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

Ensemble Learning with Support Vector Machines

高橋和子
Kazuko Takahashi

敬愛大学
Keiai university

We propose an ensemble learning with support vector machines for which we create different classifiers by changing features without re-sampling. Although these classifiers predict some different classes, we decide the final class predicted by the best classifier selected by evaluation of class membership probabilities for each sample. Experiments showed that the proposed method was better than the best single classifier in precision.

1. はじめに

本稿では, サポートベクターマシン（SVM）による文書分類の精度を高めるために, 素性選択を変化させた複数の分類器を構築し, 各事例ごとに最適な分類器を選択するアンサンブル学習を提案する. このとき, 分類器の選択方法としては, クラス所属確率の利用を提案する.

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て, 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し, 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方法である [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方法である [元田 06].

しかし, これらの方法を分類精度の高さが評価されている SVM [Joachims 98] に適用する場合, 以下の点が問題となる. まず, バギングにおいては, 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バイアス－バリアンスの理論 [Breiman 96] によると, SVM のような高バイアスのモデルはもともとバリアンスの占める要素が少ないために, 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 [Torii 07, 神嶌 08]. また, ブースティングにおいては, バリアンスの問題に加え, SVM 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という問題もある [Li 08].

複数の分類器があるときに, 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）の平均が最も高い分類器がつねに他の分類器を上回って正解するわけではなく, この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も観察される. したがって, もし各事例ごとに正解の可能性が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の数が増え, 分類精度の向上が期待できよう. 本稿ではこのような考えに基づき, SVM において多様な結果が得られるような分類器を複数個構築し, 事例ごとに正解の可能性が高いと考えられる分類器を選択する方法を提案する. ここで, 多様な分類器の構築のためには, 事例の選択ではなく素性の選択を変化させる. この理由は, SVM においてはサポートベクターが分類に大きく関与するために, リサンプリングにより事例を変化させる方法より有効であると考えたためである. 提案手法においては, 複数の分類器の中から正解する可能性が高い分類器をうまく選択できることが重要である. 今回は, 多数決による方法, 分類器の出力するスコア（分類スコア）を用いる方法, 分類スコアにより推定したクラス所属確率を用いる方法の 3 つについて検討し, クラス所属確率を用いる方法の有効性を実験的に示す.

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.

連絡先: 高橋和子，takak@u-keiai.ac.jp

2. 関連研究

関連研究として, [神嶌 08] および [Torii 07] について述べる.

[神嶌 08] で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. [Torii 07] では 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する. 分類スコアの和が大きいクラスに決定する. 素性の選択を変化させて分類器を構築する点では本稿と同様であるが, クラスの決定方法が異なる.

3. 提案手法

提案手法の概略は, 次の通りである.

STEP1　素性の選択を変化させて複数の分類器を構築する
STEP2　各事例に対して分類器ごとにクラスを予測する
STEP3　各事例ごとに適切な分類器を選択し, その分類器が予測したクラスを最終決定とする

3.1 分類器の選択方法

適切な分類器の選択方法として, 今回は次の 3 つを検討する.

- 多数決による方法（以下, 「多数決法」と略す）
- 分類スコアを用いる方法（以下,「分類スコア法」と略す）
- ［提案手法］分類スコアにより推定したクラス所属確率を用いる方法（以下, 「クラス所属確率法」と略す）

「多数決法」は, 多数決により決定されたクラスを予測した分類器を選択する. 多数決の考え方はバギングにおいて一般的であるため, 分類器の選択方法としてではないが, アンサンブル学習における従来手法として扱う. 「分類スコア法」は, 各事例に対して予測されたクラスに付随して出力される分類スコア（SVM においては分離平面からの距離）の中で最も大きな値をもつ分類器を選択する. 本来, 分類器が異なる分類スコア同士を比較することはできないが, [Torii 07] においても用いられているため [*1], 本稿においても検討する. 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, 第 1 位に予測されたクラスのもののみを対象とする. 「クラス所属確率法」は, 本稿における提案手法で, 事例が分類器が予測したクラスに属する確率（クラス所属確率) を推定し [Platt 99, Zadrozny 02, Takahashi 08], 最も大きなクラス所属確率をもつ分類器を選択する. クラス所属確率法においても, 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としては次の 2 つがあり, （）内に示した分類スコアの利用が有効である [Takahashi 08].

- ロジスティック回帰式を利用（第 1 位から第 3 位に予測されたクラスの分類スコア）
- 「正解率表」を作成・利用（第 1 位および第 2 位に予測されたクラスの分類スコア）

パラメトリックな方法であるロジスティック回帰式の利用は, 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 [*2].

ノンパラメトリックな方法である「正解率表」の作成・利用においても,「正解率表」の作成のために, 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 [*3].

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 [*4].

4. 実験と考察

提案手法を 2005 年 SSM 調査（社会階層と社会移動に関する全国調査）により収集された職業データを 390 個の国際標準職業分類（ISCO）コード [Bureau of Statistics 01] に分類するタスク（「ISCO 職業コーディング」）に適用し, 有効性を調査した.

4.1 実験設定

4.1.1 データセット

用いたデータセットは 16,089 サンプルで, 10 分割交差検定により訓練データと評価データに分割した. ロジスティック回帰式におけるパラメタ推定および正解率表作成のためには, 各訓練データごとに 10 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. また, 正解率表の区間幅は, [Takahashi 08] にしたがって 0.1 とした. 職業データには, すでに調査終了後に行われた職業コーディングにより, SSM コードと ISCO コードの 2 種類の職業コード（各 1 個）が付与されている. 本稿では, この ISCO コードを正解とした.

4.1.2 素性の選択

今回は職業データ以外のデータが利用可能であったため, 次のような方法により素性の選択を変化させた. まず, 職業データのうち,「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 個の選択回答）の 3 種類を基本素性とした. 次に, 基本素性に職業データ以外の素性, すなわち「学歴」（6 個の選択回答）,「性別」（2 個の選択回答）, 職業コーディングによりすでに付与された「SSM コード」（約 200 個）の 3 種類について, 組み合わせを変えて追加した. 構築した分類器は次の 8 種類である.

［分類器A］基本素性のみ

［分類器B］基本素性, 学歴

［分類器C］基本素性, 正解 SSM コード

［分類器D］基本素性, 学歴, 正解 SSM コード

［分類器E］基本素性, 性別

［分類器F］基本素性, 学歴, 性別

［分類器G］基本素性, 性別, 正解 SSM コード

［分類器H］基本素性, 学歴, 性別, 正解 SSM コード

---

*1 2 節で述べたように, 分類器の選択方法としてではない.

*2 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.

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

.

*3 「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した [Takahashi 08].

*4 [Takahashi 08] 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.

表 1: 各分類器における分類精度（平均）

分類器	A	B	C	D
分類精度	0.689	0.693	0.734	0.737
分類器	E	F	G	H
分類精度	0.692	0.696	0.735	**0.739**

4.1.3 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. また評価尺度としては, 分類精度（全クラスのマクロ平均）に加えて, 出現頻度の高い特定のクラスに注目した場合の F 値と AUC（area under ROC curve）も用いた.

4.2 予備実験

各分類器の分類精度を表 1 に示す. 表中, 太字は最も高い値である. 表 1 より, 単独の分類器の中で分類精度が最も高いのは分類器 H（73.9%）であったため, 以下ではこの分類器をベースラインとした. 全体では, 分類精度の高いグループ（分類器C, D, G, H）と低いグループ（分類器器A, B, E, F）に分かれたが, 高いグループの素性には SSM コードが含まれ, 低いグループには含まれていなかった.

提案手法の有効可能性を確認するために, 分類器の一部（AからD）に対して, 分類精度および正解の出現状況を調査した（表 2 参照）. 表中, 太字は分類器の中で最も高い値である. 分類器の正解状況を正解した分類器の数ごとにみると, 1 個の分類器だけが正解の場合 1.9%, 以下同様に, 2 個の場合 12.7%, 3 個の場合 2.5%, すべての場合 62.6% であった. 分類器が 1 個だけ正解した場合に注目すると, 事例数が少ないとはいえ, 分類精度の高さと関係なくどの分類器においてもほぼ等しい値であった. これより, 事例ごとにみれば, 分類精度の高い分類器がつねに正解するわけではなく, より分類精度の低い分類器だけが正解する場合もあることが確認できた. したがって, 本稿で提案するように, 各事例ごとに, 正解した分類器がある場合にはこの分類器をうまく発見し, その分類器が予測したクラスを最終決定とすることが効果的である. 実際, 分類器AからDに対して最適な分類器をすべて選択できた場合の分類精度は 79.7%（表 2 最右列に示した値の合計）となり, ベースラインを 6.0% 上回った. 同様に, 8 種類すべての分類器の場合は 80.5% で, ベースラインを 6.6% 上回った. 本稿ではこの値を分類精度の目標値とした.

今回構築した分類器は, 分類精度に注目すると 表 1 に示したように 2 種類にまとまっていた. この傾向をより詳細に調査するため, 各事例ごとの分類器における正誤状況（正解／不正解をそれぞれ 1 ／ 0 とする）に基づいて分類器間の相関係数を算出すると（表 3 参照）, 表 3 からも, 分類器は 2 つのグループ（A, B, E, FとC, D, G, H）にまとまっていることがわかった. したがって, 今回の実験結果を考察する際に, 分類器の多様性に問題があったことを考慮する必要がある.

4.3 実験結果と考察

分類器の選択方法 *5別の分類精度（全クラスの平均）を表 4 に示す. 表 4 には, 出現頻度の高いクラス（「4110」,「5220」,「3415」） *6に注目した場合の F 値と AUC も示す. 表中, 太字はすべての方法の中で最も高い値である. 表 4 より, 分類精度において, クラス所属確率法は正解率表を利用する場合（0.7%）, ロジスティック回帰を利用する場合（0.2%）ともにベースラインを上回った. 一方で, 多数決法と分類スコア法はそれぞれ 0.1%, 0.8% ベースラインを下回った. なお, F 値による評価では,「3415」におけるクラス所属確率法（正解率表）以外はどの手法もベースラインを下回り, AUC によっても, 非常に限定された状況での結果であるが, ベースラインを上回る手法はなかった. 以上より, 今後, より一般的なタスクに対する実験により確認する必要があるが, 提案手法は, クラス所属確率法（特に正解率表の利用）の場合には有効性を示したが *7, 分類器の選択方法によっては有効ではなかった. 特に, 分類スコア法はどの評価法によってもつねに結果が悪く, 多数決法はベースラインと類似の傾向であった. 分類精度がもっともよかったクラス所属確率法（正解率表を利用）においても, ベースラインをやや上回る程度で目標値との差が大きかった（5.9%）ため, 改善策を検討する必要がある. クラス所属確率法の場合はクラス所属確率を高精度に推定できることが重要であり, 精度の向上が提案手法の有効性につながると考えられる.

表 3: 正誤状況に注目した場合の分類器間の相関係数

	B	C	D	E	F	G	H
A	0.93	0.64	0.64	0.90	0.89	0.64	0.64
B		0.64	0.65	0.89	0.91	0.64	0.65
C			0.65	0.67	0.67	0.97	0.93
D				0.67	0.67	0.94	0.97
E					0.95	0.67	0.68
F						0.67	0.68
G							0.95

表 5: 分類器の多様性の程度別分類精度（平均）の向上

多様性の程度	小	中	大
クラス所属確率法（正解率表）	0.696	0.739	0.746
単独で最もよい分類器	0.696	0.734	0.739
両者の差	0.000	0.005	0.007

次に, 分類器の多様性の違いによる分類精度の変化を調査した. 表 5 に, 分類器の多様性の程度が大（すべての分類器）, 中（分類器A, B, C, E）, 小（分類器A, B, E, F）の 3 つのグループについて比較を行った結果を示す. ここで, 分類器A, B, C, Eのグループは他の素性と性質が異なる素性（SSM コード）をもつ分類器Cを含むため, これを含まないグループより多様性の程度が大きいと考えた. 表 5 より, 提案手法は分類器の多様性の程度が大きいほど有効性が高く, 程度が小さい場合には効果がなかった. 以上より, 提案手法の有効性を高めるには多様な分類器を構築することが重要で, これが可能なデータセットにおいて提案手法の有効性が期待できる.

*5 今回偶数個の分類器を構築したために同数の分類器が異なるクラスを予測する場合があるが, 多数決法は, 分類器Hが含まれるグループを選択する.

*6 「4110」,「5220」,「3415」はそれぞれ「Office clerks」,「Shop, stall and market salespersons and demonstrators」,「Technical and commercial sales representatives」で, 正例（負例）は 2026（14063）, 1170（14919）, 664（15425）サンプルであった.

*7 正解率表を利用する場合とロジスティック回帰を利用する場合の差は, クラス所属確率の推定精度の違いによるものであると考えられるが, 今回, 正解率表を利用する場合のクロスエントロピーが計算できなかったため, 明確にはいえない.

表 2: 単独の分類器（A～D）における分類精度（平均）と正解の出現状況

分類器	A	B	C	D	分類器を適切に選択できた場合
1 個の分類器だけ正解	0.004	**0.005**	0.004	**0.005**	0.019
2 個の分類器が正解	0.045	0.044	**0.084**	0.083	0.127
3 個の分類器が正解	0.014	0.018	0.020	**0.023**	0.025
すべての分類器が正解	**0.626**	**0.626**	**0.626**	**0.626**	0.626

表 4: 分類器の選択方法別分類精度（平均）, F 値（上位 3 クラスに注目した場合）と AUC（最多クラスに注目した場合）

分類器の選択方法	ベースライン（分類器H）	多数決法	分類スコア法	クラス所属確率法（ロジスティック回帰）	クラス所属確率法（正解率表）
分類精度	0.739	0.738	0.731	0.741	**0.746**
F 値（クラス「4110」）	**0.8569**	0.8557	0.8392	0.8524	0.8515
F 値（クラス「5220」）	**0.9271**	0.9263	0.8883	0.9028	0.9067
F 値（クラス「3415」）	0.8227	0.8218	0.8123	0.8216	**0.8248**
AUC（クラス「4110」）	**0.6773**	0.6756	0.6719	0.6754	0.6745

5. おわりに

本稿では, SVM における分類精度を高めるために, 素性を変化させて多様な分類器を構築し, 各事例ごとに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. 提案手法は分類器のタイプに関係なく適用でき, 性質の異なる多様な素性が利用できるような分類タスクにおいて有効であると考えられる. 今後の課題は, より一般的なデータセットに対する実験を行って有効性を再確認すること, および分類器の選択方法についてさらに検討を深めることである.

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

参考文献

[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).

[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 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, pp.1–4 (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 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 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).

[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).

[Takahashi 08] K. Takahashi, H. Takamura, and M. Okumura. Direct estimation of class membership probabilities for multiclass classification using multiple scores. In *Knowl Inf Syst* (doi:10.1007/s10115-008-0165-z). Springer London (2008).

[Tao 06] D. Tao, X. Tang, X. Li, and X. Wu. Asymmetric Bagging and Random Subspace for Support Vector Machines-Based Relevance Feedback in Image Retrieval. In *The IEEE Transactions on Pattern analysis and machine intelligence (TPAMI)* 28(7), pp.1088–1099 (2006).

[Torii 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 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).