# 複数の分類スコアを用いたクラス所属確率の推定

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

文書分類においては, 与えられた事例が分類器により決定されたクラスに属する確率がどの程度であるかを知ること, すなわちクラス所属確率を正確に推定することは重要である.

例えば, われわれは, 社会調査において「職業コーディング」[1] を担当するコーダを支援する目的で, 回答（職業データ）を機械学習により自動分類した結果を候補として提示するシステムを開発したが (Takahashi et als., 2005), システムを利用したコーダ達から, 提示された候補に対するシステムの確信度（クラス所属確率）を付与してほしいとの要望がつねに出されてきている (高橋他, 2005). 実際, クラス所属確率の推定は, 語の曖昧性解消 (Chan, 2006) や質問応答における質問の補充生成 (Jones, 2006) においても行われており, データマイニングにおけるコストセンシティブ学習 (Zadrozny and Elkan, 2001) や画像パターンの分類 (Devarakota, 2007) など文書分類以外のさまざまな分野においても重要な役割を果たしている.

クラス所属確率は, 分類器が出力するスコア（分類スコア）を単純に変形しても推定することができるが[2], 正確さに欠ける場合が多い (Zadrozny and Elkan, 2001). そこで, クラス所属確率をより正確に推定する方法が提案されてきたが (Bennett, 20006; Zadrozny and Elkan, 2001), Platt により提案されたシグモイド関数を用いる方法 (Platt, 1999) や, Zadrozny らにより提案された Isotonic regression による方法 (Zadrozny and Elkan, 2002) の 2 つがよく知られている (Mizil, 2005a). 特に, Zadrozny らは, 多値分類における任意のクラスについても, 推定したい（注目する）クラスとそれ以外のクラスの 2 つに分解すれば, 2 値分類と同様の方法で推定できることを示した.

しかし, 先行研究はいずれも, 注目するクラスの分類スコアしか用いていないという問題がある. これが問題であるとする理由は, 事例が分類されたクラスの正誤状況と分類スコアの間に, 次のような関連が観察されるからである. 分類スコアが各クラスごとに出力されるとき, 事例が分類されるのは分類スコアが最大（第 1 位）のクラスである. このとき, もし第 2 位のクラスの分類スコアも第 1 位の分類スコアと同程度に高ければ（第 1 位と第 2 位の分類スコアの差が小さければ）, 予測されたクラスは正解ではない場合がある. 逆に, 第 1 位のクラスの分類スコアが低くても, 第 2 位のクラスの分類スコアが非常に低ければ（第 1 位と第 2 位の分類スコアの差が大きければ）, 予測されたクラスは正解である場合がある. したがって, クラス所属確率は, 第 1 位のクラスの分類スコアだけでなく, 例えば第 2 位のクラスのような他のクラスの分類スコアにも依存すると考えられる.

以上より, Takahashi ら (Takahashi et als., 2007) は, 第 1 位のクラスについてのクラス所属確率を正確に推定するために, 第 1 位のクラスだけでなく第 2 位のクラスの分類スコアも用いることを提案し, 有効性を示した. また, 複数の分類スコアを用いる際に, あらかじめ分類スコアを軸と考えて各スコア軸を等間隔に区切って作成したセルごとに正解率を求めた「正解率表」を用意しておいて利用する方法を提案し, ロジスティック回帰による方法と比較すると, 正解率表を利用する方法は, カバレッジを重みとする移動平均法により正解率表の平滑化を行ったり, 分類スコアの区間幅をうまく決めれば, ロジスティック回帰による方法より有効であった. 一方で, ロジスティック回帰を利用する方法は安定してよい結果を示した.

しかし, 一般に文書分類は多値分類である場合が多いため, 第 1 位だけでなく第 2 位以下のクラスについてもクラス所属確率を推定できることが望ましい. 例えば, 上述した自動職業コーディングシステムにおいても, 提示する第 5 位までの候補のすべてに対して確信度の付与が要請されている. ここで, 第 2 位以下のクラスについても, 第 1 位のクラスの場合と同様に複数の分類スコアを用いることが有効であると考えると, 分類スコアのソートが前提である Isotonic regression では複数（次元）の分類スコアを扱うことが困難なため, Zadrozny らの方法には限界がある.

そこで, 本稿では, Takahashi らの方法を第 2 位以下の任意のクラスに対して拡張する. ただし, 任意のクラスに注目する場合には, 注目するクラス以外に複数個のクラスを考慮するとクラスの組み合わせ方が複雑になり過ぎるため, 今回は最も有効なクラス 1 つを追加することにする. ここで, 注目するクラスの順位が何位であっても, その分類スコアは第 1 位のクラスの分類スコアより小さい値しかとらないために, 値それ自身より第 1 位のクラスの分類スコアとの相対的な大きさが影響すると考えられる. したがって, 注目するクラスのクラス所属確率は, 第 1 位のクラスの分類スコアとも深く関連すると考えられる.

以上より, 本稿では, 任意のクラスについてのクラス所属確率を推定するために, 注目するクラスと第 1 位のクラスの分類スコアを用いてロジスティック回帰を利用する方法を提案する. 今回, ロジスティック回帰を利用する理由は, Takahashi らの実験において, 安定してよい結果を示したためである.

以下, 次節で関連研究について述べ, 3 節で提案手法を説明した後, 4 節で実験により提案手法の有効性を示す. 最後に 5 節でまとめる.

## 2　関連研究

ここでは, クラス所属確率を推定するために, 注目するクラスの分類スコアだけ用いる方法と複数のクラス

[1] 職業コーディングとは, 自由回答で収集される職業に関するデータ（「仕事の内容」や「従業先事業の種類」など）を総合的に判断し, 約 200 種類ある職業コードの中から該当するコードを 1 つ付ける作業をいう (1995 年 SSM 調査研究会, 1995).

[2] 例えば, ナイーブベイズ分類器のように, 出力する分類スコアが 0 以上 1 以下の範囲である場合には, この値をそのままクラス所属確率であると考えることができる. また, サポートベクターマシン（SVM）のように分類スコアが 0 以上 1 以下の範囲にない場合でも, 最大値と最小値を利用した簡単な式変形により確率値にすることが可能である.

を用いる方法に分けて述べる.

### 2.1 注目するクラスの分類スコアのみを用いる方法

#### 2.1.1 シグモイド関数による方法

Platt 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, これを単調増加で $[0,1]$ 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, クラス所属確率を求める方法を提案した (Platt, 1999) [3]. 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. シグモイド関数による方法は, 質問応答における質問の補充生成 (Jones, 2006) や画像パターンの分類に適用された (Devarakota, 2007). シグモイド関数による方法は, 独立変数 $f$ を多次元に拡張すれば, 複数の分類スコアを扱うことが容易である.

#### 2.1.2 Isotonic regression による方法

Zadrozny らは, データセットによっては Platt の提案する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, 当初は binning による方法 [4]を提案した (Zadrozny and Elkan, 2001). しかし, binning の方法には, 適切なビンの数を決定することができないという問題があったため, 次に, Isotonic regression を利用する方法を提案した (Zadrozny and Elkan, 2002). Isotonic regression においてはデータが順序付けられることが必須の条件である. Zadrozny らは Isotonic regression を実現するために PAV（Pool Adjacent Violators）アルゴリズムを用いた. PAV は事例を分類スコアの順にソートし, それに対応する正（1）誤（0）状況を示す値が逆転しないようにするために, 逆転する箇所があればその区間内にある事例の正誤状況を示す値の平均をとりながら修正していく方法である. Isotonic regression による方法は, Chan らにより語の曖昧性解消タスクにおいて利用された (Chan, 2006). Zadrozny らはまた, 多値分類においても, 注目するクラスとそれ以外のクラスの 2 つに分ければ Isotonic regression による方法が適用できるため, 任意のクラスについてのクラス所属確率を推定できることを示した.

Isotonic regression による方法は, Platt による方法と同様に, ブースティングの各手法に対しては有効であることが示されたが (Mizil, 2005b), 訓練事例が多くない場合は過学習になりやすいという欠点が指摘された (Mizil, 2005a; Jones, 2006). また, 分類スコアのソートを前提とするために, 複数の分類スコアを用いることが困難である.

### 2.2 複数の分類スコアを用いる方法

Takahashi らは, 第 1 位のクラスについてのクラス所属確率を推定する手法として, ロジスティック回帰により直接求める方法と,「正解率表」により間接的に求める方法の 2 つを提案した (Takahashi et als., 2007). いずれの方法も用いる分類スコアの数に制限がない.

#### 2.2.1 ロジスティック回帰による方法

ロジスティック回帰式においては, 独立変数の数を複数に拡張するのは容易である. したがって, 用いる分類スコアの数を簡単に増やすことができるという利点がある. また, 分類スコアを代入すれば, 直接, クラス所属確率を推定できるという利点もある. Takahashi らによる実験では, 第 1 位のクラスについてクラス所属確率を推定する場合は, 第 1 位のクラスの分類スコアだけを用いるより第 2 位のクラスの分類スコアも用いた方がよい結果であった.

#### 2.2.2 正解率表を利用する方法

Takahashi らの提案する正解率表の作成方法は次の通りである. まず, 分類スコアを軸として等間隔の区間に分ける. これを複数の分類スコアそれぞれに対して行ってできた区間をセルと呼ぶ. 次に, セルごとに, 含まれる事例から正解率を求めて正解率表を作成する. さらに, 正解率表の精度を高めるために, 各セルの正解率をそのセルの周囲にあるセルの正解率も利用した平滑化法 (安居院, 1991) を適用して修正する. クラス所属確率を推定したい未知の事例は, この正解率表を利用し, 分類スコアから該当するセルを探してそのセルの正解率を間接的に推定値とする. 実験の結果, 正解率表の平滑化, 特にカバレッジを重みとする移動平均法による平滑化は有効で, セルの区間幅を 0.1 にしたときに最もよい値であった. しかし, 正解率表を利用する方法はロジスティック回帰による方法と異なり, セルの区間幅により値が悪くなるという欠点があり, Zadrozny らの場合と同様に, 適切な区間幅を決定できないという問題がある. また, 用いる分類スコアの数を増やし過ぎるとセルの数が多くなるために, 事例数がゼロのセルが多数出現してしまうという問題もある.

## 3 提案手法

本稿では, 任意のクラスについてのクラス所属確率を, 注目するクラスと第 1 位のクラスの分類スコアを用いてロジスティック回帰により推定する方法を提案する. 注目するクラスを第 $k$ 位とすると, 第 1 位と第 $k$ 位の分類スコア（$f_1$, $f_k$）を用いる場合のロジスティック回帰式は,

$$P_{Log}(f_1, f_k) = \frac{1}{1 + \exp(A_1 f_1 + A_k f_k + B)} \quad (1)$$

で表される. 提案手法の手続きを次に示す.

STEP 1 （1）式に示すロジスティック回帰式におけるパラメータ $A_1$, $A_k$, $B$ を全訓練データを用いて最尤法により推定する. これには第 1 位と第 $k$ 位の分類スコアと正誤状況が各々ペアになったデータが必要で, これを得るには, 全訓練データをさらに訓練データと評価データの 2 つに分けて交差検定を行う.

STEP 2 評価事例における第 1 位と第 $k$ 位の分類スコア $f_1$ と $f_k$ を（1）式に代入してクラス所属確率を直接推定する.

提案手法においては, 各クラスの分類スコアをそのまま用いることができるために, 先行研究のように 2 値分類に分解する手間が不要になる.

## 4 実験

提案手法の有効性を以下の実験により示す.

### 4.1 実験方法

#### 4.1.1 分類器

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 (Kressel, 1999). 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.

[3]実際には, Platt は訓練データにおける過学習を避けるために, 正解を 1 より小さい値, 不正解を 0 より大きい値にした

[4]これは, 全訓練事例の分類スコアをソートし, 等サンプルごとにいくつかのビンに分け, 各ビンごとの正解率を計算しておく. 評価事例の分類スコアから該当するビンを探し, そのビンの正解率を間接的に評価事例のクラス所属確率とする方法である.

#### 4.1.2 データセット

実験に用いたデータセットは, 調査データである JGSS（日本版 General Social Surveys）[5] データセット（日本語）と, 新聞記事である 20newsgroups データセット (Nigam et al., 2000) の 2 つで, 両者は全く異なる性質をもつ. まず, JGSS データセットは, 2000 年から 2003 年まで毎年実施された調査（JGSS-2000, · · · , JGSS-2003）のうちの有職者の職業データ [6]で, 約 200 個の職業コードに分類される（23, 838 サンプル）(高橋他, 2005). 訓練データは通常のコーディングのように, 過去のデータ（JGSS-2000, · · · , JGSS-2002）(20, 066 サンプル）を用い, 評価データは新しいデータ JGSS-2003（3, 772 サンプル）を用いた. 職業データは, 調査終了後に行われた職業コーディングによりすでに職業コードが付与されており, これを正解として扱った. 次に, 20newsgroups データセットは, 記事の内容により 20 個のクラスに分類されている（18,828 サンプル）. 全サンプルの 4/5 を訓練データ, 1/5 を評価データとして用いて 5 分割交差検定を行った.

#### 4.1.3 クラスの順位

今回は, 任意のクラスとして, 第 2 位から第 20 位までのクラスを対象とした. JGSS データセットでは 10%, 20newsgroups データセットではすべてのクラスが該当する. このとき, 各クラスの分類スコアの値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 > ··· >20 位の分類スコア.

#### 4.1.4 評価尺度

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. 本稿では, 負の対数尤度 $L = -\sum_i log(p(x_i))$ を用いた. ここで, $p(x_i)$ は評価事例の予測クラス所属確率である.

### 4.2 実験１：用いる分類スコアの違いによる比較

まず, 注目するクラスのクラス所属確率を推定するために最も有効なクラスはそのクラスの正誤状況に強く関連するクラスであると考え, クラスごとに正誤状況とすべてのクラスの分類スコアとの相関係数を調査した. その結果, 関連が強い（相関係数の絶対値が大きい）クラスは, 注目するクラスが上位の場合は注目するクラスや注目する周囲のクラス, それ以外の場合は第 1 位のクラスであった [7]. これらのクラスがどの程度, 注目するクラスと最も関連が強いまたは 2 番目に強いかを表 1 に示す. 第 1 位のクラスは, 注目するクラスより関連が強い場合が多い. 以下の実験では, 注目するクラス以外の最も有効なクラスの候補として, 第 1 位のクラス, 注目するクラスの直前のクラス, 直後にあるクラスを考える.

図 1 は, JGSS データセットにおいて, 用いる分類スコアを「注目するクラスのみ」, 「第 1 位のクラスを追加」, 「直前のクラスを追加」, 「直後のクラスを追加」した 4 つのケースについて, 第 2 位以下の各クラスにおける負の対数尤度を示す. 第 10 位のクラスを除くすべてのクラスにおいて, 最もよいのは「第 1 位のクラスを追加」で, 最も悪いのは「注目するクラスのみ」であった. 両者の差は, クラスの順位が上位であるほど大きかった. ここで, 追加するクラスの数を複数にした場合にどの程度有効性が増すかを確認するため,

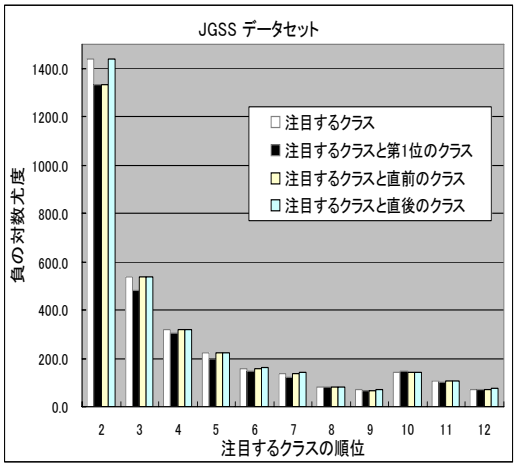


図 1: 用いる分類スコアの違いによる注目するクラスごとの負の対数尤度（JGSS データセット）. 第 13 位以下は第 12 位と同様の傾向を示すため省略した.

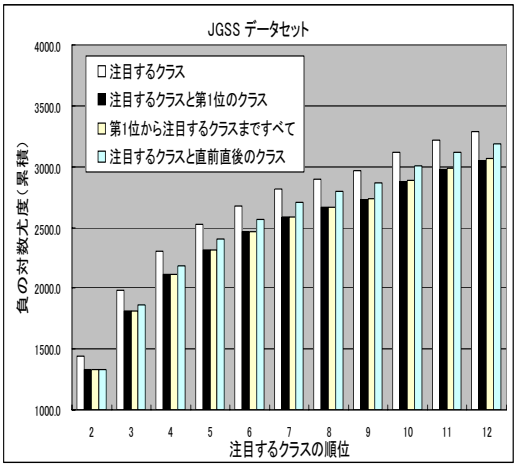


図 2: 用いる分類スコアの違いによる負の対数尤度の累積（JGSS データセット）. 第 13 位以下は第 12 位と同様の傾向を示すため省略した.

上記のケースの自然な拡張として, 「第 1 位から注目するクラスまでのすべてのクラス」「直前と直後のクラスを追加」の 2 つについても同様の実験を行った. その結果においても「第 1 位のクラスを追加」が最もよかった. 次によいのは「第 1 位から注目するクラスまでのすべてのクラス」で, 「注目するクラスのみ」は最も悪かった. この傾向は, 注目するクラス別の結果だけでなく第 20 位までの累積においても同様であった（図 2）. さらに, データセットを 20newsgroups に変えても全く同様の結果を示した. 以上より, ロジスティック回帰によりクラス所属確率を推定する際に, 注目するクラスと第 1 位のクラスの分類スコアを用いることが有効であるといえる.

### 4.3 実験２：正解率表を利用する方法との比較

ここでは, 提案手法を正解率表を利用する方法において最もよい結果を示した方法, すなわちカバレッジを重みとする移動平均法により平滑化を行った正解率表を利用する方法（分類スコアの区間幅 = 0.1）(Takahashi et als., 2007) と比較した. 第 1 位から第 5 位までの

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

[6] 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）から構成される非常に短い文書である.

[7] 注目するクラスの正誤状況と分類スコアの相関係数はすべて 1%で有意であった.

表 1: 注目するクラスの正誤状況と有効なクラス候補の関連度の強さ．データセット A は JGSS, データセット B は 20newsgroups を表す. 注目は注目するクラス, 第 1 位は第 1 位のクラス, 直前（後）は注目するクラスの直前（後）のクラスを表す．表中の数値はそれぞれの有効なクラス候補における該当数を表す.

| データ | 最も強い | | | | 2 番目に強い | | | | 合計 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| セット | 注目 | 第 1 位 | 直前 | 直後 | 注目 | 第 1 位 | 直前 | 直後 | 注目 | 第 1 位 | 直前 | 直後 |
| A | 5 | 12 | 2 | 1 | 2 | 3 | 1 | 1 | 7 | 15 | 3 | 2 |
| B（平均） | 2 | 12 | 0 | 0 | 3 | 1 | 0 | 1 | 5 | 12 | 0 | 1 |
| 合計 | 7 | 24 | 2 | 1 | 5 | 4 | 1 | 2 | 12 | 27 | 3 | 3 |

クラスについて負の対数尤度の累積は, JGSS データセットの場合, 提案手法は 4560.5, 正解率表利用の方法は 4590.2 で, 提案手法の方がよかった. 20newsgroups データセット（平均）の場合は提案手法は 3244.1, 正解率表利用の方法は 3222.6 で, 提案手法の方が悪かった. 分類スコアの区間幅を第 1 位を 0.2, 第 2 位以下を 0.5 に変えると, 正解率表利用の方法は, JGSS データセットの場合に 4556.4, 20newsgroups データセット（平均）の場合に 3201.7 と改善され, 提案手法を上回った.

以上より, 任意のクラスについても, 正解率表を利用する方法は第 1 位のクラスの場合と同様に, 分類スコアの区間幅を適切に定めるとロジスティック回帰による方法を上回る場合がある．しかし, 現在はこの値は実験的にしか決定できないために, ロジスティック回帰による方法の方が安定性がある.

## 5 おわりに

本稿では, 多値分類における任意のクラスについてのクラス所属確率を推定するために, 複数の分類スコア, 特に注目するクラスと第 1 位のクラスの分類スコアを用いて, ロジスティック回帰を利用することを提案した．調査データおよび新聞記事データによる実験の結果, 提案手法は有効性を示した．今後, 精度向上のために複雑な分類スコアの組み合わせを行う場合にも, 提案手法の拡張は容易である.

**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.

### 参考文献

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

安居院猛, 中嶋正之. 画像情報処理. 森北出版1991.

P. N. Bennett. Assessing the Calibration of Naive Bayes's Posterior Estimates. Technical Report CMU-CS-00-155, School of Computer Science, Carnegie Mellon University, pp. 1-8. 2000.

Y.S. Chan and H.T.Ng. Estimating Class Priors in Domain Adaptaion for Word Sense Disambiguation. In *Proceedings of 21th International Conference on Computaional Linguistic and 44th Annual Meeting of the ACL*, pp. 89-96. 2006.

P. R. Devarakota, B. Mirbach and B. Ottersten. Confidence Estimation in Classification Decision: A Method for Detecting Unseen Patterns. In *Proceedings of 6th International Conference on on Advances in Pattern Recognition*, pp. 136-140. 2007.

R. Jones, B. Rey, O. Madani and W. Griner. Generating Query Substitutions. In *Proceedings of International World Wide Web Conference*, pp. 387-396, 2006.

U. Kressel. Pairwise classification and support vector machines. A. J. Smola, (Eds.), *Advances in Kernel Methods Support Vector Learning*, 255-268. The MIT Press, 1999. In B. Schölkopf et al. (Eds.) *Advances in Kernel Methods Support Vector Learning*, pp. 255-268. The MIT Press, 1999.

A. Niculescu-Mizil and R.Caruana. Predicting Good Probabilities With Supervised Learning. In *Proceedings of 22nd International Conference on Machine Learning*, pp. 625-632. 2005.

A. Niculescu-Mizil and R.Caruana. Obtaining Calibrated Probabilities from Boosting. In *Proceedings of 21st International Conference on Uncertainty in Artificial Intelligence*, pp. 413-429. 2005.

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

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

K. Takahashi, H. Takamura, and M. Okumura. Automatic Occupation Coding with Combination of Machine Learning and Hand-Crafted Rules. In *Proceedings of 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 269-279, 2005.

高橋和子, 須山敦, 村山紀文, 高村大也, 奥村学. 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. JGSS 研究論文集 [4] JGSS で見た日本人の意識と行動, pp. 225–242. 2005.

K. Takahashi, H. Takamura, and M. Okumura. Estimation of Class Membership Probabilities in Document Classification. In *Proceedings of 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2007 (to appear).

B. Zadrozny and C. Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayeian classifiers. In *Proceedings of 7th International Conference on Knowledge Discovery and Data Mining*, pp. 609-616. 2001.

B. Zadrozny and C. Elkan. Learning and Making Decisions When Costs and Probabilities are Both Unknown. In *Proceedings of 7th International Conference on Knowledge Discovery and Data Mining*, pp. 204–213. 2001.

B. Zadrozny and C. Elkan. Transformation Classifier Scores into Accurate Multiclass Probability Estimates. In *Proceedings of 8th International Conference on Knowledge Discovery and Data Mining*, pp. 694–699. 2002.