半教師ありブースティングの多値分類への拡張に関する研究

1X15C048-2　阪井 優太
指導教員　後藤 正幸

1　研究背景・目的

少数のラベルありデータに加えて，多数のラベルなしデータを有効活用して分類器の学習を行う半教師あり学習は，広い適用範囲を持つ技術である．このような半教師あり学習の一手法として，データ間の類似度に基づいてラベルなしデータに仮のラベル (以下，仮ラベル) を付与しながら，ブースティングによって二値分類器を構成する SemiBoost [1] が提案されている．

半教師あり学習では，ラベルありデータは何らかの行動の結果として得られているデータであり，実際のデータ分布に従ったランダムサンプリングではない場合がある．このような場合でも，SemiBoost はカテゴリの分布が明瞭に分離されているならば，良い分類精度を示す．一方でこの手法は二値分類を対象としているため，カテゴリ数が 3 以上になった場合，直接適用することができない．そこで，複数の二値分類器を組み合わせて多値分類を行うための枠組みである，Error-Correcting Output Cording 多値判別法 [2] (以下，ECOC 法) の適用が考えられる．ECOC 法は，複数の二値分類器の出力結果を組み合わせ，新規入力データの所属カテゴリを推定する手法である．ECOC 法を適用することで，SemiBoost を多値分類へ拡張できる．しかしながら，SemiBoost を ECOC 法に適用する場合，異なるカテゴリを二つのグループにまとめて二値分類問題が構成されるため，二つの分布の境界が明瞭でなくなり SemiBoost が上手く機能せず，分類精度が悪化する可能性がある．これは，ラベルなしデータに付与される仮ラベルが誤っている割合が高まるという特徴に起因する．そのため ECOC 法の結果から，ラベルなしデータに付与する仮ラベルの信頼性が高いデータのみを活用することができれば，分類精度の向上が見込まれる．

そこで本研究では，SemiBoost に ECOC 法を用いてラベルなしデータで仮ラベルが付与されたラベルなしデータのうち，付与されたラベルの信頼性が高いもののみを仮ラベルありデータとして抽出する．そしてラベルありデータと仮ラベルありデータを用いて，ラベルありデータを増加させて再度分類を行うことで，二値分類を想定した半教師ありモデルを多値分類へ拡張可能とした半教師あり多値分類モデルを提案する．このモデルは，ラベルを持ったデータが実際のデータの分布に比べて偏りのある場合でも分類精度が大幅に悪化しない性質を持つ．最後に，UCI 機械学習レポジトリのデータを提案モデルに適用し，その有効性を示す．

2　準備

2.1　半教師あり学習

本論文では，特徴ベクトル $\boldsymbol{x} \in \mathbb{R}^d$ のカテゴリラベルを K 個のカテゴリ $\mathcal{C} = \{c_1, ..., c_k, ..., c_K\}$ の中から 1 つに決定する多値分類問題を扱う．半教師あり学習では，少数のラベルありデータ集合 $\mathcal{D}_L = \{\boldsymbol{x}_i^L, y_i^L\}_{i=1}^{N_L}$ と多数のラベルなしデータ集合 $\mathcal{D}_U = \{\boldsymbol{x}_i^U, *\}_{i=1}^{N_U}$ を用いて学習を行う．ここで，$\boldsymbol{x}_i^L \in \mathbb{R}^d$，$\boldsymbol{x}_i^U \in \mathbb{R}^d$ とする．ラベルありデータ数を N_L，ラベルなしデータ数を N_U とし，一般に $N_L \ll N_U$ を前提とする．半教師あり学習では，ラベルありデータ $\mathcal{D}_L$ を用いてラベルなしデータ $\boldsymbol{x}_i^U$ のラベル y_i^U を推定し，仮のラベルを付与し，これら全てのデータを用いて学習を行う．カテゴリが未知の入力データ $\tilde{\boldsymbol{x}}$ に対し，複数の二値分類器を組み合わせて所属カテゴリ $\tilde{y} \in \mathcal{C}$ を推定する．

2.2　ECOC 法

ECOC 法は，二値分類器の組み合わせにより多値分類を実現することのできる手法である．このとき各二値分類器の構成を符号表と呼ばれる $\{+1, -1\}$ の二値で表される数値表により表現する．いま符号表を $\boldsymbol{W}$，二値分類器の個数を R とすると $\boldsymbol{W}$ は $K \times R$ 行列となる．符号表 $\boldsymbol{W}$ の各列ベクトルは二値分類器の構成を表現しており，要素が $+1$ のカテゴリ集合と要素が -1 のカテゴリ集合を二値分類すると解釈できる．また，符号表 $\boldsymbol{W}$ の k 行目の行ベクトルをカテゴリ c_k の符号語と呼び $\boldsymbol{W}_{c_k}$ と表現する．以下では，本研究で使用するハミング距離を用いた複号法を説明する．

新規入力データ $\boldsymbol{x}$ に対する $r(1 \leq r \leq R)$ 番目の二値分類器の出力を $g_r(\boldsymbol{x}) \in \{+1, -1\}$，符号語 $\boldsymbol{W}_{c_k}$ の r 番目の値を $W_{c_k \cdot r}$ と定義する．このとき $\boldsymbol{x}$ は，符号語 $\boldsymbol{W}_{c_k}$ と分類器の出力 $\boldsymbol{g} = (g_1(\boldsymbol{x}), ..., g_R(\boldsymbol{x}))$ のハミング距離を求め，最も近いカテゴリに分類できる．

2.3　SemiBoost

SemiBoost は半教師あり学習のブースティングを用いた手法であり，カテゴリ数 $K = 2$ のときのみ適用可能な手法である．その際，c_1 のラベルを $+1$，c_2 のラベルを -1 として二値分類を行う．データ間の類似度を利用し，ラベルなしデータに仮のラベルを付与することで分類器を学習する．以前の分類器で付与されたラベルが誤っている可能性が高いデータを正しく分類できるように新たな分類器を構築することを繰り返す．このように新たな分類器を逐次的に構築し，構築されたすべての分類器の結果をアンサンブルすることにより分類を行う．この手法は，二つのデータの分布が明瞭に分かれているとき，良い分類精度が得られる．SemiBoost のアルゴリズムを以下に示す．

Step1) 全学習データ間の類似度をガウスカーネルを用いて求める．

Step2) アンサンブル分類器 $H(\boldsymbol{x})$ を初期化し，$t = 1$ とする．

Step3) 全てのラベルなしデータ $\boldsymbol{x}_i^U$ の仮ラベル y_i^U を式 (4) を用いて予測する．

Step4) 誤分類されている可能性の高いラベルなしデータ $\boldsymbol{x}_i^U$ を学習に用いるデータとして選択する．

Step5) Step4 で選んだラベルなしデータ $\boldsymbol{x}_i^U$ を用いて弱分類器 h_t を学習し，分類誤差 ε_t を用いて分類器の重みを決定する．ここで $h_{t,i}$ は弱分類器 h_t による二値分類結果である．

$$\alpha_t = \frac{1}{4} \ln \left(\frac{1 - \varepsilon_t}{\varepsilon_t} \right) \tag{1}$$

$$\varepsilon_t = \frac{\sum_{i=1}^{N_U} p_i \delta(h_{t,i}, -1) + \sum_{i=1}^{N_U} q_i \delta(h_{t,i}, 1)}{\sum_i (p_i + q_i)} \tag{2}$$

Step6) Step5 で学習した弱分類器 h_t に α_t で重みづけし，アンサンブル分類器 $H(\boldsymbol{x})$ に結合し更新する．

$$H(\boldsymbol{x}) \leftarrow H(\boldsymbol{x}) + \alpha_t h_t(\boldsymbol{x}) \tag{3}$$

Step7) $t \leftarrow t + 1$ とし，Step4 へ．$t = T$ のとき終了．

Step2 のアンサンブル分類器 $H(\boldsymbol{x})$ は，最終的に弱分類器をまとめる分類器である．Step3 において，ラベルなしデータ $\boldsymbol{x}_i^U$ がカテゴリ c_1 に所属する可能性が高いときに大きな値をとる p_i と，カテゴリ c_2 に所属する可能性が高いときに大きな値をとる q_i を用いて，ラベルなしデータがどちらのラベルに所属するかを式 (4) により求める．

$$y_i^U = \mathrm{sign}(p_i - q_i) \tag{4}$$

ただし，$\mathrm{sign}(a)$ は，$a \geq 0$ のとき $+1$，$a < 0$ のとき -1 をとる関数である．

また，このとき用いる p_i, q_i は式 (5)(6) により求める．

$$p_i = \sum_{j=1}^{N_L} s_{ij}^{UL} e^{-2H_i} \delta(y_j^L, 1) + \frac{\gamma}{2} \sum_{j=1}^{N_U} s_{ij}^{UU} e^{H_j - H_i} \tag{5}$$

$$q_i = \sum_{j=1}^{N_L} s_{ij}^{UL} e^{2H_i} \delta(y_j^L, -1) + \frac{\gamma}{2} \sum_{j=1}^{N_U} s_{ij}^{UU} e^{H_i - H_j} \tag{6}$$

$\delta(a, b)$ は $a = b$ のとき $+1$，それ以外は 0 となるインジケータ関数であり，$\gamma (> 0)$ は定数である．$s_{ij}^{UL} (\geq 0)$ はラベル

なしデータ $\boldsymbol{x}_i^U$ とラベルありデータ $\boldsymbol{x}_j^L$ の類似度，$s_{ij}^{UU}(\geq 0)$ はラベルなしデータ $\boldsymbol{x}_i^U$ と他のラベルなしデータ $\boldsymbol{x}_j^U$ の類似度とする．H_i は，アンサンブル分類器に $\boldsymbol{x}_i^U$ を入力したときの出力値である．

Step4 において，ラベルなしデータ $\boldsymbol{x}_i^U$ の中から仮ラベルを付与し，弱分類器の学習に用いるデータを選択する．ここで，$P(\boldsymbol{x}_i^U)$ はラベルなしデータ $\boldsymbol{x}_i^U$ が学習に用いられる確率であり，以下の式 (7) により求める．

$$P(\boldsymbol{x}_i^U) = \frac{|p_i - q_i|}{\sum_{i=1}^{n_U} |p_i - q_i|} \tag{7}$$

3 提案手法

3.1 SemiBoost の多値分類への拡張

提案モデルでは SemiBoost を二値分類から多値分類へ拡張する手法を提案する．ECOC 法を用いて SemiBoost を多値分類へ拡張することを考えた場合，分類器の出力を確率値とすることでカテゴリを直接推定できる．しかしながら，半教師あり学習ではラベルありデータが相対的に少なく，ラベルありデータの分布と実際のカテゴリ分布に差異がある場合がある．このとき SemiBoost を ECOC 法に直接適用すると，異なるカテゴリを二つにまとめて二値分類するため，分類する二つの分布が空間上離れた位置関係にはなくなり，SemiBoost が上手く機能せず分類精度が悪化する可能性がある．これは，正しく仮ラベルが付与されるラベルなしデータが少なくなり，誤った仮ラベルの付けられたデータの割合が多くなることに起因する．そこで本研究では ECOC 法に SemiBoost を適用し，ラベルなしデータに付与された仮ラベルの信頼性が高いデータのみを仮ラベルありデータとして抽出する．そしてラベルありデータと信頼度の高い仮ラベルありデータを学習データとして教師あり学習を行う．提案手法では，ECOC 法に SemiBoost を適用し，ラベルなしデータのうち信頼性の高いデータのみを抽出するため，ECOC 法を直接適用したときの問題点を克服し，多値分類への拡張を行うことが可能となる．

提案手法では，ECOC 法の一つである one-vs-the rest 法を用いて多値分類を行う．あるカテゴリの符号語 $\boldsymbol{W}_{c_k}$ と SemiBoost による二値分類の出力ベクトル $\boldsymbol{g}$ のハミング距離が 0 となるラベルなしデータのみにカテゴリ c_k のラベルを付与し，このデータを仮ラベルありデータとする．その結果，複数のカテゴリに所属すると判断された信頼度の低いデータは誤った分類を引き起こす要因となる可能性があるため取り除き，信頼度の高いラベルのみを学習データに追加することが可能となる．次に，ラベルありデータと仮ラベルありデータを用いて教師あり学習を行う．ラベルありデータのみのときと比較して学習に使用するデータの数が増加することで教師あり学習の精度の向上が期待される．

3.2 提案アルゴリズム

Step1) 学習データを用い，one-vs-the rest 法に従って構成された各二値分類器を，SemiBoost により学習する．

Step2) Step1 の出力結果から，ラベルなしデータと各カテゴリとのハミング距離を算出する．

Step3) ラベルなしデータのうち，ハミング距離が 0 となるカテゴリをもつデータにそのカテゴリを付与して仮ラベルありデータとする．

Step4) Step3 で得られた仮ラベルありデータとラベルありデータを用いて，教師あり学習を行いテストデータの所属カテゴリを推定する．

4 実験

提案モデルの有効性を示すために，ベンチマークデータセットを用いた分類実験を行う．

4.1 実験条件

データセットとして，UCI 機械学習レポジトリの drug consumption から，Cannabis と Nicotine の 2 種類を用いる．1885 件のデータのうち 1508 件を学習データ，残りの 377 件をテストデータとした．データセットのカテゴリ数は 7，データの次元数は $d = 12$ である．ラベルありデータは 5.2 節の方法より N_L 個抽出し，各実験でラベルありデータを再度抽出するものとした．実験結果は 5 分割交差検定を 5 回繰り返し行い，その平均を用いることとした．提案手法では，$K = 7$ の one-vs-the rest 法を用いる．提案手法と比較手法で用いる教師あり分類器は RandomForest (RF), 1 対他 SVM とする．比較手法としてラベルありデータのみを用いた教師あり分類 (ラベルありのみ)，直接 SemiBoost を ECOC 法に適用した分類 (直接分類した場合) を用いた．また参考値として，ラベルありとラベルなしデータ全てにラベルが振られたときの教師あり分類 (全データラベルあり)，ランダムにデータを分類したとき (ランダム) の分類誤り率も示す．これらは，達成可能な誤り率の下限と上限を意味する事前実験により，SemiBoost の分類器数 $T = 10$，ラベルありデータ数 $N_L = 210(30 \times 7)$ 個，偏りのあるラベルありデータ割合 $\theta = 70\%$ とした．また，SemiBoost の分類器には SVM を用いた．評価指標として分類誤り率を用いる．

4.2 ラベルありデータの抽出

本研究ではラベルありデータに偏りがある場合についても，提案手法の有効性を示すことを目的とする．そのためラベルありデータは，各カテゴリの重心ベクトルを求め，そこから最も遠い点を基準に，その近傍初期ラベルありデータのうち $\theta\%$ の点をサンプリングすることで生成する．その後，ラベルありデータのうち残りの $(100 - \theta)\%$ の点をランダムに抽出し，これらをラベルありデータとする．

4.3 実験結果と考察

実験結果を図 1,2 に示す．2 つのデータセットにおいて，ラベルありデータのみを用いた場合，ECOC 法で直接分類した場合と比較し，提案手法の分類誤り率が優れている．下限である全データラベルありにして分類を行った場合の精度とは 10% 以上の差があるものの，ラベルありのみの比較手法より分類誤り率が低く，提案手法の有効性が確認できた．

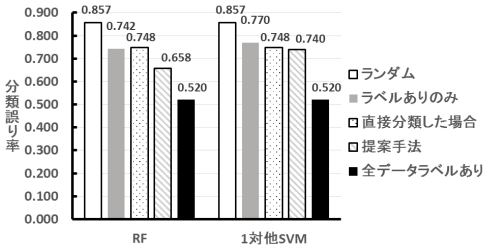

図 1：Canabis のデータを用いた実験結果

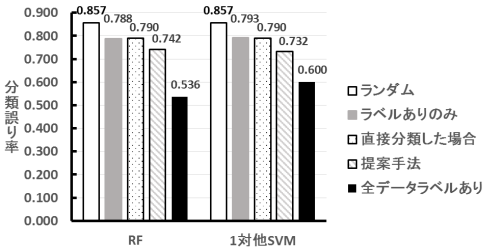

図 2：Nicotine のデータを用いた実験結果

5 まとめと今後の課題

本研究では，SemiBoost の多値分類への拡張モデルを提案した．また，偏りのあるデータに対して提案モデルを適用することでその有効性を示した．

今後の課題としては，全てのデータにラベルがある場合の精度へ近づけることや，カテゴリ数が未知の場合にも対応した半教師あり多値分類手法への拡張などが挙げられる．

参考文献

[1] P. Mallapragada, R. Jin, K. Jain, "SemiBoost: Boosting for Semi-supervised Learning," *IEEE Trans, Pattern Analysis and Machine Intelligence*, Vol. 31, No. 11, pp. 2000–2014, 2009.

[2] T. Dietterich, G. Bakiri, "Solving Multiclass Learning Problems via Error-Correcting Output Codes," *Artif. Intell.*, Vol. 2, pp. 263–286, 1995.