様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ５ 月　 ３ 日現在

機関番号：１４５０１
研究種目：基盤研究(C)（一般）
研究期間：2013～2015
課題番号：２５４２０４３８

研究課題名（和文）大次元小サンプル問題における特徴選択方式の開発

研究課題名（英文）Development of feature selection methods for large-input, small-sample problems

研究代表者
阿部　重夫（Abe, Shigeo）

神戸大学・工学研究科・名誉教授

研究者番号：５０２９４１９５

交付決定額（研究期間全体）：（直接経費）　　2,500,000円

研究成果の概要（和文）：大次元小サンプルパターン認識問題問題における特徴選択の研究を行い，以下の結論を得た。これまでに開発したブロック追加・削除（BABD）の特徴選択方式をさらに高速化するために，BABDを繰り返して適用する繰り返しBABD法と特徴量の集合をブロックに分割して，ブロックごとにBABDを適用する追加BABD法を開発した。特徴選択で用いるサポートベクトルマシン（SVM）の学習方法としてSMO（Sequential Minimal Optimization）法とニュートン法を組み合わせたSMO-NM法を開発した。ベンチマークデータを用いて提案方式の有効性を検証した。

研究成果の概要（英文）：We developed feature selection methods for small-sample, large-input problems. To speed up the block addition and block deletion (BABD) methods developed previously, we developed the iterative BABD that repeatedly iterates BABD and incremental BABD that applies BABD to a block of features divided in advance. To speed up training of support vector machines that are used for feature selection, we developed the SMO-NM method that combines the sequential minimal optimization technique and the Newton Method. We evaluated the validity of the proposed methods using benchmark data sets.

研究分野：システム工学

キーワード：パターン認識　特徴選択　ブロック追加　ブロック削除　サポートベクトルマシン

１．研究開始当初の背景

パターン認識システムの汎化能力（未知のデータに対する識別能力）はそれに用いられる入力変数，すなわち特徴量に大きく依存し，パターン認識システムが成功するか否かはどのような特徴量を選ぶかにかかるといっても過言ではない。特徴選択方式としては，認識率を特徴選択基準とするラッパー法，認識率以外の特徴選択基準によるフィルター法などがある。一般に，ラッパー法は得られる特徴量はよいが選択に時間が多くかかる問題がある。これに対してフィルター法は，ラッパー法よりも高速であるが，得られる特徴量が劣る。

ラッパー法による特徴選択を高速化する方法として，研究代表者らは複数個の特徴量を一度に追加するブロック追加（BA: Block Addition）の後に，複数の特徴量を一度に削除するブロック削除（BD: Block Deletion）を行うことにより，特徴選択を格段に高速化するBABD 方式を開発している。しかも，すべての特徴量を使用したときの認識率を閾値として，この閾値より，特徴量の選択により認識率が向上するときは，閾値を更新し，低下するときに特徴選択を終了する方式を開発した。これにより，認識に無関係な特徴量により，識別能力が低下する場合も，閾値を更新することにより，元の特徴量の集合に対して，選択された特徴量の集合は交差検定におけるデータに対して，必ず認識率が向上することが保証される。

２．研究の目的

(1)　大次元小サンプル問題に対してブロック追加・削除により特徴量を選択する方式を開発する。認識率にマージン誤差を加えた選択基準では，小サンプル問題において，テストデータに対する認識率が，元の特徴量を使ったときよりも低下する場合が起こりうる。このため，元の特徴量の集合に対するテストデータの認識率と同等以上の認識率を実現する特徴選択方式を開発することを目標とする。

(2)　方式の高速化を図る。(1)の方式は大次元小サンプル問題でなくても適用可能であり，1000～10,000 データ程度のパターン認識問題で，従来法に対して 10～100 倍高速化する方式を開発する。

(3)　マハラノビス距離の学習により特徴選択する方式を開発する。10,000 データ程度を高速に学習する方式を開発し，種々のベンチマークデータで実証する。

３．研究の方法

(1)　大次元小サンプル問題に対してブロック追加・削除により特徴量を選択する方式を開発する。　大次元小サンプル問題では，教師データによって汎化能力が大幅に変わるために，教師データの依存性が低い汎化能力の向上方式を開発する。

(2)　方式の高速化を図る。ブロック追加・削除方式では連立方程式を解いて最小二乗(LS) SVM を学習しているために，データ数が多くなると学習が遅くなる。LS SVM (Least Squares Support Vector Machine)も通常の SVM と同様に，繰り返し法で学習できることが示されており，通常の 2 個変数同時処理方式 (SMO: Sequential Minimal Optimization)ではなく，複数変数同時処理方式を開発して，1,000～10,000 個のデータに対して連立方程式の求解より 10～100 倍以上の高速化を図る。

(3) マハラノビス距離の学習により特徴選択する方式を開発する。線形計画法の分割法を適用して高速化したが，線形計画法によらずに，高速に学習する方式を開発する。

４．研究成果

研究の目的の(1)－(3)に対応した研究成果を順次以下に説明すする。

(1) 大次元小サンプル問題に対してブロック追加・削除により特徴量を選択する方式を開発する。

① これまでに開発したブロック追加・削除（BABD）の特徴選択方式では，交差検定による認識率が低下しない範囲で高速に特徴選択ができる。BABD で選択される特徴量をさらに削減するために，BABD を繰り返して適用する繰り返し BABD 法を開発した。特徴量数が 3,226 から 12,625 の９種類のマイクロアレーデータで計算機実験を行い，提案方式を評価した。その結果，7 個のマイクロアレーデータで通常の BABD に対して繰り返し BABD の効果があり，通常の BABD が 62～426 個の特徴量に削減できたのに対して，ほとんど汎化能力を低下せずに 17～88 個の特徴量に削減できた。

② 特徴量の集合をブロックに分割して，ブロックごとに BABD を適用して高速化を図る追加 BABD 法を開発した。また小サンプル問題で認識率を特徴選択基準にすると，複数の候補で選択基準が同じになる場合があるため，パターン認識問題を関数近似問題として，近似誤差を特徴選択の評価基準として採用して，複数候補が起こりにくいようにした。①と同じマイクロアレーデータを用いて計算機実験を行い，選択される特徴量数は同程度で BABD に対して最大で 4 倍程度の高速化が得られることを確かめた。

図１は乳がんデータに，追加特徴量数を変えて提案方式を適用したときのテストデータの認識率の変化を示している。ここで，乳がんデータは特徴量数が 3,226 個，教師データ数が 14 個，テストデータ数が 8 個で，教師データとテストデータの組み合わせをラ

ンダムに変えて，100 回の試行行い，認識率の平均値と標準偏差を求めた。図中の One pass, Multi pass は特徴量の追加を 1 回だけか，複数回繰り返して行ったかを示しているが，どちらの場合も追加特徴量を変えても，認識率はさほど変化していないことが分かる。

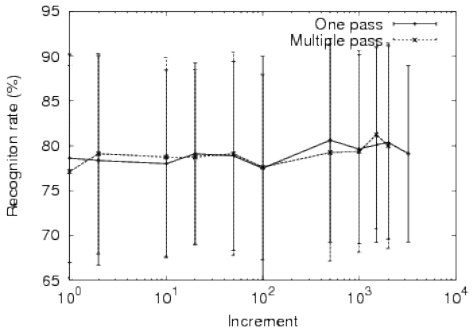

図１　乳がんデータに対して追加特徴数を変えたときの認識率の変化（雑誌論文⑤, Fig. 1 (b) 許諾を得て掲載 Copyright © 2014, Springer International Publishing Switzerland）

図２は，図 1 の条件で特徴選択したときの，特徴量の選択時間を示している。ここで特徴選択時間とは，識別器のパラメータを決定するための交差検定を含んだ時間である。Multi pass の方が選択時間が長くなるのは当然であるが，どちらの場合も，100 個付近に追加特徴量数の最適値があることが分かる。

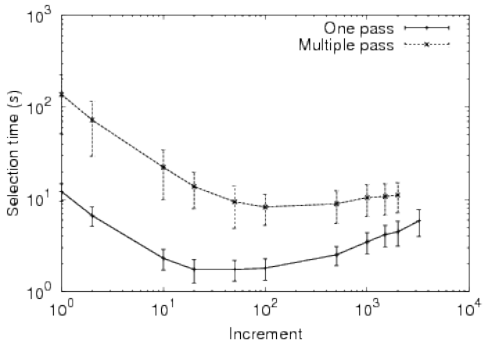

図 2　乳がんデータに対して追加特徴数を変えたときの特徴選択時間の変化（雑誌論文⑤, Fig. 1 (c) 許諾を得て掲載 Copyright © 2014, Springer International Publishing Switzerland）

③ 追加 BABD 法を関数近似に拡張した。このとき，特徴量を追加して BABD を実行したときに汎化能力が低下する可能性がある。そのような場合は，その追加 BABD を無効化する（追加する前の状態に戻す）ことにより，汎化能力の低下を防止した。この方式を入力数が 27〜7,129 個の関数近似問題に適用し，ほとんど汎化能力を低下することなく，選択される入力数をさらに減少させることを確認した。このとき，最大で 3 倍程度 BABD より高速化できた。

(2) 方式の高速化を図る。

① サポートベクトルマシン（SVM）の学習方法として SMO 法とニュートン法（NM: Newton Method）を組み合わせた SMO-NM 法を開発した。この方式では，SMO において同じ変数が繰り返して修正されるときに，ループが検出されたと判定して，その間に修正された変数をワーキングセットに設定して NM 法で学習を行う。これにより，ワーキングセットの大きさが，自動的に設定されるために，学習を高速化できる。教師データが 6,197〜20,000 個の 8 個のパターン認識問題を用いた計算機実験において，最大で SMO に対して 100 倍の高速化が得られた。

図１に血球データに対する提案方式と従来方式の学習時間の比較を示す。従来方式ではマージンパラメータの値が大きくなるにしたがい，学習時間が増加するが，提案方式ではマージンパラメータの値の変化に対して，学習時間はあまり変化しないことが分かる。

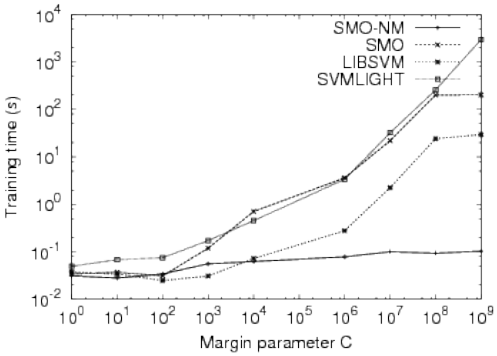

図１　血球データに対する提案方式（SMO-NM）と他方式の学習時間の比較（雑誌論文④, Fig. 10 (a) 許諾を得て掲載 Copyright © 2014, Springer-Verlag Berlin Heidelberg）

② 上記の SMO-NM 方式を絶対値を含む関数近似の学習に拡張した。最適値が正負のどちら側になるかの情報を用いて絶対値の微分を行うことにより，微分が不定になる問題を解決した。ワーキングセットの選択はパターン認識における方法を用いた。入力数が 27 個から 7,129 個の 7 つのベンチマークデータで学習時間の比較を行い，SMO に対して提案方式が多くのマージンパラメータの値に対して高速化されることを確かめた。

③ 追加 BABD，および識別器の学習の高速化に加えて，並列処理による高速化の可能性を検討した。BABD および SMO-NM の並列処理による高速化の可能性を検討し，BABD では 1 個の特徴量の追加あるいは削除において並列処理ができること，また SMO-NM では逐次処理の候補を計算が並列処理できる

ことが分かった。並列処理計算により高速化が可能かどうかの，実機での検証は今後の課題とする。

(3) マハラノビス距離の学習により特徴選択する方式を開発する
① SVM ではデータ分布に対する仮定を行っていないために，データ分布に対する事前情報がある場合, SVM の汎化能力を向上できる可能性がある。その１つとしてマハラノビス距離の導入がある。研究着手当初は，マハラノビス距離を陽に表現することが，最善であると考えていたが，その後のベンチマークデータによる評価で，マハラノビス距離の導入による汎化能力の向上はデータ依存性が大きいことが分かり，汎化能力を向上する他の方式の検討を進めた。

② SVM では，最小のマージンを最大化することにより，汎化能力を向上することに成功している。しかしながら，AdaBoost などの複数識別器システムでは，最小マージンの最大化よりも，教師データの分布を制御することが重要であることが知られるようになっている。この考えを SVM の分野に導入したものとして，マージンの平均を最大化し，マージンの分散を最小化する LDM（Large margin Distribution Machine）方式が注目されている。このため，マハラノビス距離の代わりに，データの分布に基づいて識別器を構成する方式の検討を進めた。LDM では SVM より汎化能力が高いが，パラメータ数が SVM より多く，モデル選択に時間がかかる問題がある。このため，パラメータ数が少ないモデルの開発を行った。またマージンの平均を最大化する MAMC (Maximal Average Margin Classifier) 法が提案されているが，この方式では SVM より汎化能力が低いために，汎化能力を向上する方式を提案した。今後は，提案方式の評価とともに，高速な特徴選択方式を開発することが課題である。

③ データを複数のクラスに分類するマルチラベル問題に対して，汎化能力を向上するファジィマルチラベル SVM を開発した。１対他方式の SVM をマルチラベル問題に適用すると，定義されていないマルチラベルに分類されたり，どのクラスにも分類されないことが起こりうる。この問題を解決するために，各マルチラベルのクラスに対して，ファジィ領域を定義し，ファジィ領域に対する近さを示すメンバーシップ関数を定義したファジィマルチラベル SVM を開発した。未知のデータに対する識別においては，最もメンバシーシップ値が大きいマルチラベルに属すると判定する。入力数が 72 個から 47,236 個の 12 個のマルチラベルのベンチマークデータで評価して，従来の１対他方式よりも汎化能力が向上し，またこれまでに提案されている方法と遜色がないことを確認した。今後はこの方式に対する特徴選択方式を検討することが課題である。

５．主な発表論文等

〔雑誌論文〕（計　７件）

① Shigeo Abe, Resolving Unclassifiable Regions in Multilabel Classification by Fuzzy Support Vector Machines, Proc. the LWA 2015 Workshops: KDML, FGWM, IR and FGDB, 査読有, 2015, p. 158
http://ceur-ws.org/Vol-1458/E22_CRC2_Abe.pdf

② Shigeo Abe, Fuzzy Support Vector Machines for Multilabel Classification, Pattern Recognition, 査読有 Vol. 48, No. 6, 2015, pp. 2110－2117
http://www.lib.kobe-u.ac.jp/repository/90003293.pdf

③ Shigeo Abe, Optimizing Working Sets for Training Support Vector Regressors by Newton's Method, IJCNN 2015, 査読有, 2015, pp. 93–100
http://www.lib.kobe-u.ac.jp/repository/90003295.pdf

④ Shigeo Abe, Fusing Sequential Minimal Optimization and Newton's Method for Support Vector Training, International Journal of Machine Learning and Cybernetics (10.1007/s13042-014-0265-x), 査読有, 2014, pp.1－20
http://www.lib.kobe-u.ac.jp/repository/90003294.pdf

⑤ Shigeo Abe, Incremental Feature Selection by Block Addition and Block Deletion Using Least Squares SVRs, Proc. ANNPR 2014, LNAI 8774, 査読有, 2014, pp. 35–46

⑥ Shigeo Abe, Incremental Input Variable Selection by Block Addition and Block Deletion, Proc. ICANN 2014, LNCS 8681, 査読有, 2014, pp. 547–554

⑦ Shigeo Abe, Feature Selection by Iterative Block Addition and Block Deletion, Proc. IEEE SMC Conference, 査読有, 2013, pp. 2677‑2682

〔その他〕
ホームページ等
http://www2.kobe-u.ac.jp/~abe/
https://www.researchgate.net/profile/Shigeo_Abe
https://scholar.google.com/citations?user=RpDCOwQAAAAJ&hl=en
https://www.scopus.com/authid/detail.uri?authorId=7403335651

６．研究組織

(1)研究代表者

阿部　重夫（ABE, Shigeo）
神戸大学・工学研究科・名誉教授
研究者番号：50294195