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　　‡ 東京工業大学 精密工学研究所
takak@u-keiai.ac.jp　{takamura,oku}@pi.titech.ac.jp

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に, 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, その事例の候補となるクラスにどの程度そのクラスらしいかという確率が付与されれば, 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の要望で最も多かったのは, 候補となるクラスに対してシステムの確信度を付与することであった (高橋他, 2005b). また, 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 (Platt, 1999; 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センシティブ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) .

クラス事後確率はまた, 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用いられ (Nigam et al., 2000), これ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).

さらに, クラス事後確率は, 誤分類された事例の発見にも有用である. 「誤分類された事例は正しく分類された事例より予測されたクラスに属する確率が低い」と仮定すると, クラス事後確率が低い事例ほど誤分類された可能性が高いと考えることができる.

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, 分類器がサポートベクタマシン（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. また, 本稿でも, 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とする.

先行研究における問題点の 1 つは, 利用する分類スコアをすべて 1 位のクラスに対するものに限定していることである. 実際には, 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）, 正解ではないことがある. 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）, 正解になりやすいであろう. 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, 本稿では, 複数の分類スコアを利用することにする. この他にも, 本稿では分類スコアに基づくクラスの事後確率の推定を行うために, 新たな視点による方法も提案し比較を行う.

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 $P(f) = 1/\{1 + exp(Af + B)\}$ に代入して, 直接確率値を求める. Platt(1999) は, サポートベクターマシン（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る実験の結果, 良好な確率値を求めることができたとしている. 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 の方法は離散型のノンパラメトリックな方法で, 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[2] による平滑化や, curtailment

[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) .

[2] $P = (k + b * m)/(n + m)$ により計算される P を正解率とす

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．

3 方法

本稿では, 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より行う．

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

正解率表を作成するためのデータの用い方として, 次の 2 つの方法が考えられる．1 つは, このデータをすべて用いて分類器を構築し, 同じデータ自身を分類器に与える方法である (高橋他, 2005c)．もう 1 つは, データを例えば 5 つに分割し, このうちの 4/5 を用いて分類器を構築し, 残りの 1/5 を分類して分類スコアを算出する手続きを, データを変えて 5 回行う 5 分割の交差検定による方法である．今回はこの方法により正解率表を作成することにする．

以下では, 正解率表の作成（STEP1）方法を述べる．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[3]．本稿では, これらの区間をセルと呼ぶ. 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 つ決定されたものが, 正解率表である．

正解率表を作成する際, 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．これを防止するためには平滑化を行う必要がある．本稿では, 次の 4 つの方法により平滑化を行う．

- ラプラス推定 (北, 1999) による平滑化（以下, ラプラス法）．
- 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）．
- 移動平均法による平滑化（以下, 移動平均法）．
- メディアンを用いた平滑化（以下, メディアン法）．

ラプラス法は, 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 $N_p(c(f))$ とし,

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

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:

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

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が, 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方法も有効ではないかと考え, 移動平均法やメディアン法 (安居院, 1991) による平滑化も行うことにする．この場合, 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する[4]．

移動平均法およびメディアン法は, 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それぞれ平均およびメディアンを計算し, この値を対象とするセルの正解率とする方法である．それぞれ次式により計算される:

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

$$P_{Med}(f) \quad = \quad median\Big(\frac{N_p(c(f))}{N(c(f))}, \{\frac{N_p(s)}{N(s)}\}\Big). \tag{4}$$

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]．

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 r 位まで利用する場合（f_1, $\cdots$, f_r）のロジスティック回帰式は,

$$P_{Log}(f_1,\cdots,f_r) \quad = \quad \frac{1}{1+\exp(\sum_{i=1}^{r}A_if_i+B)} \tag{5}$$

で表される．ただし, (5) 式におけるパラメタは, あらかじめ最尤法により推定しておく必要がある．

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]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 位と 2 位の分類スコアを別々に等間隔に区切る．

[4]隣接するセルの数は, 利用する分類スコアが 1 位のみ, 2 位まで, 3 位までの場合にそれぞれ 2 個, 8 個, 26 個となる．ただし, 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．

[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

4.1.1 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）(高橋他, 2005b) 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．

4.1.2 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.

4.1.3 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．これを先ほど述べた 5 分割の交差検定により, 事例の分類スコアとそれが正解であるかどうかのデータを得た．セルの区間幅は, どの分類スコアも 0.25 刻みの等間隔に分割した[7] ．以上により正解率表を作成した後, 本稿で提案する平滑化の手法を用いて新たな正解率表を作成した．なお, リッドストーン法においては, 事前に JGSS-2000, JGSS-2001, JGSS-2002 により補正値 δ の予測最適値を決定した．

4.2 各手法における対数尤度の比較

各手法の評価は対数尤度により行い, この値が大きいものほどよい手法であると判断する．

4.2.1 正解率表を利用する方法

リッドストーン法における δ の予測最適値は, 分類スコアを 1 位のみ利用する場合は 6, 2 位まで利用する場合は 0.6, 3 位まで利用する場合は 0.9 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．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．リッドストーン法は最適な予測値 δ に対する値を示す．また, *を付けた値は, 欠損値など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[8] ．

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

[7] 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, 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．

[8] ただし, この方法の妥当性については議論が必要である．

表から次のことが明らかである．すべての手法において, 分類スコアを 1 位のみ利用する場合より複数利用する場合の方がよい．特に, リッドストーン法を除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に最もよい結果が得られる．これは, 今回と正解率表の構築法が異なる高橋ら (2005c) の結果と同様である．

4.2.2 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（(5) 式）におけるパラメタの推定結果は, それぞれ次の通りであった．

$A_1 = 1.768$, $B = 0.325$,

$A_1 = 1.810$, $A_2 = -1.789$, $B = -0.879$,

$A_1 = 1.760$, $A_2 = -1.542$, $A_3 = -1.115$, $B = -1.691$

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．また, ロジスティック回帰分析による方法を正解率表を利用する方法と比較すると, 最も結果がよい, 分類スコアを 3 位まで利用する場合以外は, 正解率表を利用する手法の方がよい．この結果はいずれも高橋ら (2005c) と同様である．

4.2.3 セルの区間幅を変化させた場合

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区間幅を 0.1, 0.2, 0.3, 0.5 の 4 通りに変化させて実験を行った．

まず, すべてのセルの区間幅, 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 1 位のみ利用する場合よりよい結果であった．また, 利用した分類スコアの数に関係なく, すべての手法において区間幅を 0.2 または 0.25 にした場合が最もよかった．これらは高橋ら (2005c) と同様の結果であるが, 今回は, 区間幅によりよい手法が変化する点が異なっている．すなわち, 高橋ら (2005c) では, 分類スコアを 1 位のみ利用した場合はリッドストーン法, 2 位まで利用した場合は移動平均法がつねによい結果を示したのに対して, 今回は, 利用した分類スコア数に関係なく, セルの区間幅が小さい場合は移動平均法がよく, 区間幅を 0.25 または 0.5 にした場合にラプラス法またはリッドストーン法がよかった．特に分類スコアを 2 位まで利用し, 区間幅を 0.25 にした場合のラプラス法は, すべての中で最もよかった．

4.2.4 セルの区間幅が等しい場合とサンプル数が等しい場合の比較

利用した分類スコアの数が 1 位のみの場合において, 我々の提案するセルの区間幅が等しい手法と, Zadrozny ら（2001, 2002）により提案されたセル内のサンプル数が等しい手法についての比較を行った．セルの数を 30, 16, 13, 12, 7 個にして実験した結果, すべての場合においてセルの区間幅を等しくした方がよい結果であった．特にセルの数が多いほど, 両者の差は大きかった．

4.3 クラス事後確率の予測推定値とその誤り事例発見能力に対する評価

ここでは, 今回最もよい結果が得られた, 分類スコアを 2 位まで利用した場合のラプラス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評価を行った．

まず, 全評価データをクラス事後確率の推定値の大きい順に並べ, 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した．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 は, クラス事後確率の推定値を降順に並べて実測値と比較した結果である．これより, 区間ごとの平均値ではあるが, 推定値は実測値にほぼ等しい．昇順に並べた場合も同様の結果であった．

次に, 誤り事例を発見する能力について, 分離平面からの距離をそのまま利用する方法（以下, 直接法）(Schohn and Cohn, 2000) と比較した．図 2 は, 全評価データをラプラス法ではクラス事後確率の推定値, 直説法では 1 位の分類スコアが小さい順に並べ, 累積カバー率を増加させたときに両手

表 1: 各手法における負の対数尤度（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	2361.3	2361.1	2361.3	2362.9	2363.8	2367.5
2 位まで	3772.0	2234.2*	2229.6	2237.0	2250.2	2246.9	2246.9
3 位まで	3772.0	2234.2*	2232.7	2231.3	2275.0	2251.8	2232.9

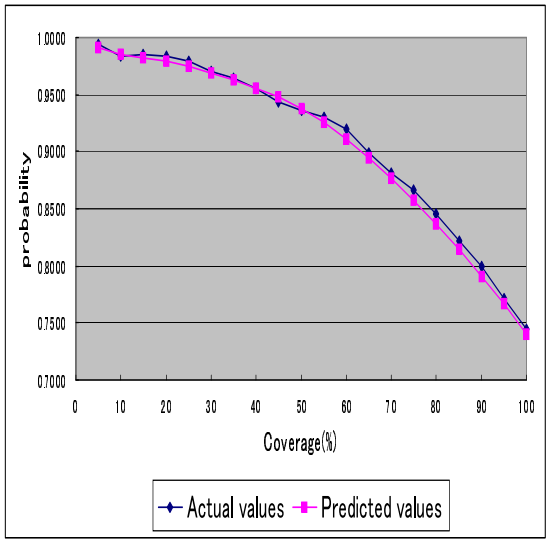

図 1: 分類スコアを 2 位まで利用したラプラス法における累積カバー率別クラス事後確率の推定値と実測値（降順）

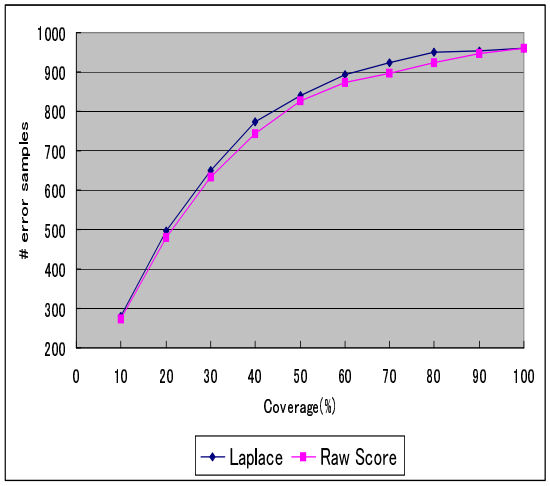

図 2: 分類スコアを 2 位まで利用したラプラス法と直接法における累積カバー率別誤り事例発見数

法が正しく発見できる誤り事例数を示す．これより，ラプラス法はつねに直接法を上回っており，ラプラス法により推定された事後確率の低いものから 40%をとると，誤り事例全体の約 80%以上を発見することができる．

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，ラプラス法などにより平滑化を行った正解率表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，セルの区間幅の決め方によりよい手法は異なるが，すべての場合の中で，分類スコアを 2 位まで利用しラプラス法による平滑化を行った正解率表を利用する方法が最もよかった．また，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もよかった．ラプラス法により作成される正解率表を利用する手法は，区間内での平均値ではあるが実測値とほぼ等しい推定を行うことができ，誤り事例の発見においても，分離平面からの距離をそのまま用いる方法を上回っていた．しかし，セルの区間幅が小さい場合には，利用する分類スコア数に関係なく移動平均法により平滑化を行う手法がよい結果を示すため，今後，この手法についての調査を深める必要がある．

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．

参考文献

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.

G. Schohn and D. Cohn. 2000. Less is More: Active Learning with Support Vector Machines. *17th ICML*, pp. 839–846.

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.

高橋和子，高村大也，奥村学．2005c．分類スコアに基づいたクラス事後確率の推定．情報処理学会研究報告 2005-NL-170(16), pp. 97–104.

Y. Tsuruoka and J. Tsujii. 2003. Training a naive Bayes Classifier via EM Algorithm with a Class Distribution Constraint. *7th CoNLL*, pp. 127–134.

B. Zadrozny and C. Elkan. 2001. Learning and Making Decisions When Costs and Probabilities are Both Unknown. *KDD'01*, pp. 204–213.

B. Zadrozny and C. Elkan. 2002. Transformation Classifier Scores into Accurate Multiclass Probability Estimates. *KDD'02*, pp. 694–699.