1A1-3

クラス所属確率を利用したアンサンブル学習

Ensemble Learning with Support Vector Machines by Using Class Membership Probabilities

高橋和子
Kazuko Takahashi

敬愛大学
Keiai university

We propose an ensemble learning with multi-class support vector machines by using class membership probabilities. In the proposed method, we decide the final class of each sample predicted by the classifier selected by evaluation of class membership probabilities. Experiments showed that the proposed method performed better than both a voting method and a method using directly scores from classifiers in precision.

1. はじめに

本稿では, 多値分類のサポートベクターマシン（SVM）における分類精度を向上させるアンサンブル学習として, 複数の分類器からクラス所属確率を用いて事例ごとに適切な分類器を選択し, 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 02], 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 96]. 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 06]. しかし, これらのアンサンブル学習は, 文書分類をはじめ多くの分類問題に適用される SVM [Joachims 98] におい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 96] *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 07, 神嶌他 08] や, また, SVM ではブースティングで必要な重みを直接的に反映させることができないこと [Li et al. 08] が理由である.

そこで, [高橋 09] では観点を変えて, 複数の分類器の中から, 事例ごとに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, 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, もし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のではないかと考えたためである. *2.

この方法においては, 多岐にわたる事例の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することおよび, 正解の可能性が高い最適な分類器をうまく選択できることの 2 つが重要である. [高橋 09] 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を変化させること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, 最適な分類器を選択する方法として, 予測クラスごとに出力されるスコア（分類スコア）を比較し最も大きな値の分類器を選択する方法を検討したが, 分類器が異なる場合には値を比較しにくいという欠点があった. そこで, 分類スコアから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 99, Zadrozny and Elkan 02] を推定し, 最も大きな値の分類器を選択する方法を検討した. クラス所属確率は確率であるために, 分類器が異なる場合でも値の比較が可能である. この他に多数決により選択する方法があるが, これはバギングに該当する.

[高橋 09] において, これら 3 つの方法を自由回答を含む調査データを用いて実験した結果, 提案手法であるクラス所属確率を用いる方法の有効性が示された. しかし, 分類器を構築にかかる手間に対して有効性の程度が高くないという問題点があった. そこで, 本稿では, 分類器の構築をリサンプリングにより行い, 調査データ以外のデータセットも用いた実験を行って, 本手法についてより詳細な調査を行う.

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

まず, [Torii and Liu 07] は, SVM（2 値分類）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て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. 分類器の選択に分類スコアが用いら

連絡先: 高橋和子, 〒 263-8588 千葉市稲毛区穴川 1-5-21
敬愛大学国際学部, takak@u-keiai.ac.jp

*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 09] 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.

れているが, 和であるために, 1 節で述べた問題は生じない.

次に, [神嶌他 08] はバギングを改造した BaggTaming と呼ばれる方法を提案した. これは, 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であるが, SVM における実験は行っていない.

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

分類スコアに基づいたクラス所属確率の推定方法には, パラメトリックな方法 [Platt 99] とノンパラメトリックな方法 [Zadrozny and Elkan 02] の 2 つがあるが, いずれも [Takahashi et al. 08] により, 多値分類への拡張が提案されている. すなわち, パラメトリックな方法はロジスティック回帰式の拡張, ノンパラメトリックな方法は「正解率表」を作成・利用する. このとき, ランク付けされたどの予測クラスに対しても, 複数個の分類スコアによる推定が有効であることが実験的に示された. 例えば, 第 1 位に予測されたクラスに対しては, ロジスティック回帰式を利用する方法では第 1 位から第 3 位に予測されたクラス, 「正解率表」を作成・利用する方法では第 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であった.

ロジスティック回帰式を利用する方法では, 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必要がある[*3].

「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必要がある. 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 [Takahashi et al. 08]. [*4]

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. 多値分類の場合には, 予測クラスはクラスの数だけランク付けされて出力されるが, 今回は第 1 位に予測されたクラスのみに注目した.

4.1 実験設定

4.1.1 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」 [SSM 06] 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 07] [*5]の 2 種類である.

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 01] 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 [SSM 07] 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 [高橋 08]. 本稿ではこの ISCO コードを正解として扱った. 素性は, 素性選択を変化させた [高橋 09] において最も分類精度が高かった素性を用いた [*6]. SVM を単独で適用した場合の分類精度は 73.9% であったｔめ, 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% であったため, 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.

今回, クラス所属確率の推定は, 安定性の点からロジスティック回帰式を利用する方法を用いた. パラメタ推定のために各訓練データをさらに訓練データと評価データに分割する際, 職業データは 10 分割, 20Newsgroups データセットは 5 分割の交差検定を行い, この評価データにおける正解／不正解の状況をそれぞれ 1 ／ 0 として用いた.

∗3　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.

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

.

∗4　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各クラス所属確率の和が 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　http://people.csail.mit.edu/jrennie/20Newsgroups/

∗6　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: 素性選択の変化による分類器構築における選択方法別分類精度（職業データ）　baseline：0.7392

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

4.1.2　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 99]. 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構築した. 評価尺度は分類精度（全クラスのマクロ平均）を用いた.

4.2　実験結果と考察

4.2.1　素性選択の変化による分類器構築

今回の実験結果を示す前に, 表 1 に前回, 人手により素性選択を変化させて分類器を構築し, 職業データにより実験を行った結果 [高橋 09] の一部を示す. 表の上段は 5 種類, 下段は 8 種類の分類器を構築した場合である. 表中, 提案手法（表）とはクラス所属確率の推定に正解率表を作成・利用した方法で, 正解率表の区間幅は [Takahashi et al. 08] にしたがい 0.1 とした. 太字は, 分類器数が等しい場合に分類精度が最も高くかつその値が単独の分類器を上回っていることを示す（以下同様である）.

表 1 より, 提案手法は分類器の数が少なくてもベースラインを上回り, 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有効であった. ただし, 0.7% しか高くなく, 訓練データ作成の手間を考慮すると, 効果的であるとはいえない.

なお, 次節以下の実験では, ここで構築した 8 種類の分類器の中で最も分類精度が高かった分類器で用いられた素性（注 6 参照）を用いることにした.

4.2.2　リサンプリングによる分類器構築

職業データおよび 20Newsgroups データセットによる結果をそれぞれ表 2, 3 に示す. 数値右の*印は単独の分類器を有意（有意水準 1%）に上回っていることを示す. また, 表中右列の目標値は, 正解であった分類器すべてをうまく選択できた場合の分類精度で, 本稿ではこれを目標とした（以下同様である）.

表 2 より, 職業データにおいては, 多数決法の一部を除き, すべての場合で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, 最大で 1.4% ベースラインより有意に高かった. ただし, 目標値にはるかに及ばず, つねに約 10%〜12% 程度低かった. 分類スコア法は 分類器の数が増えても値の変化が小さいために, 分類器が増えるにつれて順位が下がった. なお, 表 2 と表 1 において分類器の数がほぼ同数の場合を比較すると, 表 2 の方が値がよいことから, 素性選択を変化させるよりリサンプリングにより分類器を構築する方が有効であるといえる.

表 3 より, 20Newsgroups データセットにおいても, 多数決法や分類スコア法は一部でベースラインを下回る場合があったが, 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, 有意な差はなかった. また, 提案手法は多数決法と同様に, 分類器の数が増えるにつれて分類精度が上昇したが, 上昇の程度は多数決法の方が大きいため, 分類器が少ない場合は提案手法の方がよいが, 多くなると数決法に逆点された. ここでも, 提案手法は目標値よりつねに 12% 程度低かった. なお, 分類スコア法の有効性はこの場合も低かった.

表 2: 分類器の選択方法別分類精度（職業データ）（細分類）　baseline：0.7392

分類器数	提案手法	分類スコア法	多数決法	目標値
3	0.7395	**0.7423**	0.7335	0.8010
9	**0.7528***	0.7456	0.7451	0.8570
1 5	**0.7532***	0.7463	0.7443	0.8692
2 1	**0.7528***	0.7448	0.7452	0.8776

表 3: 分類器の選択方法別分類精度（20Newsgroups）（18,828 サンプル）　baseline：0.8730

分類器数	提案手法	分類スコア法	多数決法	目標値
9	**0.8753**	0.8733	0.8719	0.9837
1 5	**0.8760**	0.8730	0.8750	0.9927
2 1	0.8775	0.8728	**0.8805**	0.9980
2 5	0.8778	0.8736	**0.8816**	1.000

以上より, 提案手法は分類器をどのように構築しても, 分類器の数が少ない時点で有効性を示すこと, ただし, データセットやタスクの違いによりその傾向が異なることがわかった. 2 つのデータセットを比較すると, 分類精度だけでなく, クラスの分布状況も異なっていた. すなわち, 職業データは, 出現頻度が第 1 位のクラスが全体の 9.1% を占め, 以下 6.1%, 3.9% と続くが, 多くのクラスが 1% 未満で偏りがあったのに対し, 20Newsgroups データセットは, ほとんどのクラスの分布が約 5% で偏りがなかった. これより, 提案手法は, （1）クラスの分布状況に偏りがあるデータセットで（2）分類が困難なタスクにおいて有効性が高いと判断できる. 以下では, （1）（2）の要因別に実験を行い, 提案手法の性質をさらに調査する.

4.2.3　分類精度が高くクラス分布に偏りがある場合

分類精度が高くクラス分布に偏りがある場合として, 職業データを ISCO コード大分類（クラス数 11 個）に分類するタスクに対する実験を行った. このタスクは分類精度が 87.4（ベースライン）と高いが, クラスの分布は出現頻度が 23.9%, 15.3%, 13.9% と続き, 3 個のクラスは 2% 未満であり, 偏りが大きい. 表 4 に結果を示す. 表 4 より, どの手法も有効ではなく, ベースラインを上回ったのは分類器を 15 個以上構築した場合の提案手法のみであった. これより, 分類精度が高い場合は, クラス分布に偏りがあっても, 提案手法の有効性は高いとはいえない.

4.2.4　分類精度が低くクラス分布に偏りがない場合

分類精度が低くクラス分布に偏りがない場合として, 20Newsgroups データセットの一部を無作為に抽出し, 訓練データのサイズを小さくした実験を行った. この実験設定でも, クラスの分布は 3.3% から 6.1% で偏りがなかった. 単独の分類器の分類精度は 74.3 で, この値をベースラインとした. 表 5 に結果を示す. 表 5 より, すべての場合でベースラインを有意に上回ったが, 特に提案手法は最大で 7.5% 高かった. この値は, 分類精度が低く偏りがあった場合（表 2 参照）より大きく, また多数決法との差も大きかった.

以上より, 提案手法はクラスの分布に偏りがなく分類が困難

表 4: 分類器の選択方法別分類精度（職業データ）（大分類）
baseline：0.8740

分類器数	提案手法	分類スコア法	多数決法	目標値
3	0.8722	0.8698	0.8707	0.9062
9	0.8723	0.8695	0.8734	0.9343
1 5	**0.8805**	0.8770	0.8736	0.9482
2 1	**0.8792**	0.8749	0.8731	0.9537

表 5: 分類器の選択方法別分類精度（20Newsgroups）（3,500 サンプル）
baseline：0.7425

分類器数	提案手法	分類スコア法	多数決法	目標値
3	**0.8120***	0.8086*	0.7640*	0.8877
9	**0.8183***	0.8131*	0.7814*	0.9311
1 5	**0.8151***	0.8149*	0.7809*	0.9403
2 1	0.8117*	**0.8123***	0.7820*	0.9480
2 5	**0.8111***	0.8106*	0.7837*	0.9514

なタスクにおいて特に有効であるといえる.

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できにくい 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た. 性質の異なる 2 つのデータセットによる実験を行った結果, 提案手法は構築した分類器の数が少ない場合にも有効で, クラスの分布に偏りがなく分類が困難なタスクで特に有効であった. したがって, 提案手法は, 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. 今後の課題は, 最適な分類器の選択方法を検討するために, 提案手法のさらなる改善を行うことである.

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.

参考文献

[SSM 07] 2005 年社会階層と社会移動調査研究会. 2005 年 SSM 日本調査　コード・ブック (2007).

[SSM 06] 2005 年社会階層と社会移動調査研究会. 2005 年 SSM 調査　日本・韓国・台湾調査票 (2006).

[Asuncion and Newman 07] A. Asuncion and D. J. Newman. 2007. UCI Machine Learning Repository (2007).

[Bureau of Statistics 01] Bureau of Statistics; International Labour Office. Coding Occupation and Industry. Bureau of Statistics; International Labour Office (2001).

[Breiman 96] L. Brieman. Bagging predictors. In *Machine Learning* 24(2), pp.123–140 (1996).

[Dong and Han 04] Y-S. Dong and K-S. Han. 2004. A comparison of several ensemble methods for text categorization. In *Proceedings of IEEE 2004 International Conference on Services Computing (SCC 2004)*, pp.419–422 (2004).

[Joachims 98] T. Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proceedings of the European Conference on Machine Learning*, pp.137–142 (1998).

[神嶌他 08] 神嶌敏弘, 濱崎雅弘, 赤穂昭太郎. 飼い慣らし－飼育・野生混在データからの学習. 第 22 回人工知能学会発表論文集 (2008).

[kressel 99] U. Kressel. Pairwise classification and support vector machines. In *Advances in Kernel Methods Support Vector Learning*, pp.255–268. MIT Press (1999).

[Li et al. 08] X. Li, L. Wang, and E. Sung. AdaBoost with SVM-based component classifiers. In *Engineering Applications of Artificial Intelligence* 21(5) pp.785–795 (2008).

[元田他 06] 元田浩, 津本周作, 山口高平, 沼尾正行. データマイニングの基礎. オーム社 (2006).

[Platt 99] 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).

[Sebastiani 02] F. Sebastiani. Machine Learning Automated Text Categorization. In *ACM Computing Surveys* 34(1), pp.1–47 (2002).

[Takahashi et al. 08] K. Takahashi, H. Takamura, and M. Okumura. Direct estimation of class membership probabilities for multiclass classification using multiple scores. In *Knowl Inf Syst*, pp. 185–210. Springer London (2008).

[高橋 08] 高橋和子. 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 (2008).

[高橋 09] 高橋和子. 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 回全国人工知能学会大会発表論文集 (2009).

[Torii and Liu 07] M. Torii and H. Liu. Classifier ensemble for biomedical document retrieval. In *Proceedings of the Second International Symposium on Languages in Biology and Medicine (LBM 2007)* (2007).

[Zadrozny and Elkan 02] B. Zadrozny and C. Elkan. 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 (2002).