

## 多クラス SVM におけるクラス所属確率を用いたアンサンブル学習の提案

高橋 和子<sup>†1</sup>

本稿では、多クラスの SVM (サポートベクターマシン) における分類精度のさらなる向上を目的に、分類器が予測クラスとともに出力する分類スコアからクラス所属確率を推定して用いるアンサンブル学習を提案する。提案手法では、まず複数の分類器を構築し、次に未知の事例に対して各分類器が予測したクラスに付随して出力される分類スコアを複数個用いて、予測クラスに対するクラス所属確率を推定する。最後に、事例ごとに最も高いクラス所属確率をもつ分類器を選択し、その分類器が予測したクラスを最終決定とする。提案手法を性質の異なる 2 つのデータセットを用いた 9 つのタスクに適用した結果、分類精度が低い場合ほど有効性が高かった。また、分類器の数が少ない場合にも有効であった。

### Ensemble Learning by Using Class Membership Probabilities with Multi-class SVMs

KAZUKO TAKAHASHI <sup>†1</sup>

We propose an ensemble learning by using class membership probabilities with multi-class SVMs (Support Vector Machines). In the proposed method, we first make various classifiers by resampling or feature selection. Then we select the best classifier by using class membership probabilities estimated from classifier scores for each sample, which are outputted with predicted classes. Finally we assign the class predicted by the selected classifier to the sample. We applied the proposed method to two different dataset, which have nine tasks. The results of experiments showed that the proposed method was more effective than bagging if classification accuracies were lower than 80%, and that it was also effective for small number of classifiers.

### 1. はじめに

本稿では、多クラスの SVM (サポートベクターマシン) 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、各分類器が第 1 位の候補として予測したクラスに対してクラス所属確率を推定し、事例ごとにその中から最も高い値をもつ分類器を選択して、この分類器が予測したクラスを最終的なクラスとして決定する方法を提案する。

機械学習においては、複数の分類器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、代表的な方法として、バギングやブースティングがある<sup>17)</sup>。しかし、これらの方法は、文書分類などで多用される SVM<sup>10)</sup> においては有効でない場合が多い<sup>5),8),16),26)</sup>。この理由として、バギングについては、バイアス - バリアンス理論<sup>\*1</sup>により、SVM のような高バイアスの分類器ではバリアンスの占める要素がもともと小さいために、バギングによる効果が低バイアスのモデルほどには期待できないと説明される。また、ブースティングにおいても、SVM の場合は、事例に対する重み付けを直接反映させることが困難であるという問題がある<sup>15)</sup>。したがって、SVM においては、バギングやブースティングとは別の視点によるアンサンブル学習について検討する必要があると考えられる。ところで、アンサンブル学習全般における問題として、結果の解釈が困難になり、理解しやすさの点で問題を呈しているとの指摘があるために<sup>17)</sup>、検討すべき方法は、単純かつ理解可能性が高いものであることに留意する必要がある。

この検討を行うために、本稿では、次のような場合、すなわち、ある事例に対して複数の分類器が存在する場合の正解状況を観察すると、分類精度 (全クラスのマクロ平均) が高い分類器が不正解の事例に対して、より低い値の分類器が正解する場合があることに注目する。このような場合に、事例が分類されるクラスを正しく予測するには、バギングにおける最終決定法として多用される多数決 (やその改良形) より、正解した分類器をうまく見つけて、その分類器が予測したクラスを最終決定とする方が効果的であると考えられる。また、このような事例に限らず、正解した分類器がある場合に、つねにその分類器をうまく見つけることができれば、全体の正解率を高めることができるのは明らかである。この方法は、複数の分類器の結果の統合として各分類器の値を組み合わせる方法ではないが、考え方が単純で理

<sup>†1</sup> 敬愛大学国際学部

Faculty of International Studies, Keiai University

\*1 文献 4) によれば、誤差はバイアス (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 つの部分に分解できるとする。

解しやすいという利点がある。しかし、問題は、どのようにして事例ごとに正解の可能性が高い分類器を見つければよいのかということである。

SVM においては、分類器が予測したクラスの正解の可能性は、予測クラスとともに出力される分類スコア（事例と分離平面との距離）の大きさにより評価することができる。したがって、分類器を選択する方法として、最も単純には分類スコアの利用が考えられるが、分類器が異なっている分類スコアを比較することはできないという問題がある。そこで、異なる分類器同士での比較を可能にするために、分類スコアを確率に変換する。例えば、クラス所属確率は事例がそのクラスに所属する確率であり、予測したクラスに対する分類器の確信度、言い換えれば分類器がどの程度信頼性できるかを示すものである。クラス所属確率は推定値であるが、分類スコアを用いたパラメトリックな方法<sup>18)</sup>またはノンパラメトリックな方法<sup>32)</sup>が提案されている。さらに、いずれの方法に対しても、多クラス分類問題に対する拡張方法が提案されている<sup>21),22)</sup>。

この拡張方法の利用により、文献 23), 24) では、多クラス SVM においてクラス所属確率が最も高い分類器を選択する方法を提案し、2つのデータセットを用いたタスクによる実験を行った。提案手法における「複数の分類器の中から最も信頼できる分類器を選択する」という考え方は、理解しやすいものである。実験の結果、提案手法は従来手法であるバギング（多数決）より有効な場合があることがわかった。本稿では、さらに多くの実験を行って、提案手法における有効性を詳細に示すと同時に、計算量やクラス所属確率の推定精度についても考察する。本稿によって提案手法の有効性が明確に示されれば、SVM における有効なアンサンブル学習としての貢献が期待できる。

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

SVM におけるアンサンブル学習に関する研究として次の3つについて述べる。まず、Kim ら<sup>11)-13)</sup> は、多くの研究と異なり、SVM においてバギングやブースティングが有効であったとする。リサンプリングにより分類器を構築し、最終クラスの決定方法として、多数決法、LSE-based weighting, double-layer hierarchical combining の3つを提案した。多クラス SVM においては、多項式カーネルや Gauss カーネルを用いて、IRIS (3 クラス) や Hand-Written Recognition (10 クラス)<sup>3)</sup> および Mobile telecommunication payment fraud detection (2 クラスと 4 クラス) のタスクによる実験により実験し、分類精度を評価尺

度として比較した結果、いずれの手法も単独の SVM より有効であったが、手法間ではタスクにより異なっていた。すなわち、IRIS および Hand-Written Recognition では double-layer hierarchical combining, Fraud Detection では多数決法 (LSE-based weighting は計算量が多く適用できなかった) が最もよい結果であった。

文献 28) は、バイアス - バリアンズ理論による誤差の分解の観点から、SVM におけるバイアスを小さくするために、カーネルのパラメタを調節した Lobag (Low Biased Bagged) SVM を提案した。Lobag, Bagged SVM, 単独の SVM を線形カーネル, 多項式カーネル, Gauss カーネルを用いて、UCI リポジトリにおける letter recognition など 7 種類のデータセット<sup>3)</sup> により実験し、平均推定誤差および win-tie-loss を評価尺度として比較した結果、Lobag, Bagged SVM, 単独の SVM の順によかった。

文献 30) は、SVM におけるアンサンブル学習として、バギング, AdaBoost, Arc-X4, 変形 AdaBoost の 4 種類を提案した。線形カーネル, 多項式カーネル, Gauss カーネルを用いて、UCI リポジトリにおける 20 種類のデータセット<sup>3)</sup> により実験し、分類精度を評価尺度として比較した結果、バギングが最もよかった。

## 3. 提案手法

提案手法の手順は次の通りである。

- STEP1 リサンプリングまたは素性選択を変化させて複数の分類器を構築する
- STEP2 各分類器は未知の事例に対してクラスを予測する
- STEP3 各分類器の予測クラス (第 1 位) に対するクラス所属確率を推定する
- STEP4 クラス所属確率が最も大きな分類器を選択し、この分類器の予測クラスを最終決定とする

STEP3 におけるクラス所属確率の推定は、パラメトリックな方法 (ロジスティック回帰式を利用して直接的に推定) およびノンパラメトリックな方法 (「正解率表」を作成し利用して間接的に推定) の 2 種類がある。文献 21), 22) による実験では、いずれの方法においても、第 1 位だけでなく第 2 位または第 3 位に予測されたクラスの分類スコアまで利用する方法が有効であったため、本稿においても複数の分類スコアを用いる。各方法について、3.1 節および 3.2 節で述べる。

### 3.1 クラス所属確率の推定方法（ロジスティック回帰式の利用）

ロジスティック回帰式によりクラス所属確率を直接的に推定する方法において、最もよい結果が得られたのは、第1位から第3位までに予測されたクラスの分類スコア ( $f_1, f_2, f_3$ ) を利用した場合であった。したがって、未知の事例において第1位に予測されたクラスに対するクラス所属確率は、これら3個の分類スコアを次のロジスティック回帰式

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

に代入して計算すればよい。

このとき、(1)式における4個のパラメタ ( $A_1, A_2, A_3, B$ ) は最尤法により、あらかじめ推定しておく必要がある。このためには、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た結果を用いる<sup>\*1</sup>。

### 3.2 クラス所属確率の推定方法（正解率表を作成・利用）

正解率表を作成・利用する方法において最も結果がよかったのは、第1位および第2位に予測されたクラスの分類スコア ( $f_1, f_2$ ) を利用した場合であった。正解率表を作成する方法について、これら2個の分類スコアを利用する場合を例に説明する。この場合の正解率表は、第1位に予測されたクラスの分類スコアを横軸、第2位に予測されたクラスの分類スコアを縦軸とし、両軸とも、適切に設定された等間隔の幅（例えば、0.1）で区切られることにより生じるセル<sup>\*2</sup>から構成される2次元の表である。セルの内容は正解率で、この値は、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、各セルに該当する評価データを振り分けて、セルごとに、正解事例数を全事例数で割った値である。

未知の事例において第1位に予測されたクラスに対するクラス所属確率は、第1位および第2位の予測クラスに対する分類スコアに基づいて正解率表の該当するセルを探し、そのセルの正解率を間接的に用いる。

ロジスティック回帰式を利用する方法と正解率表を作成・利用する方法を比較すると、正

\*1 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、 $Y^1, \dots, Y^n$  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<sup>25)</sup>。

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

\*2 文献 32) では第1位の分類スコアしか扱わないために軸を1つしか必要とせず、ピンとよばれている。

解率表を作成・利用する方法は、分類スコアを表す各軸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、結果の安定性の点において、ロジスティック回帰式を利用する方法の方が優れていた<sup>21), 22)</sup>。したがって、本稿におけるクラス所属確率の推定には、ロジスティック回帰式を利用する方法を用いる。

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 (1) 式で計算された値を正規化する必要があるが、本稿では、注目するクラスに関してそのクラスに所属するかどうかに関心があるため、正規化までは行わない<sup>\*3</sup>。

## 4. 実験と考察

提案手法の有効性を調査するために、多数決により最終的なクラスを決定する方法（以下、「多数決法」と略す）を従来手法であるバギングとみなして比較を行った。また、提案手法と基本的には同じ考え方であるが、分類器の選択に分類スコアをそのまま利用する方法（以下、「分類スコア法」と略す）との比較も行った。なお、単独の分類器の中で最も高い値を baseline とした。

### 4.1 実験設定

#### 4.1.1 データセットとタスク

実験には、調査データである自由回答および英文ニュース記事という性質の全く異なる2種類のデータセットを用いた。自由回答は、2005年SSM調査（社会階層と社会移動に関する全国調査）により収集されたデータのうち、職業に関するデータ（以下、職業データと略す）<sup>2)</sup>で、5語程度（平均）の非常に短い日本語と複数の選択回答から構成される。英文ニュース記事は、UCIリポジトリ<sup>3)</sup>における文書分類用データセットの20Newsgroupsである。

本稿では、2つのデータセットに対して、それぞれ複数のタスクを実験した。まず、職業データ（16,089サンプル）におけるタスクは、国際標準職業分類であるISCO（International Standard Classification of Occupations）コード<sup>6)</sup>を付与することである。例えば、「仕事の内容」が「JA協同組合（肥料、融資、農機具等のあっせん）」、「従業先事業の種類」が「各農家を巡回して注文（肥料）、農作物の生産指導（先輩と同行）」、「従業上の地位・役職」が「2（常時雇用の一般従事者 役職なし）」、「従業先事業の規模」が「5（30～99人）」

\*3 文献 21) においては、正規化した値としない値の違いは大きくなかった。

という回答には、「3213」(Farming and forestry advisers)なるISCOコードが付与される。ISCOコードは、すでに調査終了後に行われたISCO職業コーディングにより付与されているため、本稿においてはこのコードを正解として扱った。学習に用いた素性は、職業データを構成する「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・役職」(13種類の選択回答)であるが、分類精度を高めるために、「学歴」(6種類の選択回答)および国内標準職業分類であるSSM(Social Stratification and social Mobility)コード<sup>\*1)</sup>も追加した。ISCOコードの付与は、例で示した4桁の小分類(390個)が一般的であるが、ISCOコードでは、小分類の上位に3桁の中分類(116個)、2桁の垂大分類(28個)、1桁の大分類(10個)が階層的に存在するため<sup>\*2)</sup>、本稿ではこれら4つのコードを付与するタスクについての実験を行った。各タスクにおけるbaselineは順に、74.9%、79.2%、84.2%、87.4%であった<sup>\*3)</sup>、

次に、20Newsgroupsデータセット(18,828サンプル)におけるタスクは、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ことであり、素性はネットニュース記事に出現する単語unigramを用いた。分類精度を変化させるために、本稿ではノイズを10%から20%まで適宜混入させ、4種類のタスクについての実験を行った。各タスクにおけるbaselineは順に、69.2%、77.9%、79.2%、87.3%であった<sup>\*4)</sup>。

以上に述べたように、2つのデータセットは、言語や内容、用いられる素性以外にも、クラス分布の偏りの程度が、職業データでは大きく、20Newsgroupsデータセットでは小さいという違いがあった。

実験のための訓練データと評価データの分割は、職業データでは10分割交差検定、20Newsgroupsデータセットでは5分割交差検定により行った。すなわち、職業データの訓練データは14,480サンプル、評価データは1,609サンプルで、20Newsgroupsデータセットの訓練データは15,062サンプル、評価データは3,766サンプルであった。

以上に述べた計8つのタスクによる実験に加え、本稿では小さなサイズの訓練データにおける有効性を調査するために、20Newsgroupsデータセット(ノイズなしの場合)の中から

\*1 ISCO 職業コーディングに先立ち、SSM 職業コーディングも実施された。

\*2 例えば、例で示した小分類3213の上位に、中分類321(Life science technicians and related associate professional), 垂大分類32(Life science and health associate professionals), 大分類3(TECHNICIANS AND ASSOCIATE PROFESSIONALS)が位置する。

\*3 クラス分布の平均(標準偏差)は順に、0.3%(0.9%), 0.8%(1.5%), 3.1%(4.0%), 9.1%(7.3%)であった。

\*4 クラス分布の平均(標準偏差)は、いずれも5.0%(0.5%)であった。

表1 分類器の個数別正解した分類器をすべて選択できた場合の分類精度(目標値)の変化(単位:%)

タスク	baseline	3個	5個	7個	9個	11個	15個	17個	19個	21個
職業データ	73.9	80.1	82.4	85.1	85.7	86.0	86.5	87.2	87.5	87.8
20Newsgroups	87.3	95.0	96.8	97.8	98.4	98.7	99.0	99.5	99.6	99.8

無作為に3,500サンプルを抽出し、実験を行った。この実験では、訓練データと評価データを5分割交差検定により分割した。すなわち、訓練データは2,800サンプル、評価データは700サンプルで、訓練データのサイズは前述した20Newsgroupsデータセットの1/5である。

#### 4.1.2 分類器と評価尺度

分類器は文献17)を参考にし<sup>\*5)</sup>、いずれのタスクにおいても、リサンプリングにより最大21個まで構築した。その際、SVMは2値の分類器であるために、one-versus-rest法を用いて多クラスの分類器へと拡張した<sup>14)</sup>。カーネル関数は、文献19)において有効性の高かった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)を用いた。

#### 4.2 実験結果と考察

##### 4.2.1 提案手法による分類精度向上の可能性

実験を行う前に、職業データ(小分類)タスクと20Newsgroupsデータセット(ノイズなし)タスクに対して、正解した分類器をすべて選択できた場合の分類精度向上の程度を調査した。分類器数別の結果を表1に示す。表1より明らかなように、分類器が増えるほど分類精度が向上したが、3個の場合にすでに、職業データ(小分類)タスクと20Newsgroupsデータセット(ノイズなし)タスクにおいてそれぞれ6.2%、7.7%向上しており、21個の場合にそれぞれ13.9%、12.5%まで向上した。これより、提案手法は、正解した分類器をうまく選択できれば有効性が高く、また少ない分類器でも効果のある手法であることがわかった。実際には、正解の分類器が存在しない事例があり得ることや、クラス所属確率は推定値であるために、推定精度の問題もあることから、表1に示した値を実現することは困難であると考えられるが、本稿ではこの値を目標とした。

\*5 文献17)においては、分類精度70%の2項分布において分類器を最大21個まで構築した場合を例として、多数決法により分類器が11個以上(半分以上)間違える確率が2.6%であることから、この場合の誤差が1/10以下に減少することを示した。

表 2 手法別分類精度 (職業データ) (16,089 サンプル) (単位: %)

タスク (baseline)	手法	3 個	9 個	15 個	21 個	向上度 (最大値)
大分類 (87.4)	提案手法	87.2	87.2	<b>87.9</b>	<b>88.1</b>	<b>0.7</b>
	多数決法	87.1	87.3	87.3	87.4	0.0
	分類スコア法	87.0	87.0	87.0	87.7	0.3
亜大分類 (84.2)	提案手法	<b>84.3</b>	84.4	84.3	84.4	0.2
	多数決法	84.2	<b>84.5</b>	<b>84.4</b>	<b>84.5</b>	<b>0.3</b>
	分類スコア法	84.3	84.3	84.2	84.2	0.1
中分類 (79.2)	提案手法	<b>79.7</b>	79.7	80.0	80.0	0.8
	多数決法	79.5	<b>80.0</b>	<b>80.1</b>	<b>80.1</b>	<b>0.9</b>
	分類スコア法	79.6	79.6	79.6	79.5	0.4
小分類 (73.9)	提案手法	74.0*	<b>75.3</b>	<b>75.3</b>	<b>75.3</b>	<b>1.4</b>
	多数決法	72.6	74.3	74.6	74.7	0.8
	分類スコア法	<b>74.2</b>	74.6	74.6	74.5	0.7

表 3 手法別分類精度 (20Newsgroups) (18,828 サンプル) (単位: %)

タスク (baseline)	手法	3 個	9 個	15 個	21 個	向上度 (最大値)
ノイズなし (87.3)	提案手法	<b>87.5</b>	<b>87.5</b>	87.6	87.8	0.5
	多数決法	87.2	87.5	<b>87.8</b>	<b>88.1</b>	<b>0.8</b>
	分類スコア法	87.0	87.0	87.7	87.5	0.4
ノイズ 10% (79.4)	提案手法	<b>80.9*</b>	<b>82.5*</b>	<b>82.7*</b>	<b>83.1*</b>	<b>3.7</b>
	多数決法	79.3	81.3	81.6	81.6	2.2
	分類スコア法	80.5	82.0	82.2	82.5	3.1
ノイズ 12.5% (77.9)	提案手法	<b>79.0*</b>	<b>81.3*</b>	<b>81.5*</b>	<b>83.1*</b>	<b>5.2</b>
	多数決法	77.5	79.9	80.2	80.2	2.3
	分類スコア法	76.6	80.6	80.9	81.2	3.3
ノイズ 20% (69.2)	提案手法	<b>73.4*</b>	<b>76.2*</b>	<b>76.9*</b>	<b>77.1*</b>	<b>7.9</b>
	多数決法	71.4	74.6	74.9	74.9	5.7
	分類スコア法	73.0	75.6	76.2	76.6	7.4

#### 4.2.2 職業データのタスクにおける実験結果

職業データを用いた 4 つのタスクにおける実験結果について、構築した分類器が 3 個、9 個、15 個、21 個の場合を表 2 に示す。表において、太字は、baseline を上回りかつ 3 つの手法の中で最も高い値であること、数値右上の\*印は、提案手法の値が多数決法より有意 (5% 水準) に高いことを示す (表 3、表 4 においても同様)。

表 2 より、提案手法は、単独の SVM における分類精度が低い場合 (73.9%)、または分類器が少ない場合 (3 個) に、多数決法を上回った。一方で、多数決法は分類精度が高かったり分類器が多い場合に提案手法を上回ったが、両手法間に有意な差は認められなかった。また、分類スコア法が 3 つの手法の中で最も高い値を示したのは、分類精度が最も低く分類器が最も少ないタスクでのみで、3 つの手法の中で最も悪い結果であった。なお、表 1 に目標値を示した小分類タスクにおいて、提案手法だけでなく他の手法も目標値との差が大きく、その差は分類器が増えるほど広がった \*1。

#### 4.2.3 20Newsgroups データセットのタスクにおける実験結果

20Newsgroups データセットを用いた 4 つのタスクにおける実験結果について、構築した分類器が 3 個、9 個、15 個、21 個の場合を表 3 に示す。

表 3 より、20Newsgroups データセットにおいても、提案手法は、単独の SVM における分類精度が低い場合 (80% 未満) または分類器数が少ない場合 (9 個以下) に、多数決法を上

\*1 提案手法の場合、6.1% から 12.5% に広がった。

回った。20Newsgroups データセットの場合は、多数決法は提案手法だけでなく分類スコア法をも下回る場合が多く、提案手法を上回ったのは、分類精度が高く (87.8%) かつ分類器が多い (15 個以上) 場合のみであった。表 2 および表 3 から、2 種類のデータセットにおける結果を比較すると、提案手法だけでなく、どの手法も職業データのタスクの場合より向上度が大きく、よい結果であった。ただし、20Newsgroups データセット (ノイズなし) の場合も職業データと同様に、いずれの手法も表 1 に示した目標値との差が大きく、また、分類器が増えるほど差が広がった \*2。

#### 4.2.4 サイズの小さな訓練データにおける実験結果

20Newsgroups データセット (ノイズなし) において、訓練データを 1/5 のサイズとした実験結果を表 4 に示す。構築した分類器は、第 4.2.2 節および第 4.2.3 節と同様に、3 個、9 個、15 個、21 個であり、表中、最下段における目標値も、第 4.2.1 節における目標値と同様である。このタスクにおける baseline は 74.8% で、18,828 サンプルの場合のノイズ 12.5% タスクとノイズ 20% タスクの間であった。また、このデータセットにおけるクラス分布の偏りの程度は、18,828 サンプルの場合と同様に小さかった。

表 4 は 1 つのタスクのみという限界はあるが、提案手法はつねに多数決法を上回り、特に分類器数が少ないほど手法間の差が大きかった。多数決法は分類スコア法をも下回り、3 つの手法の中でつねに最も低い値であった。分類スコア法は提案手法の次によく、特に分類器

\*2 提案手法の場合、7.5% から 12.0% に広がった。

表 4 手法別分類精度 (20Newsgroups) (3,500 サンプル)(単位: %)

タスク	手法	3 個	9 個	15 個	21 個	向上度 (最大)
ノイズなし (74.3)	提案手法	81.2*	81.8*	81.5*	81.2*	7.5
	多数決法	76.4	78.1	78.1	78.2	3.9
	分類スコア法	80.9	81.3	81.5	81.2	7.2
	目標値	88.8	93.1	94.0	94.8	20.5

が 21 個の場合には 3 つの手法の中で最も高い値であった。なお、このタスクにおいても他の場合と同様にいずれの手法も目標値との差が大きく、その差は分類器が増えるほど広がった\*1。

#### 4.2.5 考 察

ここでは、提案手法の性質について考察を行った後、計算時間を考慮した上で、再度、多数決法 (バギング) と比較を行う。さらに、構築すべき分類器の数や、提案手法において核となるクラス所属確率の推定精度と有効性の関係についても考察する。

まず、表 2 および表 3 より、提案手法は、データセットの性質に関係なく、分類精度が低い場合 (80% 未満) に有効性が高いことがわかった。また、表 3 および表 4 からは、提案手法は訓練データのサイズが小さい場合にも有効であることがわかった。したがって、分類が困難なタスクや、学習が必要な訓練データの量が少ない場合には、バギングより提案手法の方が適していると考えられる。

次に、提案手法は、データセットの性質やサイズに関係なく、分類器が少ない場合 (3~9 個) に有効性を示した (表 2 および表 3 参照)。このことは、提案手法においては、分類器の構築時間が少なくすむことを意味する。しかし、提案手法は、分類器の結果をそのまま利用できる多数決法や分類スコア法と異なり、クラス所属確率を計算するための時間が余分に必要になるため、以下で検討を行う。

提案手法で必須のロジスティック回帰式のパラメタ推定 (または正解率作成) のためには、訓練データを訓練データと評価データに分割し、再度、SVM を実行する必要がある。例えば、この作業を 5 分割交差検定により行う場合には、SVM を余分に 5 回実行する必要が生じる。ただし、この場合の訓練データ数は全訓練データ数の 4/5 であるために、計算時間が単純にデータ数の 3 乗<sup>7)</sup> に比例すると仮定すると、5 回の実行を行うためには、全訓練データを用いた場合の約 2 倍の計算時間となる。したがって、単純に比較すると、提案手法は多数決法

の約 3 倍の計算時間が必要なために、他の方法と同じ分類器数ではなく、3 倍の個数と比較する必要がある。例えば、提案手法の分類器 3 個は多数決法 9 個、同様に、提案手法 9 個は多数決法 27 個の結果と比較しなければならない。

以上より、表 2、表 3、表 4 において比較が可能なペアとして、提案手法 3 個と多数決法 9 個の Win-Tie-Loss を調査すると、1-1-7 で多数決法の方がよい結果となった。ただし、両手法の差 (平均) は、Win で 3.1、Loss で 0.5 で、Loss の場合の差は小さかった。同様に、提案手法 9 個 (多数決法 27 個\*2) の場合は、5-0-4 で提案手法の方がよく、両手法の差 (平均) も、Win で 1.4、Loss で 0.4 で、この場合も Loss における差は小さかった。Win の内容は、職業データ (小分類) タスクと 20Newsgroups データセットのノイズなしを除くすべてのタスクで、いずれも分類精度が 80% 未満であった。なお、提案手法を多数決法と計算時間が等しいと考えられる分類スコア法と比較すると、提案手法 3 個 (分類スコア法 9 個) の場合の Win-Tie-Loss は 2-1-6 であるが、提案手法 9 個 (分類スコア法 27 個\*3) の場合は 6-1-1 であった。これらの結果より、提案手法の計算時間を考慮しても、提案手法は分類器数が少ない場合に有効であるといえる。

ここで、今回の実験で構築した分類器は十分な数であったのかという点について、文献 4) では、カテゴリ型の分類 (クラス数は 3, 6, 7, 26) に対してリサンプリングの回数を 50 としたバギングの実験結果から、リサンプリングは、クラスの数が増えるにつれより多くの回数が必要であるとしており、クラスの数に応じて構築すべき分類器数が検討される必要があったと考えられる。しかし、職業データのタスクでは、クラスの数 390 (小分類)、116 (中分類)、28 (亜大分類)、10 (大分類) と大きく異なっていたにもかかわらず、すべて同数の分類器で実験を行ったため、クラスの数が多いタスクほど、良好な結果を得るための分類器数が不足した可能性が高い。実際、バギングやブースティングが有効であったとする Kim ら<sup>11)-13)</sup> の実験では、クラス数 3, 4, 10 に対する分類器はそれぞれ 5 個、11 個、11 個であった。また、本稿で参考にした文献 17) の分類器は 21 個であったが、多クラスではなく 2 クラスであった。しかし、一方で、文献 26) では、2 クラスに対して 10 個の分類器を構築したが、バギングは有効性を示さなかった。クラス数と実際に必要な分類器数の関係については、今後の課題である。

最後に、クラス所属確率の推定精度について考察する。提案手法においてクラス所属確

\*1 提案手法の場合、7.6% から 13.6% に広がった。

\*2 今回は分類器を 21 個までしか構築しなかったため、多数決法における分類器 27 個の結果は単回帰式により推定した。

\*3 多数決法における分類器 27 個の場合と同様に、単回帰式により推定した。

率は重要であるが、推定値であるために、その推定精度は有効性に大きく影響すると考えられる。これについて、今回得られた実験結果より調査した。データセットは異なるが、表 2 および表 3 において分類精度がほぼ等しい(約 79%) 職業データ(中分類)タスクと 20Newsgroups データセット(ノイズ 10%)タスクについて、クラス所属確率の推定精度をクロスエントロピー  $H^{*1}$ により比較した。計算の結果、それぞれ 0.5811 と 0.3562 であり、20Newsgroups データセット(ノイズ 10%)タスクの方が推定精度がよかった。提案手法の有効性も 20Newsgroups データセット(ノイズ 10%)タスクの方が高かったことから、クラス所属確率の推定精度がよいほど提案手法の有効性が高いといえる。ただし、今回はデータセットが異なっているために他の要因の影響もあり得るため、今後の課題として、同一のデータセットによる実験を行って確認する必要がある。

## 5. おわりに

本稿では、文書分類において分類精度の高さが評価されているサポートベクターマシンにおける単純かつ理解可能性の高いアンサンブル学習として、各分類器が予測したクラスに対して最も高いクラス所属確率をもつ分類器を選択し、この分類器が予測したクラスを最終的なクラスとして決定する方法を提案した。提案手法を性質の異なる 2 種類のデータセットを用いた 9 種類のタスクに適用した結果、単独の分類器より有効で、特に、分類精度の低い場合には多数決法によるバギングより有効であった。また、分類器の数が少ない場合にも有効であった。しかし、すべてのタスクにおいて、提案手法は正解したすべての分類器を選択できた場合の分類精度と 10% 程度の差があり、この値を縮めることが今後の大きな課題である。

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。本研究は平成 22 年度科研費(22530516)の助成を受けたものである。

## 参 考 文 献

- 1) 1995 年 SSM 調査研究会: SSM 産業分類・職業分類(95年版) 修正版, (1996).
- 2) 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).
- 3) Asuncion, A. and Newman, D. J.: UCI Machine Learning Repository, Univer-

\*1 クロスエントロピー  $H$  は次式で計算される。  $H(y, p) = \frac{1}{N} \sum_{i=1}^N \{-y_i \log(p_i) - (1 - y_i) \log(1 - p_i)\}$ ,  
ただし、 $N$  は評価データの事例数、 $p_i$  は  $i$  番目の事例におけるクロス所属確率の推定値、 $y_i$  は  $i$  番目の事例における正誤状況で、正解の場合に 1、不正解の場合に 0 とする。 $H$  の値が小さいほど推定精度は高い。

- sity of California, Irvine, School of Information and Computer Sciences (online), available from (<http://www.ics.uci.edu/~mllearn/MLRepository.html>) (accessed 2007-04-01).
- 4) Breiman, L.: Bagging predictors, *Machine Learning*, Vol.24, No.2, pp.124-130 (1996).
  - 5) Buciu, I., Kotropoulos, C. and Puas, I.: Combining Support Vector Machines for Accurate Face Detection, *IEEE 2001 International Conference on Image Processing (ICIP '01)*, Thessaloniki, Greece, pp.1054-1057 (2001).
  - 6) Bureau of Statistics; International Labour Office: Coding Occupation and Industry, (2001).
  - 7) Christianini, N. and Shawe-Taylor, J.: *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, (2000).
  - 8) Dong, Y-S. and Han, K-S.: A comparison of several ensemble methods for text categorization, *IEEE 2004 International Conference on Services Computing (SCC 2004)*, Shanghai, China, pp.419-422 (2004).
  - 9) Duda, R. O., Hart, P. E. and Stork, D. G.: *Pattern Classification*, A Wiley-Interscience Publication, 2nd edition(2000).
  - 10) Joachims, T.: Text Categorization with Support Vector Machines, *the European Conference on Machine Learning*, Chemnitz, Germany, pp.137-142 (1998).
  - 11) Kim, H-C., Pan, S., Je, H-M., Kim, D. and Bang, S-Y.: Support Vector Machine Ensemble with Bagging, *LNCS 2388*, Niagara Falls, Canada, pp.131-141 (2002).
  - 12) Kim, H-C., Pan, S., Je, H-M., Kim, D. and Bang, S-Y.: Pattern Classification Using Support Vector Machine Ensemble, *The 16th International Conference on Pattern Recognition (ICPR (2)'02)*, Quebec City, Canada, pp.160-163 (2002).
  - 13) Kim, H-C., Pan, S., Je, H-M., Kim, D. and Bang, S-Y.: Constructing support vector machine ensemble, *Pattern Recognition 36*, pp.2757-2767 (2003).
  - 14) Kressel, U.: Pairwise classification and support vector machines, *Advances in Kernel Methods Support Vector Learning*, MIT Press, pp.255-268(1999).
  - 15) 工藤 拓, 松本裕治: Support Vector Machine を用いた Chunk 同定, *自然言語処理*, Vol.19, No.5, pp.3-22 (2002).
  - 16) Lorena, A. C. and de Carvalho, A. C. P. L. F.: Human Splice Site Identification with Multiclass Support Vector Machines and Bagging, *LNCS 2714*, Istanbul, Turkey, pp.234-241 (2003).
  - 17) 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).
  - 18) Platt, J. C.: Probabilistic Outputs for Support vector Machines and Comparisons to Regularized Likelihood Methods, *Advances in Large Margin Classifiers*, MIT Press, pp.1-11(1999).
  - 19) 高橋和子, 高村大也, 奥村 学: *機械学習とルールベースの組み合わせによる自動職業*

- コーディング, 自然言語処理, Vol.12, No.2, pp.3-38 (2005).
- 20) 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 1 2 社会調査における測定と分析をめぐる諸問題, 2005 年 SSM 調査研究会, pp.47-68 (2008).
  - 21) 高橋和子, 高村大也, 奥村 学: 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 (2008).
  - 22) Takahashi, K., Takamura, H. and Okumura, M.: BDirect estimation of class membership probabilities for multiclass classification using multiple scores, *Knowledge and Information Systems*, Springer London, Vol.19, No.2, pp.185-210 (2008).
  - 23) 高橋和子: クラス所属確率を利用したアンサンブル学習, 人工知能学会第 24 回大会 長崎 (online), available from (<https://kaigi.org/jsai/webprogram/2010/pdf/260.pdf>) (accessed 2011-04-01).
  - 24) 高橋和子: クラス所属確率を用いた多クラス SVM におけるアンサンブル学習, 情報処理学会第 73 年全国大会, 東京 pp.2-25-2-26 (2011).
  - 25) 東大教養学部統計学教室 (編): 基礎統計学 III 自然科学の統計学, 東京大学出版会 (1992).
  - 26) Torii, M. and Liu, H.: Classifier ensemble for biomedical document retrieval, *The 2nd International Symposium on Languages in Biology and Medicine (LBM 2007)*, Matrix, Biopolis, Singapore, (2007).
  - 27) Valentini, G., Muselli, M. and Ruffino, F.: Bagged Ensembles of Support Vector Machines for Gene Expression Data Analysis, *the International Joint Conference on Neural Networks (IJCNN2003)*, Oregon, USA, pp.20-24 (2003).
  - 28) Valentini, G. and Dietterich, T. G.: Low Bias Bagged Support Vector Machines, *The 20th International Conference on Machine Learning (ICML-2003)*, Washington, DC, pp.752-759 (2003).
  - 29) Valentini, G. and Dietterich, T. G.: An experimental bias-variance analysis of SVM ensembles based resampling, *IEEE Trans Syst Man Cybern B Cybern*, Vol.35, No.6, pp.1252-1271 (2005).
  - 30) Wang, S., Mathew, A., Chen, Y., Xi, L., Ma, L. and Lee, J.: Empirical analysis of support vector machine ensemble classifiers, *Expert Systems with Applications* 36, Vol.38, No.3, pp.6466-64761 (2009).
  - 31) Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng A., Liu, B., Yu, P. S. Zhou, Z-H., Steinbach, M., Hand, D. J. and Steinberg, D.: Top 10 algorithms in data mining, *Knowledge and Information Systems*, Springer London, Vol.14, No.1, pp.1-37 (2008).
  - 32) Zadrozny, B. and Elkan, C.: Transformation Classifier Scores into Accurate Multiclass Probability Estimates, *Proceedings of the Eighth International Conference on Knowledge Discovery and Data Mining (KDD'02)*, pp.694-699 (2002).