Vol.2011-MPS-85 No.13
2011/9/16

多クラス分類のためのデータ分布に基づく階層化手法の提案

久須田　樹哉[†1]　渡　邉　真　也[†2]
施　　建　明[†2]　Paul Horton [†3]

本論文では，2 値分類手法を多クラス分類へ拡張するための新たな階層的分類メカニズムの提案を行う．従来までに提案されている One-Against-One（OAO）や One-Against-All（OAA）と異なり，提案手法では学習用として事前に与えられた各クラスのデータ分布に基づいた分類の階層化を行っている．そのため，従来までの手法に比べより高い精度の識別が期待でき，特に少数データからなるクラスの識別においてその効果を期待できると考えている．また，階層化構造がクラス間の類似度を表わすことになるため，得られた分類構造から相対的なクラス特性を推定することができる．本論文では，UCI レポジトリに含まれるいくつかの例題に対して，k-NN，OAO に基づく多クラス SVM との比較実験を行い提案する階層化メカニズムの有効性の検証を試みた．

A proposal of hierarchization method based on data distribution for multi class classification

Tatsuya Kusuda,[†1] Shinya Watanabe,[†2]
Jianming Shi[†2] and Paul Horton[†3]

This paper proposes a new hierarchical method for multi-class classification using binary classifiers. Unlike existing extension methods, such as One-Against-One (OAO) and One-Against-All (OAA), our proposed method makes a hierarchic structure of classification according to distributions of each class data as given training data. Thus, a more accurate multi-class classification can be expected than by existing methods. In particular, our proposed method is expected to be more effective for classes with few samples. Since a hierarchical structure of classification derived by our proposed method is formed on the basis of similarities among classes, relative features of each class can be inferable through the classification structure. In this paper, the effectiveness of our proposed method is discussed through some examples from UCI repository, based on comparison with that of k-NN and OAO.

1. はじめに

SVM，ニューラルネットに代表される識別手法に関する研究は数多く行われており，中でも多クラス分類はその応用範囲の広さと複雑さから様々な手法が提案されている[1)–4)]．そのアプローチは k 近傍法（k-nearest neighbor algorithm，k-NN）に代表される一層多クラス識別法を用いる方法と SVM などの 2 値分類手法を多クラスへ拡張する方法の 2 つに大別する事ができ，サポートベクターマシン (Support Vector Machine, SVM)[1)] などの 2 値分類手法を利用した One-Against-One（OAO）や One-Against-All（OAA）[5)] は後者の代表的な手法として広く一般的に利用されている[6)]．

OAO および OAA は単純なメカニズムでありながら比較的高い識別率を実現することができる一方，識別手順にクラス間の類似性といった学習データの情報が考慮されず，最適な階層化が実現されていないという問題点がある[5)]．

そこで本研究では，多クラス識別においてクラス間の特性の差異を考慮した新たなアプローチとして，各クラスのデータ分布に基づく階層化手法の提案を行う．

提案手法は，学習データの各クラスの分布情報に基づき段階的に大まかな分割から徐々に詳細な分割を実現する階層的な分類メカニズムに基づいている．具体的には，各階層において各クラスがどちらか一方のグループに完全に振り分けられ，クラス全体がより均等になる様な特徴空間の算出を行い，階層的にそれらを繰り返す事で階層的クラス分類を実現している．

提案手法の実現により，クラス間の類似性を考慮したクラス階層化が可能となり下記のメリットを期待する事ができる．

- 各クラス間の近接度合いに関する特性可視化．
- 詳細な分類分けによる識別精度の向上．

本提案手法の適用により得られる階層構造には，クラス間の大局的な類似度，各階層の特徴づけに関する情報といった対象データの特性が明確化されていると考えられる．また，

†1 室蘭工業大学 大学院　情報電子工学系専攻
Graduate School of Information and Electronic Engineering, Muroran Institute of Technology
†2 室蘭工業大学　しくみ情報系領域
Department of Information and Electronic Engineering, Muroran Institute of Technology
†3 産業総合研究所　生命情報工学研究センター
National Institute of Advanced Industrial Science and Technology Computational Biology Research Center (AIST CBRC)

データ分布に基づく階層化分類の実現により精度向上，特にデータ数の少ないクラスに対する識別精度向上を期待することができる．

提案手法の有効性を検証するために，UCI レポジトリ[7] から引用したデータを対象に重み付け k-NN，多クラス SVM（OAO）との比較実験を行った．実験では，既存手法に対する識別性能の優位性を検討するとともに，本手法の適用により得られた階層構造に対する妥当性についても検証を行い提案手法のデータ特性の可視化に関する有用性を考察した．

本論文の構成を示す．まず，第 2 章において提案する階層型多クラス分類手法について説明する．第 3 章では UCI レポジトリに含まれるいくつかの例題に対する数値実験及びその結果と考察について述べ，最後に，本研究の結論を示す．

2. データ分布に基づく階層化手法

提案する階層化手法は，データの分布特性に基づいて段階的に分類の粒度を細かくする思想に基づいており，階層上位において大まかなグループ分割を行い，下位の階層では類似したクラス間での分割を行う．具体的には，類似クラスのクラスタリングによるグループ統合を行い，グループ間分類を実現している．提案手法の概念図を**図 1** に示す．提案手法では図 1 に示す木構造の階層型学習器を生成している．

図 1 から分かる様に，提案手法ではデータの分布から類似しているクラス同士をグループ化しグループ間で分類を行うという操作を繰り返すことで階層化分類を実現している．図中の例では，1 層目においてクラス 1・3 のグループとクラス 2・3・4 のグループを変数増加法により求めた最適な特徴空間で分割し，2 層目においてクラス 1 とクラス 3，クラス 2 とクラス 3・4 を分類，最下層の 3 層目ではクラス 3 と 4 を分類している様子が示されている．

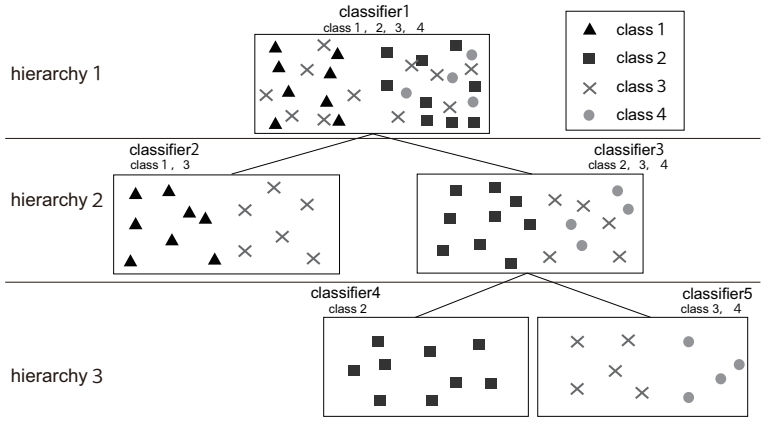

図 1 提案手法の概念図

Conceptual diagram of proposal method

本手法において重要なのは，階層構造からクラス間の相対的な類似度が分かるだけでなく，各階層の分類器ごとに特徴空間（類似度）の定義が異なるため，どの属性がクラス全体の特徴づけに対してどのような影響を持っているのか把握できる点である．また，下の階層ではより詳細な分類が実現されることになるため，従来手法ではデータ数の多いクラスに埋もれてしまうデータ数の少ないクラスに対する高精度識別を期待することができる．

2.1 提案アルゴリズムの手順

提案する階層化手法は学習用データセットから階層型学習器を生成する過程とテストデータの予測を行う過程の 2 段階に分類する事ができる．階層型学習器の生成とテストデータに対する予測の具体的な手順についてそれぞれ以下で説明する．

階層型学習器の生成

階層型学習器を生成する具体的な手順を以下に示す．

Step1: 学習用データセットの入力．

Step2: 得られたデータ分布からデータをクラスタリングによってグループ化し，グループ間の分割を実現する最良な特徴空間を算出（2.2 節）．

Step3: Step2 で求めた類似度（選択した特徴量および重み）に基づいて SVM で分類．

Step4: Step3 で分類されたデータに対して，下階層の分類を行うかを判定．分類を行う場合には Step2 へ，そうでなければ階層型学習器の生成を終了．

テストデータに対する予測

テストデータの予測を行う具体的な手順を以下に示す．

Step1: 学習用データに基づき生成した階層型学習器に対して，テストデータを入力．

Step2: 各階層においてテストデータがどの分岐に属するか SVM で識別．

Step3: 最下層において OAO を適用した SVM により最終的な推測結果を出力．

本手法の核となっているのは階層型学習器の生成における Step2 のクラスのグループ化である．次節において，このグループ分割メカニズムの詳細について説明する．

2.2 階層化のためのグループ分割

提案する階層化手法では，類似したクラスをクラスタリングによりグループに統合し，グループ間分類を行っている．具体的には，2-means（k-means，k=2）によって得られたグ

ループ毎に各クラスのデータ含有率を求め，含有率が閾値以上のクラスを同一グループに決定している．しかし，単純な 2-means を使ったグループ分割では，初期点依存の問題が生じるため，2-means を複数回試行しグループ内分散とグループ間分散の比が最小となるグループ分割を採用している．

本手法では，グループ分割を行う際に変数増加法[8] を用いることで，グループ分類を実現する特徴空間の算出を行っている．本手法では，変数増加法を選択するかしないかの 2 値ベクトルではなく，0 以上 2 未満の整数ベクトル（0,1,2）とすることで重み付けも行えるようにし，ビームサーチの概念を導入することで設定したビーム幅の分だけ特徴の組み合わせを保存しながら最適な特徴空間を探索できるように拡張している．さらに，変数増加法を用いて特徴選択を行う場合，予測精度だけで評価を行うが，本手法では各クラスが 2 グループに精度良く分離しグループに含まれるクラス数がより均等になるよう式（1）に示す評価式を定義した．式（1）を用いる事により，単に精度のみが良くなる様なグループ化ではなく，極力グループ間にクラスが均等に分かれる様なグループ化を実現することができる．これは，階層構造をよりシンプルに保つため階層が深くなりすぎることを防ぐためである．また，クラスがどちらのグループに属するか判断できない場合には両方のグループに属する事も許容している．

以下，上記で述べた変数増加法に基づくグループ分割の手順を示す．

Step1: データを 2-means で 2 個のグループにクラスタリング.

Step2: Step1 で生成したグループ毎に各クラスのデータ含有率を算出.

Step3: グループ毎にデータが $\alpha\%$ 以上含まれるクラスを所属クラスとし，所属クラス以外のデータを削除.

Step4: 片方のグループだけに所属しているクラスの数 $\sharp N_{G_j}$ を算出.

Step5: 評価値を以下の式で算出．以下では $\sharp C$ はクラス数，$\sharp G$ はグループ数，Balance はグループ間のクラスの均等性，Accuracy は予測精度を意味するものとする.

$$\mathrm{Eval} = \beta \times \mathrm{Balance} + (1-\beta) \times \mathrm{Accuracy} \tag{1}$$

$$\mathrm{Balance} = \prod_{j=1}^{\sharp G} (\sharp N_{G_j} \times \frac{\sharp G}{\sharp C}) \tag{2}$$

$$\mathrm{Accuracy} = \frac{\text{正解したデータ数}}{\text{全データ数}} \tag{3}$$

Step6: Step1〜Step5 を特徴毎に行い，評価値が最も高い特徴を選択.

Step7: 選択した特徴と残りの特徴を組み合わせて Step1〜Step6 を実行.

Step8: 評価値に向上が見られなくなったら終了.

式 (1) の評価式では第 1 項で所属クラス数のバランスを評価し，それに第 2 項の予測精度を加えている．つまり，所属クラスのバランスが良く，予測精度が高いグループが高評価を得る様になっている．パラメータ β は両項の重み付けを表しており，本論文の実験では 0.3 を用いた．また，Step3 におけるパラメータ α は，グループ分割によるクラス分離の精度の下限を表している．つまり，α の値が高ければグループ分割が実現しやすくなり，逆に低ければ完全にクラスがどちらか一方に分離される場合しかグループ分割が生じなくなる．

3. 数 値 実 験

提案する階層化手法の有効性を検証するために UCI レポジトリに含まれる幾つかのデータを対象に，重み付き k-NN，SVM を使用した OAO，提案手法の比較実験を行った．

3.1 対象データ

対象データとして UCI レポジトリ[7] の 6 つのデータ（Iris,Wine,Heart Disease,Glass,Vowel,Car Evaluation）を使用した．各データの特徴を表 1 に示す.

3.2 使用パラメータについて

提案手法では以下の 5 つのパラメータを使用する.

- グループ判定パラメータ α：グループ分割時にグループに所属するクラスを判定するための閾値.
- 評価式の重みパラメータ β：グループ分割時の評価式 Balance の重み（式 (1)）.
- 交差検定の分割数：精度計算に使用する交差検定の分割数.
- クラスタリング回数：グループ分割時に 2-means を試行する回数.

使用する各パラメータの値を表 2 に示す．ただし，Car Evaluation においては α が 0.2 では全てのクラスが両方のグループに重複してしまいグループ分割ができなかったため，α の値を 0.25 として使用した.

3.3 実 験 結 果

UCI レポジトリの 6 つのデータ（Iris，Wine，Heart Disease，Glass，Vowel，Car Evaluation）に対して提案する階層化手法を適用し，識別性能および得られた階層構造の妥当性についての検証を行った.

⋆1 Car Evaluation のみ 0.25

表 1 使用データの特徴
The characteristics of the used data

Dataset	Number of data	Features	Classes
Iris	150	4	3
Wine	178	13	3
Heart Disease	270	13	5
Glass	214	10	6
Vowel	528	10	11
Car Evaluation	1728	7	4

表 2 使用パラメータ
Used parameters

Parameter	Values
Determination of group parameter α	$0.2^{\star 1}$
Weight of evaluation formula parameter β	0.3
Number of partitions in cross validation	10
Number of clustering	30

表 3 予測精度結果
The results of predictive accuracy

Dataset	Weighted k-NN	OAO	Proposed method
Iris	95.33%	96.00%	**97.33**%
Wine	93.26%	97.19%	**97.75**%
Heart Disease	55.56%	55.22%	**56.57**%
Glass	65.42%	66.36%	**70.56**%
Vowel	92.02%	**99.60**%	95.04%
Car Evaluation	88.60%	**99.36**%	99.07%

表 4 クラス毎の予測精度（Glass）
The predictive accuracy of each class in Glass

Id	Number of data	Weighted k-NN	OAO	Proposed method
1	70	81.43%	71.43%	**84.29**%
2	76	67.11%	**76.32**%	71.05%
3	17	0.00%	11.76%	**17.65**%
5	13	46.15%	**53.85** %	46.15%
6	9	22.22%	22.22%	**44.44**%
7	29	82.76%	79.31%	**86.21**%
全体	214	65.42%	66.36%	**70.56**%

表 5 α による予測精度への影響
The effects of α value for predictive accuracy

Dataset	Weighted k-NN	OAO	Proposed method				
α values			0.4	0.3	0.2	0.1	0.0
Iris	94.00%	94.67%	96.00%	95.33%	95.33%	96.00%	**96.67**%
Wine	93.26%	97.19%	97.75%	97.19%	97.75%	**98.31**%	97.19%
Heart Disease	54.88%	54.88%	54.88%	54.88%	**55.22**%	54.88%	54.88%
Glass	65.42%	64.95%	64.49%	64.49 %	64.95%	**67.76**%	64.95%
Vowel	92.02%	**99.60**%	95.45%	95.15%	95.05%	95.75%	**99.60**%
Car Evaluation	82.81%	**99.71**%	99.42%	99.54%	**99.71**%	**99.71**%	**99.71**%

3.3.1 識別性能に関する検証

識別精度の観点から検証を行う．10 分割交差検定 (10-fold Cross-Validation) を用いた場合の各データに対する重み付き k-NN，SVM を用した OAO，SVM を提案する階層化手法に適用した場合の予測精度結果を表 3 に示す．なお，表中における太字は，3 手法のうち最も高い予測精度であることを意味する．

表 3 より，提案手法は Iris，Wine，Heart Disease，Glass データにおいて予測精度が最良である一方，Vowel，Car Evaluation データでは OAO に劣っていることが分かる．

Vowel，Car Evaluation データの予測精度が OAO に劣った原因は，変数増加法に基づく分離では一定の誤りが生じてしまうためと考えられる．これらの原因については，後述のパラメータ α に関する実験でより明らかとなる．

また，提案手法における階層化では，データ分布に基づくグループ分類が実現されているためデータ数の少ないクラスに対する予測精度の向上を期待する事ができる．この点を確認するため，クラスに含まれるデータ数にばらつきがある Glass データに対して詳細な分析を行った．Glass データの各クラスの予測精度を表 4 に示す．

表 4 より Glass データの各クラスにおいてデータ数が一番多いクラス 2 では OAO より予測精度が下回っているものの，データ数の少ないクラス 3，6，7 で提案手法の予測精度が他の手法を上回る結果となった．このことから，提案手法が相対的にデータ数の少ないクラスの予測精度向上に有効であることが確認できた．

パラメータ α の結果への影響について

提案手法では α によりグループ分割におけるクラス分離精度の下限を設定している．α の値を高くした場合，グループ分割が生じやすくなり階層構造が生成されやすくなる一方，グループ分割時の誤差が大きくなってしまうため予測精度の低下を招くと思われる．そこで，α の値による予測精度に対する影響の検証を行った．各データに対して α の値を変更した場合の予測精度を表 5 に示す．

表 5 より，α の値を下げることで予測精度の向上が見られた．α の値を下げると両方のグループに重複するクラスが増え，片方のグループにのみ所属するクラスの判定は厳しくなる．そのため，全てのクラスが重複し階層化されないケースも見られるが，分岐による分類ミスが軽減されるため予測精度が向上したと思われる．また，最下層の識別に OAO を使用

Vol.2011-MPS-85 No.13
2011/9/16

表 6 Glass データのクラス
Classes of Glass data

id	Details of classes
1	building_windows_float_processed
2	building_windows_non_float_processed
3	vehicle_windows_float_processed
5	containers
6	tableware
7	headlamps

表 7 Glass データの属性
The attributes of Glass data

id	Features name	Details of features
1	RI	refractive index
2	Na	Sodium
3	Mg	Magnesium
4	Al	Aluminum
5	Si	Silicon
6	K	Potassium
7	Ca	Calcium
8	Ba	Barium
9	Fe	Iron

しているためパラメータ値を厳しくする事で少なくとも OAO と同じ予測精度を実現できる事が確認された．

3.3.2 階層化構造に対する検証

提案手法のデータ分析ツールとしての有効性を検証するために Glass と Heart Disease に対して階層化手法を適用し，得られた階層構造について考察する．階層構造の検証のための実験では，対象データの全データを用いて階層型学習器の生成を行った．

Glass データは UCI レポジトリにあるガラスの酸化物含有量に関するデータであり，このデータのクラスと属性をそれぞれ表 6，表 7 に示す．Glass データに対して提案手法を適用し得られた階層図を**図 2** に示す．図 2 では提案手法によって得られた階層型分類器を木構造の形で可視化しており，各階層におけるクラスタには分類されたデータの数とグループ分割で分類されたクラスが記されている．さらに，各分岐において分類に使用された属性についても記されている（太字で書かれている属性は重みを 2 に設定）．

図 2 よりクラス 1，2，3 の建物や車の窓ガラスは類似した特徴を持っていることが分かる．そのため，他のヘッドランプ，食器，容器について順に分類していく結果となった．また，各階層での分類使用属性からどのクラスがどのような属性で特徴づけられているのか分かるとともに，クラス 1，2，3 間ではこれらの属性だけでは適切に分離できないことも読み取れる．

次に表 8，表 9 に示す心疾患に関する Heart Disease データに対する結果について考察する．Heart Disease データに対して提案手法を適用する事で得られた階層図を**図 3** に示す．図 3 では図 2 と同様に提案手法によって得られた階層型分類器を木構造の形で可視化している（クラス及び属性の id 番号の対応については**図 9** を参照）．

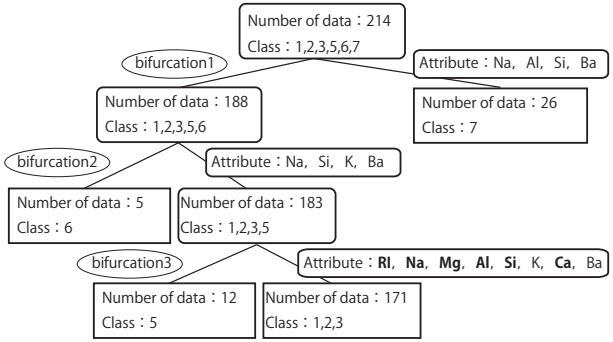

図 2 提案手法の階層図（Glass）
Hierarchy diagram of proposal technique(Glass)

表 8 Heart Disease データのクラス
Classes of Heart Disease data

Classes id	Details of classes
0	person who hasn't heart disease
1〜4	person who has heart disease（1〜4 is degree）

表 9 Heart Disease データの属性
The attributes of Heart Disease data

Features id	Features name	Details of features
1	age	age in years
2	sex	sex (1 = male; 0 = female)
3	cp	chest pain type
4	trestbps	resting blood pressure
5	chol	serum cholestoral in mg/dl
6	fbs	fasting blood sugar > 120 mg/dl
7	restecg	resting electrocardiographic results
8	thalach	maximum heart rate achieved
9	exang	exercise induced angina
10	oldpeak	ST depression induced by exercise relative to rest
11	slope	the slope of the peak exercise ST segment
12	ca	number of major vessels (0-3) colored by flourosopy
13	thal	3 = normal; 6 = fixed defect; 7 = reversable defect

Vol.2011-MPS-85 No.13
2011/9/16

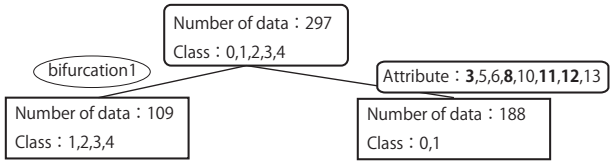

図 3 提案手法の階層図（Heart Disease，0.2）

Hierarchy diagram of proposal technique(Heart Disease，0.2)

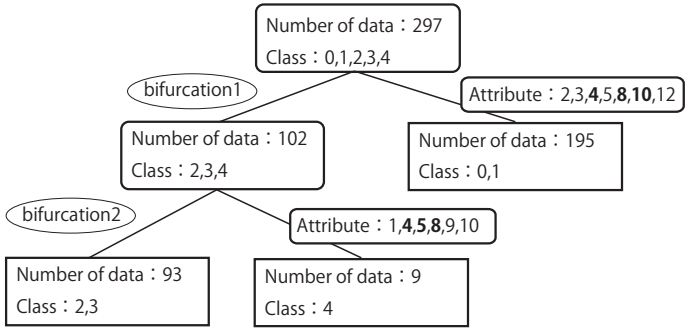

図 4 提案手法の階層図 (Heart Disease，0.4)

Hierarchy diagram of proposal technique(Heart Disease，0.4)

図 3 の分岐においてクラス 2，3，4 とクラス 0 が分割されているため，心疾患であるかどうかを分類していると読み取ることができる．ただし，クラス 1 が重複している事から，軽度の心疾患の患者を分類する事は難しい事が分かる．

次に，より詳細な問題分析を行うために α の値を 0.4 に緩和し，より階層分類を生じやすくした場合について実験を行った．結果を**図 4** に示す．

図 4 ではクラス 1（軽度の心疾患）がクラス 0（正常）の側に所属されているため，軽度の心疾患の患者の症状は正常の人と見分ける事が難しい事が分かる．さらに，属性 4，5，8，10 が分岐 1，分岐 2 の両方で選択されている事から，血圧や血清値，最大心拍数などは心疾患によって変化しやすいと推測される．

4. おわりに

本研究では，データ分布に基づく新たな階層化手法を提案した．提案手法はクラス間の類似性に基づく分類階層化を最大の特徴としており，単なる予測精度の向上だけでなく，クラス間の近接度合いなどの問題特性の可視化を目的としている．提案手法の有効性を検証するため，UCI レポジトリに含まれるいくつかの例題に対し重み付き k-NN，OAO との比較実験を行い，以下の事柄を明らかにすることができた．

- データ数の少ないクラスに対して，提案手法は特に効果的．
- パラメータ調整により，少なくとも OAO と同等以上の予測精度を実現．
- 提案手法で得られた階層構造結果による，クラス間の類似性，クラス間分類に本質的に効いている属性の明確化

今後は提案手法の単純化について検討を行い，より大規模なデータに対する応用を進めたいと考えている．

参 考 文 献

1) G. Nalbantov P.J.F.Groenen and J.C.Bioch. a majorization approach to linear support vector machines with different hinge errors. Advances in Data Analysis and Classification, Vol. 2, pp. 17-43, 2008.
2) J.A.K. Suykens and J. Vandewalle: Least squares support vector machine classifiers, Neural Netherlands, Vol. 9, No. 3, pp. 293-330, 6.1999.
3) Doumpos, M. Zopounidis, C. Golfinopoulou, V: Additive Support Vector Machines for Pattern Classification, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, Vol. 37, No. 3, pp. 540-550, 2007.
4) Fukumizu, K: Special statistical properties of neural network learning. Proc, NOLTA'97, pp. 747-750, 1997.
5) Jonathan Milgram and Mohamed Cheriet and Robert Sabourin: " One Against One " or " One Against All ":Which One is Better for Handwriting Recognition with SVMs?, Ecole de Technologie Superieure, Montreal, Canada, 2006
6) Chih-Wei Hsu Chih-Jen Lin: A comparison of methods for multiclass support vector machines, Neural Networks, IEEE Transactions on, Vol. 13, No. 2, pp. 415-425, 2002.
7) C.L.Blake and C.J.Merz: UCI repository of machine learning databases, University of California, Department of Information and Computer Science, 1998, http://www.ics.uci.edu/MLRepository.html
8) 森 裕一, 垂水 共之, 田中 豊：変数の一部に基づく主成分分析：変数選択手法の数値的検討, 計算機統計学,1988
9) 田村 坦之，益永 健一郎，鳩野 逸生，馬野 元秀，外嶋 成留，杉原 誠，平山 克己，中川 義之：遺伝的アルゴリズムとラグランジュ緩和法を併用したビーム探索法によるスケジューリング問題の解法, シンポジウム 日本オペレーションズ・リサーチ学会,1994
10) Baldi P et al: Assessing the accuracy of prediction algorithms for classification, an overview, Bioinformatics, Vol. 16, No. 5, pp. 412-424, 2000