クラス所属確率を用いたアンサンブル学習

高橋 和子
敬愛大学 国際学部
takak@u-keiai.ac.jp

1　はじめに

本稿では, 多値分類におけるサポートベクターマシン（SVM）の分類精度を高めるアンサンブル学習として, 事例ごとにクラス所属確率を用いて適切な分類器を選択し, この分類器の予測クラスを最終決定とする方法を提案する.

機械学習においては, 複数の分類器を組み合わせ, それらの結果を統合することで個々の分類器よりも予測精度を上げるアンサンブル学習が有効な場合が多く (Sebastiani, 2002), 代表的な方法としてバギングやブースティングがある. バギングは, リサンプリングにより元のデータセットと同じサイズのデータセットを複数個作成し, 各データセットに同じアルゴリズムを適用してバリアンスの異なる複数の分類器を構築する. 個々の分類器による予測結果に対して, カテゴリ型の場合には多数決により, 連続値である回帰問題の場合には平均値や中央値により最終決定を行う (Breiman, 1996). また, ブースティングは, 逐次的に事例の重みを変化させながら分類器を構築していき, 個々の分類器による予測結果に異なる重み付けをして最終決定を行う (元田他, 2006). しかし, これらのアンサンブル学習は, 文書分類で多用される SVM (Joachims, 1998) に対しては有効性が高くないことが指摘されている. これは, SVM のような高バイアスのモデルは, バイアス－バリアンス理論 (Breiman, 1996) [1]におけるバリアンスの占める要素がもともと少ないために, 低バイアスのモデルほどにはリサンプリングによる効果が期待できないこと (Torii and Liu, 2007; 神嶌他, 2008) や, また, ブースティングで必要な重みを SVM では直接的に反映させることができないこと (Li et al, 2008) が理由である.

そこで, 高橋（2009a）では観点を変えて, 事例ごとに複数の分類器の中から正解の可能性が最も高い分類器を選択し, この分類器が予測したクラスを最終決定とする方法を提案した. これは, 複数の分類器における各事例の正解状況を比較すると, 全クラスについての分類精度（分類器が正解した事例数を全事例で割った値）が最も高い分類器が不正解の事例に対して, もし, 分類精度がより低い分類器が正解する場合が観察されるために, 事例ごとに正解の可能性が最も高い分類器を選択することができれば, 全体として正解事例数が増え, 分類精度が向上すると考えたことによる [2].

この方法においては, 多岐にわたる事例に対してそれぞれが正解となる可能性を高めるためにできる限り多様な分類器を構築しておくことおよび, 正解の可能性が高い最適な分類器をうまく選択することの２つが重要である. 高橋（2009a,2009b）では, まず, 多様な分類器の構築のためには, リサンプリングではなく素性選択の変化が有効であると考え, 人手により用いる素性の種類を変化させた分類器を構築した. 次に, 最適な分類器の選択の判断基準としては, 簡単には予測クラスごとに出力されるスコア（分類スコア）を用いる方法が考えられるが, 事例が予測クラスに所属する確率（確信度）すなわちクラス所属確率 (Platt, 1999; Zadrozny and Elkan, 2002; Niculescu-Mizil and Caruana, 2005) の大きさにより判断する方が自然であると考えた. ここで, クラス所属確率の推定は, 多値分類において複数個出力される分類スコアを用いることが有効であることが実験的に示されている (高橋他, 2008; Takahashi et al, 2008). 最適な分類器の選択としては, 他にも多数決による方法があるが, これはバギングに該当する. 高橋（2009a,2009b）において, これらの 3 つの方法を実験により比較した結果, クラス所属確率を用いる方法の有効性が示唆された. しかし, 適用した分類タスクがやや特殊であったため [3], 得られた結果を一般化するまでには至らなかった.

本稿では, この手法の有効性をより一般的に示すため, 3 つの方法を性質の異なる 2 種類のデータセットに適用して比較する. また, 分類器の構築方法については, 素性の選択をさまざまに変化させて構築しても, 正誤の出現状況に注目すると, 2 つのグループにまとまっていて多様性があるとはいえなかったため (高橋, 2009b), 本稿ではリサンプリングにより構築する.

以下, 次節で関連研究について述べた後, 3 節で提案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

関連研究として, 神嶌他（2008）および Torii and Liu（2007）について述べる.

神嶌他（2008）はバギングを改造した方法で, より多様な事例が多数含まれると考えられる野生データ（整合性のある概念に基づいてラベル付けされた事例事例とそうではない事例が混在する）に注目したリサンプリングにより分類器を構築し, 各分類器の正解率を重みとする多数決によりクラスを決定する方法を提案した. Torii and Liu（2007）は, SVM においてはバギングが有効ではないとして, bag-of-words に対する情報利得により, 利用する素性を上位からのランキングにより変化させることで多様な分類器を構築し, 分類スコアの和が大きいクラスに決定する方法を提案した.

3　提案手法

3.1　提案手法の手順

提案手法の手順は, 次の通りである.

STEP1　リサンプリングにより複数の分類器を構築する

STEP2　個々の分類器ごとに未知の事例に対するクラスを予測する

STEP3　個々の分類器ごとに未知の事例に対する予測クラスのクラス所属確率を推定し, 最も大きな値をも

[1]バイアス－バリアンス理論においては, 誤差をバイアス（予測に用いたモデルに由来する誤差）, バリアンス（学習に用いた訓練データのサンプリングの揺らぎに由来する誤差）, 基本的に減らせない誤差の 3 つに分解できるとする.

[2]実際に, 高橋（2009b）における実験では, すべての事例で最適な分類器を選択できれば, 計算上は分類精度が 73.9% から 80.5% に 6.6% 向上した.

[3]自由回答と選択回答から構成される職業に関する調査データを約 400 個の国際標準職業（ISCO）コードに分類するタスクであった（4.1 節参照）.

つ分類器の予測クラスを最終決定とする

3.2 クラス所属確率の推定方法

第 1 位に予測されたクラスに対するクラス所属確率の推定方法は, パラメトリックな方法とノンパラメトリックな方法の 2 つがある (高橋他, 2008; Takahashi et al, 2008).

- ロジスティック回帰式を利用（パラメトリックな方法）
  第 1 位から第 3 位に予測されたクラスの分類スコアの利用が有効
- 「正解率表」を作成・利用（ノンパラメトリックな方法）
  第 1 位と第 2 位に予測されたクラスの分類スコアの利用が有効

ロジスティック回帰式を利用して推定する方法では, 第 1 位から第 3 位に予測されたクラスの分類スコア（f_1, f_2, f_3）を, 次のロジスティック回帰式

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

に代入して直接計算する. ただし, (1) 式におけるパラメタ（4 個）を最尤法により推定するために, 訓練データをさらに訓練データと評価データに分けて学習しておく[4].

「正解率表」を作成・利用して推定する方法では, あらかじめ正解率表を作成しておく. その際, ロジスティック回帰式におけるパラメタ推定の場合と同様に, あらかじめ訓練データを訓練データと評価データに分割して学習を行い, 評価データの正誤状況と第 1 位および第 2 位に予測されたクラスの分類スコアを調査しておく. 各分類スコアを等間隔（例えば 0.1）に分け, 各区間（セル）ごとに正解率（各セル内の正解事例数／各セル内の全事例数）を算出したものが正解率表で, クラス所属確率法の推定は, 評価事例の分類スコアから正解率表内の該当セルを探し, そのセル内の正解率を間接的に用いる. 正解率表を用いる方法は, 分類スコアの区間設定が適切であればロジスティック回帰を用いる方法より良好な結果が得られたが, 安定性の問題が存在する (高橋他, 2008).

なお, クラス所属確率を事後確率と考えるためには, すべてのクラスに対してそれぞれのクラス所属確率を求めて和が 1 になるように正規化する必要があるが, 今回は正規化までは行っていない [5].

4 実験と考察

提案手法, 分類スコアの最も大きな値の分類器を選択する方法（以下, 「分類スコア法」と略す）, 多数決により分類器を選択する方法（以下, 「多数決法」と略す）の 3 つの方法を性質の異なる 2 種類のデータセットに適用して結果を比較し, 提案手法の有効性を調査した. ここで, 多値分類の場合には, 分類スコアはクラスの数だけ出力されるが, 今回は簡単のため, 分類スコア法においては第 1 位に予測されたクラスのもののみを対象とした. また, 提案手法においても第 1 位に予測されたクラスの推定値のみを対象とした.

4.1 実験設定

データセットとタスク

用いたデータセットは,「2005 年社会階層と社会移動に関する全国調査」(2005 年 SSM 調査) (2005 年 SSM 調査研究会, 2006) により収集されたデータのうち職業に関するデータおよび 20Newsgroups データセット (Asuncion and Newman, 2007) [6]の 2 つである.

職業データ（16,089 サンプル）のタスクは, 390 個の国際標準職業分類（ISCO）コード (Bureau of Statistics, 2001) に分類するもので, 調査終了後の作業により, すべての事例に対して, 国内標準職業分類である SSM コード (2005 年 SSM 調査研究会, 2007) と ISCO コードの 2 種類の職業コードが各 1 個ずつ付与されている. 本稿ではこの ISCO コードを正解として扱った. 素性としては, 高橋（2009b）において最も分類精度が高かった分類器で用いられたものを用いた [7]. この分類期に SVM を単独で適用した場合の分類精度は 73.9% で, 本稿ではこの値をベースラインとした. 訓練データと評価データの分割は, 10 分割交差検定により行った.

20Newsgroups データセット（18,828 サンプル）のタスクは, ネットニュース記事を 20 個のディスカッショングループ・カテゴリに分類するもので, 素性としては, ネットニュース記事に出現する単語 unigram を用いた. SVM を単独で適用した場合の分類精度は 87.3% で, 本稿ではこの値をベースラインとした. 訓練データと評価データの分割は, 5 分割交差検定により行った.

クラス所属確率を推定するためのロジスティック回帰式におけるパラメタ推定や正解率表作成のためには, 各訓練データそれぞれに対してさらに 10 分割交差検定（20Newsgroups データセットでは 5 分割交差検定）を行って訓練データと評価データに分割し, この評価データにおける正解／不正解の状況（2 値）を用いた.

分類器と評価尺度

SVM は本来 2 値分類器であるため, one-versus-rest 法を用いて多値分類器に拡張した (kressel, 1999). カーネル関数は線型カーネルを用いた. 分類器はリサンプリングにより最大 25 個まで構築した.

今回の実験においては, クラス所属確率の推定はロジスティック回帰式を利用する方法を用い, データの分布状況への依存度が高いと考えられる正解率表を作成・利用する方法は用いなかった. ただし, 高橋（2009b）における結果を示す際には, 正解率表を作成・利用する方法も用いた. また, 評価尺度としては, 分類精度（全クラスのマクロ平均）を用いた.

4.2 実験結果と考察

リサンプリングによる場合

職業データにおける結果を表 1, 20Newsgroups データセットにおける結果を表 2 に示す. 表中の太字は, 同じ

[4]簡単のため, 分類スコアが 1 個の場合におけるパラメタの推定方法を以下に示す. 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.

$$L(A, B) = \prod_{Y^i=1} P_{Log}(f^i; A, B) \times \prod_{Y^i=0} [1 - P_{Log}(f^i; A, B)]. \quad (2)$$

.

[5]高橋他（2008）の実験においては, 正規化した場合の方がややよい結果であったことが報告されている.

[6]http://people.csail.mit.edu/jrennie/20Newsgroups/

[7]職業データである「仕事の内容」（自由回答）,「従業先事業の種類」（自由回答）,「従業上の地位と役職」（13 種類の選択回答）に,「学歴」（6 種類の選択回答）,「性別」（2 種類の選択回答）,「付与ずみの SSM コード」（約 200 種類）を追加したものである.

表 1: リサンプリングによる場合（職業データ）
分類器の選択方法別分類精度（全クラスの平均）
baseline : 0.7392

分類器数	提案手法	分類スコア法	多数決法
3	0.7395	**0.7423**	0.7257
9	**0.7528***	0.7456	0.7432
1 5	**0.7532***	0.7463	0.7459
2 1	**0.7528***	0.7448	0.7468

表 2: リサンプリングによる場合（20Newsgroups）
分類器の選択方法別分類精度（全クラスの平均）
baseline : 0.8730

分類器数	提案手法	分類スコア法	多数決法
9	**0.8753**	0.8733	0.8719
1 3	**0.8760**	0.8730	0.8750
1 5	**0.8755**	0.8677	0.8730
2 1	0.8775	0.8728	**0.8805**
2 5	0.8778	0.8736	**0.8816**

表 3: 素性選択の変化による場合（職業データ）
分類器の選択方法別分類精度（全クラスの平均）
baseline : 0.7392

提案手法	提案手法*	分類スコア法	多数決法
0.7415	-	0.7269	0.7361
0.7410	**0.7460**	0.7310	0.7380

分類器数の中で最も分類精度の高かった値を示す（以下, 同様である）. また, 数値右の*印は, 単独の分類器を有意に上回っていることを示す（有意水準 1%）.

表 1 より, 職業データにおいては, 多数決法の一部（分類器が 3 個の場合）を除き, すべての手法がベースラインを上回った. 特に, 提案手法は 3 つの方法の中で分類精度が最も高く, ベースラインとは最大で 1.4% の有意な差があった. 分類器の数を増やすにつれて多数決法では分類精度が上昇したのに対し, 提案手法では 15 個までは上昇したが, それ以降は変化しないため, 提案手法と多数決法における分類精度は, 分類器の数が少ない場合（3 個, 9 個）には有意な差があったが（有意水準 10%）, 多い場合には（15 個以上）有意な差が認められなかった. ここで, 提案手法と多数決法における分類精度と分類器の数との相関係数は, それぞれ 0.774 と 0.859 であった. なお, 分類スコア法は, 分類精度と分類器の数との関係が最も低く（相関係数 0.607）, 分類器を増やしても分類精度の値の変化は小さかったため, 分類器の数が増えるにつれて順位が低下した.

表 2 より, 20Newsgroups データセットにおいては, 多数決法は一部（分類器が 9 個の場合）でベースラインを下回り, 分類スコア法もベースラインを下回るかほぼ同様の値であったのに対し, 提案手法はつねにベースラインを上回った. ただし, 職業データの場合とは異なり, 有意な差は認められなかった. また, 20Newsgroups データセットにおいては, 提案手法は多数決法と同様に, 分類器の数を増やすにつれて分類精度も上昇したが, その程度は緩やかであった. ここで, 提案手法と多数決法における分類精度と分類器の数との相関係数は, それぞれ 0.942 と 0.945 であった. 分類器が 15 個以下の場合は提案手法の分類精度が最も高く, ベースラインを最大で 0.2% 上回ったが, 21 個以上の場合は多数決法の方が高かった. ただし, 両手法とも上昇の程度は小さく, また手法間に有意な差は認められなかった. なお, 分類スコア法は, 職業データの場合と同様に, 分類器の数に関係なくほぼ一定の値であった（相関係数 0.164）.

表 1, 2 より, 提案手法は分類器の数が非常に少ない場合でも, 分類器を単独で適用する場合の分類精度を上回ることがわかった. ただし, 表 1 から明らかなように, 他の方法より有効であるためには, ある程度（今回は 9 個）以上の分類器が必要であった. また, 表 2 で示されるように, 分類器の数が多くなると多数決法に劣る傾向がみられたが, 表 2 においてはデータセットの分類精度がもともと高く, 表 1 の場合と約 15% の差がある. この点を考慮すると, 提案手法は, 分類が困難なタスクに対する有効性が高いと考えられる.

素性選択の変化による場合

表 3 は, 職業データに対して人手により素性の選択を変化させ [8], 分類器を 5 個, 8 個構築した場合の結果である (高橋, 2009b). 上段が 5 個の場合, 下段が 8 個の場合を示す. 表中の提案手法*は, クラス所属確率の推定を正解率表を作成・利用した場合で, 正解率表の区間幅は, 高橋他（2008）にしたがって 0.1 とした.

表 3 においては, 構築できた分類器の数が少なく, 多数決法はベースラインを下回ったが, 提案手法はつねにベースラインを上回った. 特に, クラス所属確率の推定に正解率表を作成・利用する方法は最もよく, ベースラインを 0.7% 上昇した. 分類スコア法は, ここでも順位が最も低かった.

表 3 より, 提案手法は, 素性選択の変化による場合において分類器の数が少ない場合でも, 分類器を単独で適用する場合の分類精度を上回ることが確認できた. なお, 表 3 を表 1 において分類器の数がほぼ等しい場合の結果と比較すると, 表 1 の方がよかった. これより, 高橋（2009a,2009b）における予想と異なり, 素性選択を変化させた場合も分類期の数が少なければ効果的でないことが確認できた.

以上に示したように, 提案手法は多数決法に比較すると, 困難なタスクや構築できる分類器の数が少ない場合に特に有効性を発揮すると考えられる. したがって, 分類器の構築に時間を要する大容量のデータセットや分類が困難なデータセットに SVM を適用する場合の有効な手法として期待できる. しかし, 今回の実験では, 分類精度の上昇が, 計算上見込める値（6.6%）にはるかにおよばなかったため, クラス所属確率の推定精度を高めることも含めて, 最適な分類器を発見する方法をさらに検討する必要がある.

5 おわりに

本稿では, バギングやブースティングの効果が期待しにくい SVM における分類精度を高めるために, 各事例ごとに最適な分類器としてクラス所属確率が最も高い分類器を選択するアンサンブル学習を提案し, 性質の異なる 2 つのデータセットにより実験を行った. その結果, 提案手法は, 分類が困難なタスクにおいて, 構築できる分類器の数が多くない場合に特に有効であった. しかし, 分類精度がもともと高いタスクにおいて分類器の数

[8]注 6 に挙げた素性のうち, 職業データを共通とし, これに「学歴」,「性別」,「付与ずみの SSM コード」のそれぞれを追加する／しないとして変化させた (高橋, 2009b).

が多い場合には多数決法に劣り, 有効性が高いとはいえなかった.

提案手法は SVM に対するアンサンブル学習として考えられたものであるが, SVM に限定される手法ではない. 今後の課題は, 提案手法の核である最適な分類器の選択方法をさらに検討することおよび, SVM 以外の機械学習に対しても提案手法を適用し有効性を確認することである.

謝辞 2005 年 SSM 調査データの利用に関して, 2005 年 SSM 調査研究会の許可を得た.

References

2005 年 SSM 調査研究会. 2007. 2005 年 SSM 日本調査コード・ブック.

2005 年 SSM 調査研究会. 2006. 2005 年 SSM 調査　日本・韓国・台湾調査票.

A. Asuncion and D. J. Newman. UCI Machine Learning Repository. 2007.

Bureau of Statistics; International Labour Office. 2001. Coding Occupation and Industry. Bureau of Statistics; International Labour Office.

L. Brieman. 1996. Bagging predictors. In *Machine Learning* 24(2), pp. 123–140.

Y-S. Dong and K-S. Han. 2004. A comparison of several ensemble methods for text ctegorization. In *Proceedings of IEEE 2004 International Conference on Services Computing (SCC 2004)*, pp. 419–422.

T. Joachims. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proceedings of the European Conference on Machine Learning*, pp. 137–142.

神嶌敏弘, 濱崎雅弘, 赤穂昭太郎. 2008. 飼い慣らし－飼育・野生混在データからの学習. 第 22 回人工知能学会全国大会論文集.

U. Kressel. 1999. Pairwise classification and support vector machines. In *Advances in Kernel Methods Support Vector Learning*, pp. 255–268. MIT Press.

工藤拓, 松本裕治. 2002. Support Vector Machine を用いた Chunk 同定. 自然言語処理　Vol.19 No.5, pp.3–22.

X. Li, L. Wang,and E. Sung. 2008. AdaBoost with SVM-based component classifiers. In *Engineering Applications of Artificial Intelligence* 21(5) pp.785–795.

松田博義, 滝口哲也, 有木康雄. 2007. 弱識別器に SVM を用いた AdaBoost の検討. 信学技報　Vol.107 No.405, pp.109–114.

元田浩, 津本周作, 山口高平, 沼尾正行. 2006. データマイニングの基礎. オーム社.

A. Niculescu-Mizil and R. Caruana. 2005. Predicting Good Probabilities With Supervised Learning. In *Proceedings of the 22nd International Conference on Machine Learning (ICML'05)*, pp. 625–6323.

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.

F. Sebastiani. 2008. Machine Learning Automated Text Categorization. In *ACM Computing Surveys* 34(1), pp.1–47.

高橋和子. 2008. 機械学習による ISCO 自動コーディング. 2005 年 SSM 調査シリーズ１２　社会調査における測定と分析をめぐる諸問題, pp.47–68.

高橋和子, 高村大也, 奥村学. 2008. 複数の分類スコアを用いたクラス所属確率の推定. 自然言語処理　Vol.15 No.2, pp. 3–38.

高橋和子. 2009a. クラス所属確率を用いた事例ごとの分類器選択. 第 16 回言語処理学会年次大会発表論文集, pp. 709–712.

高橋和子. 2009b. サポートベクターマシンにおけるアンサンブル学習の提案. 第 23 回人工知能学会全国大会論文集.

K. Takahashi, H. Takamura, and M. Okumura. 2008. Direct estimation of class membership probabilities for multiclass classification using multiple scores. In *Knowl Inf Syst* 19(2), pp.185–210. Springer London.

D. Tao, X. Tang, X. Li, and X. Wu. 2006. Asymmetric Bagging and Random Subspace for Support Vector Machines-Based Relevamce Feedback in Image Retrieval. In *The IEEE Transactions on Pattern analysis and machine intelligence (TPAMI)* 28(7), pp.1088–1099.

M. Torii and H. Liu. 2007. Classifier ensemble for biomedical document retrieval. In *Proceedings of the Second International Symposium on Languages in Biology and Medicine (LBM 2007)*.

X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLarchlan, A. Ng and B.Liu, P. S. Yu, Z-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg. 2008. Top 10 algorithms in data mining. In *Knowl Inf Syst* 14, pp.1–37. Springer London.

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.