 2 位まで利用するのがよく, またこの場合の移動平均法による平滑化手法が最もよい手法であった. この結果を確認するために, 分類スコアを 1 位のみ利用する場合と 2 位まで利用する場合において, すべての手法に対してセルのサイズを 0.2, 0.3, 0.5 の 3 通りに変化させ, 実験を行った.

実験の結果を表 2 および表 3 に示す. 順位はセルのサイズが等しい場合を表す.

まず, セルのサイズに関係なく, すべての手法において分類スコアを 2 位まで利用する場合は, 1 位のみ利用する場合よりよい結果が得られた.

次に, 分類スコアを 2 位まで利用する場合において, セルのサイズが 0.5 の場合を除いて移動平均法による平滑化手法が最もよかった（セルのサイズが 0.5 の場合は 2 番目であった）. 最もよいものから 3 番目によいものまでは, 移動平均法による平滑化手法, ロジスティック回帰分析, メディアンによる平滑化手法の 3 つがセルのサイズの違いにより順位を入れ替えるだけで, ラプラス法やリッドストーン法はつねに下位であった.

以上をまとめると, クラス事後確率の推定には, 分類スコアを 2 位まで利用した移動平均法による平滑化の手法が安定してよい. ただし, 正解率表におけるセルのサイズにより, 対数尤度の値が異なるため, 適切に決める必要がある. また, 分類スコアを細分化し過ぎると欠損値の問題が生じる可能性があることに注意する必要がある.

4.4 移動平均法におけるクラス事後確率の推定

ここでは, 今回最もよい結果が得られた, 分類スコアを 2 位まで利用した移動平均法による平滑化手法において, どの程度正確な推定が行えるかについての評価を行う. 評価の方法は, 全評価データをクラス事後確率の推定値の大きさの順に並べ, 累積カバー率を増加させたときの各区間における推定値を実測値と比較する. ここで, クラス事後確率の実測値とは, 対象とする区間における正解率とする.

図 1 に, クラス事後確率の推定値を降順に並べ, 累積カバー率を 10%から 100%まで 10%刻みに増加させた場合の, 各区間における推定値と実測値の平均を示す. 同様に, 図 2 は, クラス事後確率の推定値を昇順

[9]ただし, パラメタの数に比較して事例数が少ない場合にはパラメタの推定ができない可能性がある.

表 2: セルのサイズを変えた場合の各手法における負の対数尤度と順位（1 位のみのスコアを利用）

セルのサイズ	ベースライン	平滑化なし	ラプラス法	リッドストーン法	移動平均法	メディアン法	ロジスティック回帰分析
0.2	3772.07	3233.26	3214.95	2622.91	3125.42	3204.94	3172.43
0.3	3772.07	3227.15	3216.34	2694.31	3068.12	3233.16	3172.43
0.5	3772.07	3184.85	3178.64	2733.81	3101.12	3279.16	3172.43

表 3: セルのサイズを変えた場合の各手法における負の対数尤度と順位（2 位までのスコアを利用）

セルのサイズ	ベースライン	平滑化なし	ラプラス法	リッドストーン法	移動平均法	メディアン法	ロジスティック回帰分析
0.2	3772.07	3107.6*6	2622.64	2624.05	2408.61	2480.32	2597.33
0.3	3772.07	2990.8*6	2814.55	2694.14	2525.71	2602.43	2597.32
0.5	3772.07	2919.0*6	2887.85	2738.74	2602.02	2702.93	2597.31

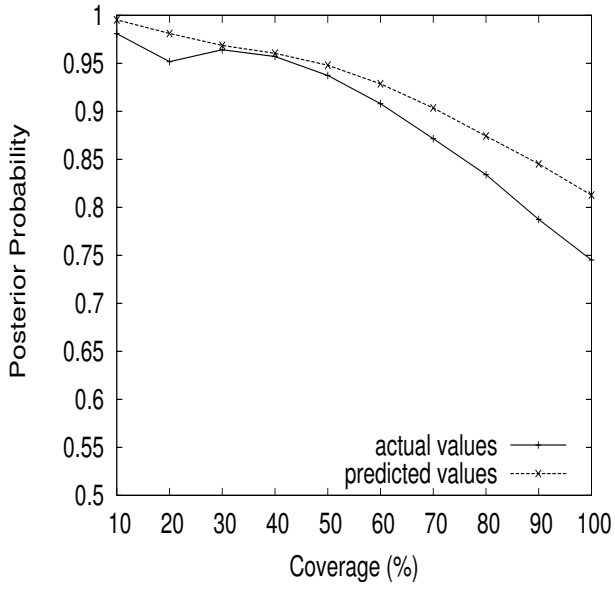

図 1: 累積カバー率別クラス事後確率の推定値と実測値（降順）（分類スコアを 2 位まで利用する場合）

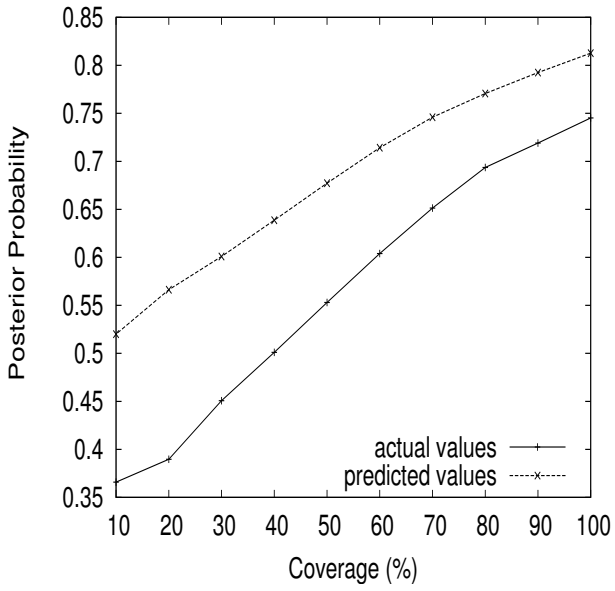

図 2: 累積カバー率別クラス事後確率の推定値と実測値（昇順）（分類スコアを 2 位まで利用する場合）

に並べた場合の結果を示す.

まず, 図 1 および図 2 より, クラス事後確率の推定値はつねに実測値より高めであることがわかる. 図 1 において, 推定値が高いほど実測値との差が小さく, 低くなるに従って差が広がる傾向がある.

これより, 分類スコアを 2 位まで利用した移動平均法による平滑化手法は, 実際より高めに推定を行うが, 推定値が高いほど正しい値に近く, 特に上位 50%まではほぼ正確な推定を行うことができる.

図 2 では, 推定値が低い場合にどの程度正確な推定を行えるかがわかる. 推定値が低いほど実測値との差が大きいが, 特に下位 10%の区間では, 推定値と実測

値の差は 0.1541 である.

実際にクラス事後確率の推定を不正解である事例の発見に用いるためには, 推定値が低い場合においてより正確さが要求される. 推定値と実測値の差に対する評価については, 現段階では明確でなく, 今後, 調査を行って明らかにする必要がある.

5 おわりに

本稿では, クラス事後確率の推定を行うために, 分類器の出力するスコアを複数利用することおよび, 正解率の平滑化に移動平均法などを利用することを提案した.

職業データによる実験の結果, 分類スコアは 2 位まで利用する場合が最もよかった. また, 手法は, 移動平均法により平滑化を行った正解率表を利用して, 間接的にクラス事後確率を推定する手法が最もよく, 2 番目は, ロジスティック回帰分析を利用して直接クラス事後確率を計算する手法であった. ラプラス推定やリッドストーン法による平滑化の手法は, 欠損値の問題を考慮する必要がないという長所があるものの, 分類スコアを 1 位のみ利用する場合のリッドストーン法を除き, 他の手法より劣っていた.

しかし, 正解率表を利用する方法においては, セルの決め方による影響があるために, ここでの結果をそのまま一般化することはできない. 今後の課題として, まず, 本稿で提案した以外の方法による場合についての調査を行う必要がある. 本稿においては, セルの大きさをデータの分布状況から恣意的に決めたが, 次の段階として, 例えば最小記述長原理などの情報量基準を用いて自動的に決定する方法 (下平他, 2004) を適用していく予定である. また, 正解率表の平滑化の方法として, 最大エントロピー法による手法の実験も行う予定である. 次に, 本稿で用いたデータセット以外のデータセットに対しても同様の実験を行い, われわれの提案する手法の有効性について確認したい.

6 謝辞

日本版 General Social Surveys（JGSS）は, 大阪商業大学比較地域研究所が, 文部科学省から学術フロンティア推進拠点としての指定を受けて（1999-2003 年度）, 東京大学社会科学研究所と共同で実施している研究プロジェクトである（研究代表：谷岡一郎・仁田道夫, 代表幹事：佐藤博樹・岩井紀子, 事務局長：大澤美苗）. 東京大学社会科学研究所附属日本社会研究情報センター SSJ データアーカイブがデータの作成と配布を行っている.

References

1995 年 SSM 調査研究会. 1995. SSM 産業分類 · 職業分類（95 年版）.

安居院猛, 中嶋正之. 1991. 画像情報処理. 森北出版

R. K. Ahuja and J. B. Orlin. 2001. A Fast Scaling Algorithm for Minimizing Separable Convex Functions Subject to Chain Constraints. *Operations Research* Vol.49, pp. 784–789.

北研二. 1999. 言語と計算４　確率的言語モデル. 東京大学出版会.

K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning* 39(2/3), pp. 103–134.

J. C. Platt. 1999. Probabilistic Outputs for Support vector Machines and Comparisons to Regularized Likelihood Methods. In *Advances in Large Margin Classifiers*, pp. 1–11. MIT Press.

大阪商業大学比較地域研究所・東京大学社会科学研究所（編）. 2005. 日本版 General Social Surveys 基礎集計表・コードブック JGSS-2003. 大阪商業大学比較地域研究所.

下平英寿, 伊藤秀一, 久保川達也, 竹内啓. 2004. モデル選択. 岩波書店.

高橋和子, 須山敦, 村山紀文, 高村大也, 奥村学. 2005a. 職業コーディング支援システム（NANACO）の開発と JGSS-2003 における適用. 日本版 General Social Surveys 研究論文集［4］ JGSS で見た日本人の意識と行動, pp.225–242.

高橋和子, 高村大也, 奥村学. 2005b. 機械学習とルールベース手法の組み合わせによる自動職業コーディング. 自然言語処理　Vol.12 No.2, pp. 3–24.

東京大学教養学部統計学教室（編）. 1992. 自然科学の統計学. 東京大学出版会.

Y. Tsuruoka and J. Tsujii. 2003. Training a naive Bayes Classifier via EM Algorithm with a Class Distribution Constraint. In *Proceedings of the Seventh CoNLL Conference*, pp. 127–134.

B. Zadrozny and C. Elkan. 2001. Learning and Making Decisions When Costs and Probabilities are Both Unknown. In *Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining (KDD'01)*, pp. 204–213.

B. Zadrozny and C. Elkan. 2002. Transformation Classifier Scores into Accurate Multiclass Probability Estimates. In *Proceedings of the Eighth International Conference on Knowledge Discovery and Data Mining (KDD'02)*, pp. 694–699.