5C-1

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

高橋　和子
敬愛大学国際学部

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，バギングやブースティングが代表的である．しかし，文書や画像の分類などで多用されるSVMはバイアスーバリアンス理論でいうバリアンスの占める要素がもともと小さいために，低バイアスのモデルほどにはバギングによる効果が期待できない．またブースティングにおいても，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．

ここで，分類器の正解状況を事例ごとにみると，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観察される．したがって，事例ごとに正解の可能性が高い分類器を選択すれば，単独の分類器より正解事例数が増えることになる．このとき，どの分類器が正解である可能性が高いかを知る方法が問題となるが，提案手法では，クラス所属確率が高いほど正解の可能性が高いと考えた．本稿の目的は，限定されたタスクにより示された提案手法の有効性［4］をより一般化して示すことである．

2. 関連研究

SVMにおけるアンサンブル学習として，[1]はリサンプリングにより分類器を構築し，多数決法，LSE-based weighting, double-layer hierarchical combiningの3つの方法を提案した．2クラスや多クラスのタスクによる実験を行ったが，LSE-based weightingは計算量が多く適用できない場合があった．［5］は bag-of-words を情報利得による素性選択の変化により多様な分類器（2 クラス）を構築し，分類器が出力する分類スコア（分離平面からの距離）の和の大きさによりクラスを決定する方法を提案した．

3. 提案手法

提案手法は次に示すように単純である．

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）に対して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には分類スコアを用いる．推定方法には，ロジスティック回帰式を直接利用するパラメトリックな方法と，「正解率表」を作成して間接的に利用するノンパラメトリックな方法の2つがある．多クラスの場合，いずれも分類スコアを複数個用いることが有効であることが実験的に示されている[3]．

4. 実験

4.1　データセットと分類タスク

今回用いたデータセットは，2005年SSM（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（16,089 サンプル）のうち職業に関するデータおよび，20Newsgroups データセット（18,828 サンプル）の2種類である．

職業データのタスクは，回答を 390 個の国際標準職業分類（ISCO）コード（小分類）に分類する [1]．今回用いたデータセットは調査終了後の作業により ISCO コードが付与済みで [2]，本稿ではこれを正解とした．ISCO コードは階層的で，小分類の上位に中分類（116 個），亜大分類（28 個），大分類（10 個）が存在するため，本稿ではこれらのタスクについての実験も行った．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．素性はネットニュース記事に出現する単語 unigram

Ensemble Learning by Using Class Membership Probabilities with Multi-class SVM
Kazuko TAKAHASHI
Department of International Studies, Keiai University.
takak@u-keiai.ac.jp

1)素性は，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）や「性別」（2 種類の選択回答）および「付与ずみの SSM コード」（約 200 種類）（注2参照）を用いた．

2)このとき国内標準職業分類である SSM コードも付与された．

を用いた．分類精度を変えるために，本稿ではノイズを10%から 20%まで適宜混入させた実験も行った．クラス分布の偏りの程度は，職業データは大きく，20Newsgroups データセットは小さかった．訓練データと評価データの分割は，職業データは 10 分割交差検定，20Newsgroups データセットは5分割交差検定により行った．

4.2 分類器構築，クラス所属確率の推定，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[2]．カーネル関数は線型カーネルを用いた．分類器は[4]での結論より，素性選択ではなくリサンプリングにより 3 個から 21 個まで構築した．

本稿では，クラス所属確率は[3]にしたがってロジスティック回帰式により推定し，分類スコアは第1位から第3位までの予測クラスに付随して出力される 3 個を利用した．ロジスティック回帰式のパラメタ（4 個）推定は，各訓練データをさらに訓練データと評価データに分割して交差検定を行い，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 とみなして最尤推定により行った．

評価尺度は分類精度（全クラスのマクロ平均）を用いた．baseline は，単独で SVM を適用した場合に最も高かった分類精度とした．

4.3 実験結果と考察

提案手法をタスク別構築した分類器数別に多数決法（バギング）や分類スコア法（分類スコアの最も大きな分類器を選択する方法）と比較した（表１，表２参照。分類スコア法は略）．表 2 の上段から順に，ノイズの混入率 0%，10%，12.5%，20%のタスクである．表中 ，太字はタスク別分類器数別に手法間で分類精度が高い方の値で，特に*印は有意な差があることを示す（有意水準 1%）．

表１，表２より，提案手法は分類精度が高い（80%以上）場合は，分類器を増やしても baseline とほぼ同じ値で，有意な差がないものの多数決法に劣る場合があるが，分類精度が低下するほど有効性が高まった．この傾向はクラス分布の偏りが小さいタスクで特に顕著であった．また，提案手法は，分類器が少なく多数決法の効果が現れない段階でも有効性を示した．なお，分類スコア法は提案手法と類似する傾向を示したが，提案手法を常に下回った（表１小分類タスクの分類器 3 個の場合を除く）．理由は，分類スコアは確率ではないため，異なる分類器同士の比較が意味をもたないためであると考えられる．

表１　職業データ　分類精度の比較（単位：%）

分類	手法	baseline	3個	9個	15個	21個
大	提案手法	87.4	87.2	87.2	**88.1**	**87.9**
	多数決法		87.1	87.3	87.4	87.3
亜大	提案手法	84.2	**84.3**	84.4	84.3	84.4
	多数決法		84.2	**84.5**	**84.4**	**84.5**
中	提案手法	79.2	**79.7**	79.7	80.0	80.0
	多数決法		79.5	**80.0**	**80.1**	**80.1**
小	提案手法	73.9	**74.0***	**75.3**	**75.3**	**75.3**
	多数決法		72.6	74.3	74.6	74.7

表２　20Newsgroups データセット　分類精度の比較（単位：%）

手法	baseline	3個	9個	15個	21個
提案手法	87.3	**87.5**	**87.6**	87.6	87.8
多数決法		87.2	87.5	**87.8**	**88.1**
提案手法	79.4	**80.9***	**82.5***	**82.7***	**83.1***
多数決法		79.3	81.3	81.6	81.7
提案手法	77.9	**79.0***	**81.3***	**81.5***	**81.8***
多数決法		77.5	79.9	80.2	80.2
提案手法	69.2	**73.4***	**76.2***	**76.9***	**77.1***
多数決法		71.4	74.6	74.9	74.9

5. おわりに

本稿では，多クラスの SVM における分類精度を高める方法として，事例ごとにクラス所属確率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的なクラスとするアンサンブル学習を提案した．実験の結果，提案手法はクラス分布が偏っていても分類精度が低い（80%未満）タスクほど有効性が高かったため，分類が困難なタスクに向くと考えられる．また，提案手法は構築した分類器が少ない場合にも有効なことやアルゴリズムが単純であることから，分類器の構築に時間を要するようなタスクにも向く．今後の課題は，提案手法が有効である理由を理論的に示すことである．

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．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]H. Kim et al. Pattern Classification Using Support Vector Machine Ensemble.. In ICPR(2), pp. 160-163. 2002.

[2]U. Kressel. Pairwise classification and support vector machines. Advances in Kernel Methods: Support Vector Learning, pp. 255-268. MIT Press. 1999.

[3]K. Takahashi et al. Direct estimation of class membership probabilities for multiclass classification using multiple scores. Knowledge and Information Systems, Vol.19 No.2, pp. 185-210. Springer London. 2008.

[4]高橋和子．クラス所属確率を利用したアンサンブル学習．人工知能学会全国大会（第 24 回）発表論文集．2010.

[5]M. Torii and H. Liu. Classifier ensemble for biomedical document retrieval. In Proceedings of the 2nd International Symposium on Languages in Biology and Medicine. 2007.