2008－NL－188 (4)
2008／11／26

素性選択によるアンサンブル学習に関する一考察

高橋 和子
敬愛大学 国際学部
takak@u-keiai.ac.jp

本稿では, サポートベクターマシンによる文書分類において, 素性を変えて構築した複数の分類器の中から各事例ごとに適切な分類器を選択するアンサンブル学習を提案し, その有効性について調査する. 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約 390 個の国際標準職業分類（ISCO）コードに分類するタスクに適用し, 分類精度について単独の分類器と比較した. その結果, 適切な分類器を選択する方法として, 多数決や分類器の出力するスコア（分類スコア）を用いる場合は有効性を示さなかったが, 分類スコアにより推定したクラス所属確率を用いる場合は, 最も分類精度の高い分類器をわずかに上回ることを確認した.

Ensemble learning by feature selection for Support Vector Machines

Kazuko Takahashi
Faculty of International Studies, Keiai University

We propose ensemble learning by feature selection for Support Vector Machines. In the proposed method, we first make some various classifiers by changing differently features to evaluate samples. We then select the most appropriate one of all the classifiers for each sample by using a voting method, a method using classifier scores, or a method using class membership probabilities. We applied the proposed method to ISCO (International Standard Classification of Occupation) coding which has approximately 390 classes in 2005SSM surveys, and empirically showed that the proposed method was slightly effective by the method using class membership probabilities, while it was not effective by other methods.

1 はじめに

本研究の目的は, サポートベクターマシン（SVM）による文書分類において, 素性を変えて構築した複数の分類器の中から各事例ごとに適切な分類器を選択するアンサンブル学習を提案し, その有効性について調査することである.

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い (Sebastiani, 2002; 麻生他, 2003; 元田他, 2006). アンサンブル学習の代表的な方法としては, 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, 1996; 麻生他, 2003; 元田他, 2006). 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り (麻生他, 2003; 元田他, 2006; Wu et al, 2008), 代表的なアルゴリズムに AdaBoost がある.

バギングやブースティングを, 文書分類において分類精度の高さが評価されている SVM (Joachims, 1998) に適用する場合, 以下の点が問題となる. まず, バギングについては, バイアスーバリアンスの理論 (Breiman, 1996) により, 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 つの部分に分解したとき, 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 and Liu, 2007; 神嶌他, 2008). また, ブースティングでは, SVM において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ため, それに代わる対策が必要

になり (工藤他, 2002; 松田他, 2007; Li et al, 2008), うまく代替できない場合には有効ではない.

複数の分類器があるときに分類精度 (分類器が正解した事例数を全事例で割った値) を全クラスの平均値で比較すると, この値が最も高い分類器は, 全体における正解事例数は最も多いが, 事例ごとにみるとつねに他の分類器を上回って正解するわけではない. すなわち, この分類器が不正解の事例に対して, 分類精度のより低い分類器が正解する場合も観察される. このような場合, 事例ごとに, 正解した分類器を選択しその予測値を最終結果にすれば, 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るであろう.

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アンサンブル学習を提案する. このとき, 構築される分類器は, できる限り多様なクラスを予測することが望ましい. このためには, 訓練事例を変化させるより, より積極的に, 学習のための素性を変化させることが効果的であると思われるため, 本稿では素性選択を変化させる. この点で提案手法はバギングと異なるが, 同時並行的に分類器を構築する点では類似する. また, 提案手法においては, 複数の分類器の中から正解の可能性が高い適切な分類器を選択できることが重要になる. 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

ここでは, 関連研究として次の 2 つの研究について述べる. 一つは, バギングの考え方を利用し, 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, 多様な分類器の構築を提案する (神嶌他, 2008). もう一つは, SVM に注目した場合にバギングが有効ではないとして, bag-of-words に対して情報利得により上位からランキングを行って, 素性として利用する順位を変化させることで多様な分類器の構築を提案する (Torii and Liu, 2007).

いずれも, 有効なアンサンブル学習のために多様な分類器を構築しようとする点で本稿と共通するが, 神嶌らの研究は事例選択である点が本稿と異なり, Torii and Liu の研究は, 素性選択である点は本稿と同様であるが, 本稿では異なる性質をもつ素性の組み合わせを検討する点および, 最終決定の方法として多数決以外の方法も提案する点で異なる.

3 提案手法

提案手法の手順は, 次の通りである.

STEP1　素性の選択方法を変化させて複数の分類器を構築する
STEP2　各事例に対して個々の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されたクラスとする

STEP3 において適切な分類器を選択する方法については 3.1 節で述べる.

3.1 分類器の選択方法

本稿では, クラスを最終的に決定するために適切な分類器を選択する方法として次の 3 つを検討する.

- 多数決による方法 (以下, 多数決法と略す)
- 分類器の出力するスコア (分類スコア) を用いる方法 (以下, 分類スコア法と略す)
- 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)

多数決法はバギングにおいてしばしば用いられる方法で, 最も頻度の多いクラスを予測した分類器を選択する.

分類スコア法は, 各事例に対する結果に付随して出力される分類スコアの中で最も大きな値をもつクラスを出力した分類器を選択する. SVM においては, 分類スコアは分離平面からの距離である. 今回, 2 値分類である SVM を one-versus-rest 法 (kressel, 1999) により多値分類器として拡張したために, 各ランクでクラスが出力され, その分類スコア間に次の関係 : 第 1 位の分類スコア > 第 2 位の分類スコア >··· があるが, 今回は第 1 位の分類スコアのみを対象とする.

クラス所属確率法は, 分類器が事例に対して予測したクラスがどの程度確からしいかを示すクラス所属確率 (Platt, 1999; Zadrozny and Elkan, 2002; Niculescu-Mizil and Caruana, 2005; Langford and Zadrozny, 2005; 高橋他, 2008; Takahashi et al, 2008) を推定し, 最も大きなクラス所属確率の推定値をもつクラスを出力した分類器を選択する. クラス所属確率の推定は, 高橋ら (高橋他, 2008; Takahashi et al, 2008) において第 1 位に予測されたクラスのクラス所属確率の推定方法として提案されたもののうち, ロジスティック回帰を用いた場合 [1] に最も

[1] 他にノンパラメトリックな方法として, 「正解率表」を用いる方法も提案された. この方法は, 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, 評価データの正誤状況から区間に分けた分類スコアごとに正解率を算出した正解率表を作成しておき, 評価事例の分類スコアから正解率表の該当する

よい結果が得られた方法, すなわち第 1 位から第 3 位までのクラスの分類スコア（f_1, f_2, f_3）を利用し, 次のロジスティック回帰式

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

により計算する. ただし, (1) 式におけるパラメタ（4 個）は最尤法により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. 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 (東大教養学部統計学教室（編）, 1992).

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

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 [2].

4 実験と考察

提案手法の有効性を調査するために, 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約 390 個の国際標準職業分類（ISCO）コード (Bureau of Statistics, 2001) に分類するタスク（「ISCO 職業コーディング」 (2005 年 SSM 調査研究会, 2006; 2005 年 SSM 調査研究会, 2007) [3]）に適用し, 単独の分類器と比較を行った.

4.1 実験設定

データセット

用いたデータセットは, 2005 年 SSM 調査データのうち有職者（本人現職, 本人初職, 配偶者職）計 16,089 サンプルである. 訓練データと評価データの分割は, 10 分割交差検定によった. すなわち, 14,480 サンプルを訓練データ, 1,609 サンプルを評価データとし, 訓練データと評価データを変化させて 10 回の実験を行った. ロジスティック回帰式におけるパラメタ推定のためには, 各訓練データごとに 10 分割交差検定を行ってこれを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.

2005 年 SSM 調査における職業データは, 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. 本稿においては, この ISCO コードを正解として扱った.

素性選択

2005 年 SSM 調査データにおける職業データは,「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 種類個の選択回答）,「従業先事業の規模」（13 種類の選択回答）の 4 種類で構成されるが, 高橋ら (高橋他, 2005) にしたがい, 今回も「従業先事業の規模」を除く 3 種類を基本素性とした.

次に, 基本素性に「学歴」（6 種類の選択回答）を追加した. これは, ISCO コードは SSM コードと異なり, 分類の際にスキルという概念が用いられるが, 直接これを測るデータが収集されていないため, 最も近いものとして学歴による代用が妥当であると考えたためである. さらに, これらの素性に, 職業コーディングにより付与された「SSM コード」を追加した. これは, 高橋ら (高橋他, 2005) により, SSM 職業コーディングに適用される機械学習の分類精度向上のために, すでにルールベース手法による自動分類システムが存在する場合にはこの手法により予測されたクラスを機械学習の素性として追加することが有効であることが報告されており, これを参考にしたためである. 今回用いられるクラスは SSM コードとは異なる ISCO コードであり, さらに追加されるコードも対象とするクラスとは別の体系のコードであるが, 職業コードであるという共通点を考慮した.

以上より, 今回は次のような素性をもつ 4 種類の分類器を用いることにした.

- ［分類器 A］仕事の内容, 従業先事業の種類, 従業上の地位と役職
- ［分類器 B］仕事の内容, 従業先事業の種類, 従業上の地位と役職, 学歴
- ［分類器 C］仕事の内容, 従業先事業の種類, 従業上の地位と役職, 正解 SSM コード
- ［分類器 D］仕事の内容, 従業先事業の種類, 従業上の地位と役職, 学歴, 正解 SSM コード

セルを探し, そのセルの正解率を間接的に用いる. 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 (高橋他, 2008; Takahashi et al, 2008).

[2] 高橋ら (高橋他, 2008) によれば, 正規化した値としない値の違いは大きくなかったことが報告されている.

[3] ISCO 職業コーディングは, 近年の国際比較研究の高まりに応じ, これまで実施されてきた我が国独自の職業コードである SSM 職業コード (1995 年 SSM 調査研究会, 1996) を付与するタスク（「SSM 職業コーディング」）に加えて, 新たに実施されるようになった. この結果, 同一のデータに対して 2 種類のコードを付与する作業が必要になり, 作業量の多さと煩雑さの問題がこれまで以上に深刻化したため, より分類精度の高い自動コーディングが要請されている (高橋, 2008).

分類器と評価尺度

今回は，分類器として SVM を用いたが，SVM は 2 値分類器であるために，one-versus-rest 法を用いて多値分類器へと拡張した (kressel, 1999)．カーネル関数は，高橋ら（2005）にしたがって線型カーネルを用いた．また，評価尺度としては，分類精度を用いた．

4.2 予備実験

提案手法についての実験を行う前に，単独の各分類器における分類精度および多様性について調査した．まず，表 1 に各分類器における分類精度を示す．表 1 において，表側の最高，最低は 10 分割交差検定において，分類精度が最も高かった値と最も低かった値を示す．太字は 4 つの分類器の中で最も高い値を示す．表 1 より，単独の分類器における分類精度はつねに，分類器 D ＞ 分類器 C ＞ 分類器 B ＞ 分類器 A の順に高く，分類器 A と B，分類器 C と D で平均の値が類似していた．

4 つの分類器の正解状況は，1 個の分類器だけが正解の場合は 1.9%，2 個の分類器が正解の場合は 12.7%，3 個の分類器が正解（1 個だけ不正解）の場合は 2.5%，すべての分類器が正解の場合は 62.6% であった．これを分類器ごとに調査した結果を表 2 に示す．太字は 4 つの分類器の中で最も高い（よい）値を示す．表 2 から明らかなように，1 個の分類器だけが正解した事例数は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かった．ただし，事例数はいずれも少なかった．2 個の分類器が正解した事例数は，分類器 A と B（以下，グループ a とよぶ），分類器 C と D（以下，グループ b とよぶ）でほぼ同数であった（実際に，各グループ内で正解した事例もほぼ一致していた）．グループをまたがって 3 個の分類器が正解した事例数は少なかった．この点からも，今回用いた分類器は 2 つのグループになり，多様性に欠けていたといえる．今後の課題としたい．

表 1，2 より，事例ごとにみると，必ずしも分類精度の高い分類器だけが正解しているわけではなく，より分類精度の低い他の分類器が正解している場合もあった．各事例において正解した分類器があればその予測クラスを最終決定とすることにすると，分類精度は 79.7% となり（表 2 参照），単独の分類器の中で最も高い分類器 D の値（73.7%）を 5.9% 上回る結果を示した．

4.3 実験結果と考察

提案手法における分類器の選択方法別の結果を，分類精度については表 3 に，正解した分類器の個数ごとの再現率の状況については表 4 に示す．表 3，4 において，多数決法 a はグループ a，多数決法 b はグループ b の分類器をそれぞれ選択する場合を意味する [4]．分類器 D（平均）との差は，分類精度（平均）と単独の分類器の中で最も分類精度の高かった分類器 D の分類精度（平均）との差を示す．また，太字は 4 つの選択方法の中で最も高い（よい）値を示す．表 4 右欄のカバレッジは，ここでは，正解した分類器の個数別の正解事例数が全体の正解事例に占める割合を示す．

表 3 より，全体では，分類器の選択方法としてクラス所属確率法を用いる場合に最も分類精度が高く（73.8%），単独で最も高かった分類器 D を 0.1% 上回った．他の方法による場合は，分類 D を 0.2%（多数決法 b の場合）～1.5%（多数決法 a の場合）下回り，いずれも有効性が示せなかった．クラス所属確率法による場合も，10 分割交差検定のそれぞれで最も高い単独の分類器を最高時 0.9% 上回る場合のあれば，最低時 1.6% 下回る場合もあり，有効性を主張できるまでにはいたらなかった．ただし，単独の分類器における最高値は，最も高い場合 0.767%，最も低い場合 0.723% で両者の間に 0.044% の差があったのに対し，クラス所属確率法による場合は表 3 に示すように 0.020% の差しかなく，安定性があった．

表 4 におけるカバレッジの値より，すべての分類器が正解した場合だけで正解事例の約 8 割近くを占めており，残り約 2 割が，正解した分類器が 1 個から 3 個の場合であることがわかった．この中からできる限り多くの分類器を選択できることが目標である．まず，1 個の分類器だけが正解した場合は，多数決法より分類スコア法とクラス所属確率法の方が再現率が高かった．しかし，その値は 20% に達しておらず，今後，改善方法を検討する必要がある．次に，2 個の分類器が正解した場合に最も再現率が高いのはクラス所属確率法で，次は多数決法 b であった．最後に，3 個の分類器が正解する場合には，多数決法はその定義から必ず正解した分類器を選択することができるため，他の 2 つの方法より優位であった．

今回は提案手法がクラスごとにどのような傾向を示すのかという分析まで行っておらず，考察が十分ではないが，提案手法の有効性が示せなかった理由として次の 3 つが考えられる．

まず，今回は，構築した分類器の個数と多様性に問題があった．　バギングの例ではあるが，Breiman (Breiman, 1996) においては，カテゴリ型の分類の場合にリサンプリングの回数を 50 として実験しており（用いたデータセットのクラスは 3 個，6 個，7 個，26 個），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説明している．今回のタスクにおいては，クラスが約 390 個あるのに対して分類器が 4 個であったことは，個数だけでなく多様性の点からも問題があったと考えられる．多様性については，例えば「性別」や「年代」などのデータも素性として積極的に活用していことが可能であるが，個数については，現在のような素性選択の変化だけで対応することは困難である．

次に，適切な分類器の選択方法として用いたクラス所属確率の推定における精度の問題がある．より

[4] 今回の実験では分類器の数が偶数であったため，多数決法においては同数（2 つ）の分類器が正解した場合にどちらを選択するかを決めておく必要がある．

表 1: 単独の分類器における分類精度

	分類器A	分類器B	分類器C	分類器D	平均	最高	最低
平均	0.689	0.693	0.734	**0.737**	0.713	0.737	0.689
最高	0.704	0.739	0.754	**0.767**	0.741	0.767	0.704
最低	0.684	0.684	0.728	**0.729**	0.706	0.729	0.684

表 2: 各分類器における正解の状況と分類精度

	分類器A	分類器B	分類器C	分類器D	分類器を適切に選択した場合
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
(1 個の分類器だけ不正解)	**(0.001)**	(0.006)	(0.005)	(0.003)	0.025
すべての分類器が正解	**0.626**	**0.626**	**0.626**	**0.626**	0.626
分類精度	0.689	0.693	0.734	**0.737**	**0.797**

高精度な推定のためには, 手間がかかっても適切な正解率表の作成が必要であることがわかった. さらには, クラス所属確率の推定方法そのものに関する研究を深める必要があろう.

最後に, 今回の実験に用いたタスクは, 事例に含まれる素性の数が少ない[5]上に, クラスの数が非常に多く, 難度が高かったために, 提案手法の効果が出にくかったのではないかと考えられる. より容易なタスクとして, 例えば今回と同様の職業データに対し, これまで国内で広く実施されてきた SSM コード（約 190 個）に分類するタスクが存在する. 提案手法をこのタスクを始めより一般的なタスクに対しても適用して, 有効性を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いて, 素性を変化させることで複数の多様な分類器を構築し, その中から各事例ごとに適切な分類器を選択するアンサンブル学習を提案し, その有効性について調査した. 提案手法は, 2005 年 SSM 調査における職業データを用いた分類実験の結果, 適切な分類器を選択する方法としてクラス所属確率を用いる場合にのみ, 単独の分類器をわずかに上回った. しかし, 今後, さらなる方法の改善とタスクを変えた実験を必要とする.

当面の課題として, まず, 本稿で提案手法の有効性を確認するために, 適用するタスクを変えて実験を行う予定である. 次に, より多様な分類器を構築するための素性選択を検討する予定である. また, これらの実験に対して, 分類精度だけでなく AUC（Area Under the ROC Curve）による評価も行う予定である.

[5]高橋ら (高橋他, 2005) によれば, 職業データにおける自由回答（仕事の内容および 従業先事業の種類）はともに 1 文が多く, 両者を併せても平均の長さが 15 文字程度である.

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.

References

1995 年 SSM 調査研究会. 1996. SSM 産業分類 · 職業分類（95 年版）.

2005 年社会階層と社会移動調査研究会. 2007. 2005 年 SSM 日本調査　コード · ブック（95 年版）.

2005 年社会階層と社会移動調査研究会. 2006. 2005 年 SSM 調査　日本 · 韓国 · 台湾調査票.

麻生英樹, 津田宏治, 村田昇. 2003. パターン認識と学習の統計学　新しい概念と手法. 岩波書店.

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, pp. 1–4.

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

表 3: 分類器の選択方法における分類精度

	多数決法 a	多数決法 b	分類スコア法	クラス所属確率法	最高 − 最低
分類精度（平均）	0.695	0.735	0.726	**0.738**	0.043
分類器 D（平均）との差	−0.042	−0.002	−0.011	**0.001**	0.043
分類精度（最高）	0.710	**0.762**	0.748	0.751	0.052
分類精度（最低）	0.685	0.722	0.717	**0.731**	0.046
最高 − 最低	0.025	0.040	0.031	**0.020**	-

表 4: 分類器の選択方法における再現状況

	多数決法 a	多数決法 b	分類スコア法	クラス所属確率法	カバレッジ
1 個の分類器だけ正解	0.037	0.037	**0.177**	0.173	0.02
2 個の分類器が正解	0.340	0.651	0.590	**0.665**	0.16
3 個の分類器が正解	**1.000**	**1.000**	0.848	0.946	0.03
すべての分類器が正解	**1.000**	**1.000**	**1.000**	**1.000**	0.79

用いた Chunk 同定. 自然言語処理　Vol.19 No.5, pp.3–22.

J. Langford and B. Zadrozny. 2005. Estimating class Membership Probabilitiesusing Classifier Learners. In *Proceedings of the Eighth International Conference on Knowledge Discovery and Data Mining (AISTAT'05)*, pp.694–699.

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, 高村大也, 奥村学. 2005. 機械学習とルールベース手法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. 3–24.

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 1 2　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.

K. Takahashi, H. Takamura, and M. Okumura. 2008. 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.

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)*.

東京大学教養学部統計学教室（編）. 1992. 基礎統計学　自然科学の統計学. 東京大学出版会.

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.