

シミュレーションデータを用いた識別器の性能の検討

堀江 俊介[†] 松原 友子^{††} 笠井 聡^{†††} 内山 良一^{††††} 原 武史[†] 藤田 広志[†]

[†] 岐阜大学大学院医学系研究科知能イメージ情報分野 〒501-1194 岐阜県岐阜市柳戸 1-1

^{††} 名古屋文理大学情報文化学部 〒492-8520 愛知県稲沢市稲沢町前田365

^{†††} Konica Minolta Medical Imaging USA, INC. 411 Newark Pompton Turnpike Wayne, NJ 07470

^{††††} 岐阜大学大学院医学系研究科医療情報学分野 〒501-1194 岐阜県岐阜市柳戸 1-1

E-mail: [†] {horie,hara,fujita}@fjt.info.gifu-u.ac.jp ^{††} tomoko@nagoya-bunri.ac.jp

あらまし コンピュータ支援診断 (CAD) 開発における偽陽性削除や良悪性鑑別などの処理では、様々な識別器が用いられている。しかし、高い識別性能が得られる識別器は一意には決まらない。そのため、用いる特徴量に基づいて、正確な分類を可能にする識別器を選択することは、重要な役割を担っていると考えられる。そこで本研究では、正規分布に基づく人工的なデータ群を作成し、識別器の識別性能を比較するシミュレーション実験を行った。識別器には、線形判別 (LDA)、二次判別 (QDA)、ニューラルネットワーク (ANN)、サポートベクターマシン (SVM)、AdaBoost の5つを比較対象として用いた。その結果、データ群の分布、次元数、データ数が識別性能に与える影響が明らかとなった。

キーワード 識別器, シミュレーションデータ, AUC, コンピュータ支援診断(CAD)

Classifier Performance Study Using Simulation Data

Shunsuke HORIE[†] Tomoko MATSUBARA^{††} Satoshi KASAI^{†††} Yoshikazu UCHIYAMA^{††††}
Takeshi HARA[†] Hiroshi FUJITA[†]

[†] Graduate School of Medicine, Gifu University 1-1 Yanagido, Gifu-shi, Gifu 501-1194, Japan

^{††} School of Information Culture, Nagoya Bunri University 365 Maeda Inazawa-cho, Inazawa-shi, Aichi 492-8520, Japan

^{†††} Konica Minolta Medical Imaging USA, INC. 411 Newark Pompton Turnpike Wayne, NJ 07470

^{††††} Graduate School of Medicine, Gifu University 1-1 Yanagido, Gifu-shi, Gifu 501-1194, Japan

E-mail: [†] {horie,hara,fujita}@fjt.info.gifu-u.ac.jp ^{††} tomoko@nagoya-bunri.ac.jp

Abstract Various classifiers are used on the elimination of false positives or diagnosis of benign and malignant lesion. However, no classifier shows consistently superior performance regardless of the nature of the data. Therefore, choosing the type of classifier should be done experimentally. In this paper, we generated simulation data based on normal distributions; and then conducted a computer simulation study to compare classification performance of five classifiers, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Neural Network, Support Vector Machine, and AdaBoost. As a result, Our results clearly show the effects of class distribution, number of features, and sample size on classifier performance.

Keywords Classifier, Simulation Data, AUC, Computer-aided diagnosis (CAD)

1. まえがき

近年、さまざまなモダリティの医用画像が普及し、病変の早期発見に期待が寄せられている。しかし、同時に読影枚数の増加から、医師にかかる負担が増加し、読影ミスが生じる危険性も考えられる。そこで、現在、このような医師の負担

を軽減させるため、計算機を用いた画像解析を行うことにより、病変と疑われる部位を示すコンピュータ支援診断 (Computer-aided diagnosis: CAD) システムの開発が注目されている。

CAD システムでは、画像解析により多くの特徴量を算出し、

識別器を用いて、病変部の検出、良悪性鑑別などを行う。CADシステムの開発において、患者の情報に基き、正常であるか、異常であるかを正しく区別できる識別器の設計は、重要な役割を担っていると考えられる。

識別器には、線形判別や二次判別を代表として様々なものがある[1]。しかし、常に高い識別性能が得られる識別器は一意には決まてはいない。よって、有効な特徴量の選択基準、特徴量の数やその分布、また、トレーニングのデータ数が識別器に与える影響が明らかになれば、CADにおける最適な識別器の選択が可能となる。

Chanらは、人工データを用いて、線形判別分析(Linear Discriminant Analysis : LDA)、二次判別(Quadratic Discriminant Analysis : QDA)、ニューラルネットワーク(Neural Network : ANN)におけるサンプルサイズが識別性能に与える影響を報告している[2]。

本研究では、これら3種類の識別器に加えて、サポートベクターマシン(Support Vector Machine : SVM)[3]とAdaBoostの5種類の識別器の識別性能をシミュレーション実験により比較することを目的とする。

2. 方法

我々が行ったシミュレーション実験の流れを図1に示す。その詳細を以下に示す。

2.1 データ群の作成

本研究では、正規分布に基づく2つのデータ群(Class A, Class B)を作成した。Class Aの平均ベクトルの成分は全て0、共分散行列は単位行列とする。Class Bは Bhattacharyya

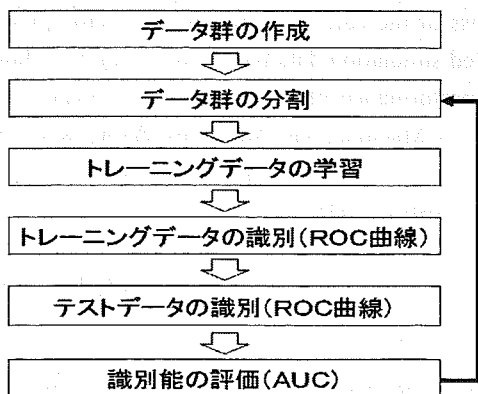


図1 シミュレーション実験の流れ

distance が 3/8 となるように、異なる平均ベクトル・異なる共分散行列(Case 1)、等しい平均ベクトル・異なる共分散行列(Case 2)、異なる平均ベクトル・等しい共分散行列(Case 3)の3つの分布に基づくデータ群とする。Bhattacharyya distance は以下の式により求められる。

$$B = \frac{1}{8} \Delta + \frac{1}{2} \ln \frac{\det[(\Sigma_1 + \Sigma_2)/2]}{\sqrt{\det \Sigma_1} \sqrt{\det \Sigma_2}} \quad (1)$$

$$\Delta = (\mu_2 - \mu_1)^T (\frac{\Sigma_1 + \Sigma_2}{2})^{-1} (\mu_2 - \mu_1) \quad (2)$$

ここで、 μ_1, μ_2 は Class A, Class B の平均ベクトルであり、 Σ_1, Σ_2 は Class A, Class B の共分散行列である。

データは Class A, Class B, 各々で 1000 個ずつとする。また、特徴量の数(次元数)は、3, 6, 9, 12, 15 とする。

例として Class A のデータ 300 個、Class B のデータ 300 個の 2 次元データの分布を図 2 に示す。

2.2 データ群の分割

2.1 で述べた Class A と Class B を、トレーニングデータとテストデータに分割する。トレーニングデータは、ランダムに Class A, Class B から各々 N 個ずつ抽出した合計 2N 個とする。また、テストデータは、Class A から 1000-N 個、Class B から 1000-N 個の合計 2000-2N 個とする。また、N は 20, 40, 60, 80, 100, 200, 400, 600, 800 の 9 パターンとする。

2.3 識別能の評価

データの分布 3 種類、特徴量 5 パターン、トレーニングデータとテストデータの組み合わせ 9 パターンの計 1 3 5 種類に対して、各々、トレーニングをし、それを用いてテストを行う。トレーニングとテストの結果を ROC 曲線下の面積 AUC(area under the curve)を用いて検討する。今回、識別器には、LDA, QDA, ANN, SVM, および AdaBoost を用いる。なお、ANN の中間層のノードの数は次元数の数と等しい。SVM の

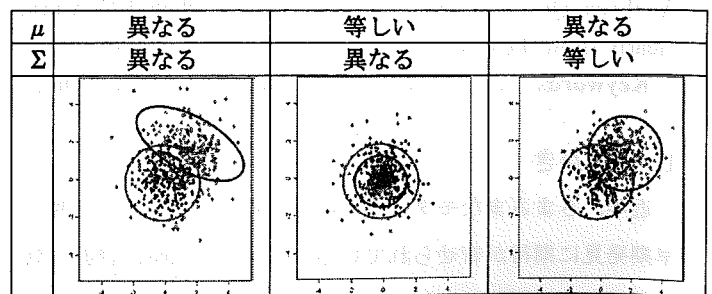


図2 2次元での正規分布に基づく3種類のデータ分布

カーネルは radial, gamma は 1/3, cost は 1 となっている。そして, AdaBoost の boosting タイプは discrete, 縮小パラメータは 1 となっている。

また, 実験結果の偏りを減少させるため, データの全パターンに対して, 10 回の実験を行った AUC の平均値を結果に示す。

3. 結果と考察

3.1 Case 1: 平均ベクトルが異なり, 共分散行列が異なる分布

平均ベクトルが異なり, 共分散行列が異なる分布での 3 次元データの AUC のグラフを図 3 (a) に示す。また, 15 次元データの AUC のグラフを図 3 (b) に示す。

全体的に, トレーニングのデータ数が多いとき, トレーニングの AUC 値は高いが, テストの AUC 値が低くなる。また, トレーニングのデータ数の増加に伴い, トレーニングの AUC 値が低くなり, テストの AUC 値が高くなる。よって, トレーニングのデータ数を多くするほど, トレーニングとテストの AUC 値の差は徐々に小さくなっているのが分かる。

また, トレーニングに 1600 個のデータを用いたときの AUC 値に注目すると, QDA が 0.888 と一番高い値であった。そして, 識別器ごとのテストの AUC 値の平均 (以下: AUC 値の平均) においても QDA が 0.879 と 5 つの中で最も高い値となり, トレーニングのデータ数の大小に関わらず, 高い AUC 値が期待できると考えられる。しかし, トレーニングに 1600 個のデータを用いたときの AUC 値は LDA が 0.872, ANN が 0.848, SVM が 0.865, AdaBoost が 0.860 であり, 多くのデータを用いてトレーニングされた状態においては, 5 つの識別器に大きな差はなかった。

また, 特徴量を増やし, 次元数を増加させると, トレーニングの AUC 値とテストの AUC 値との差は大きくなる傾向を示した。その差はトレーニングのデータ数が少ないときほど, 顕著であった。

15 次元での結果を例として示す。トレーニングに 1600 のデータを用いたとき, 3 次元の場合と同様に QDA は 0.866 と一番高い AUC 値であった。しかし, 40 のデータのみをトレーニングさせた場合, AUC 値は 0.582 となり, 5 つの識別器の中で最も低い AUC 値であった。QDA の AUC 値の平均はト

レーニングのデータ数が少ないときの値の影響が大きく, 0.767 と低い値となった。他の識別器もトレーニングのデータ数を多くすることによって, 高い AUC 値を示すことができたが, トレーニングが少ないときは AUC の値は 0.75 を超えなかった。その中で SVM はトレーニングのデータ数が最小の 40 のときでも, 0.792 を示し, AUC 値の平均も 0.858 と最も高い値を示した。このことから, SVM はトレーニングのデータ数への依存が少なく, トレーニングのデータ数の大小に関わらず, 安定した結果が得られると考えられる。

以上より, この分布においては多くのトレーニングデータを用いることが可能であるならば, QDA が最良の識別器であり, トレーニングのデータ数が少なければ, SVM が最良の識別器である。

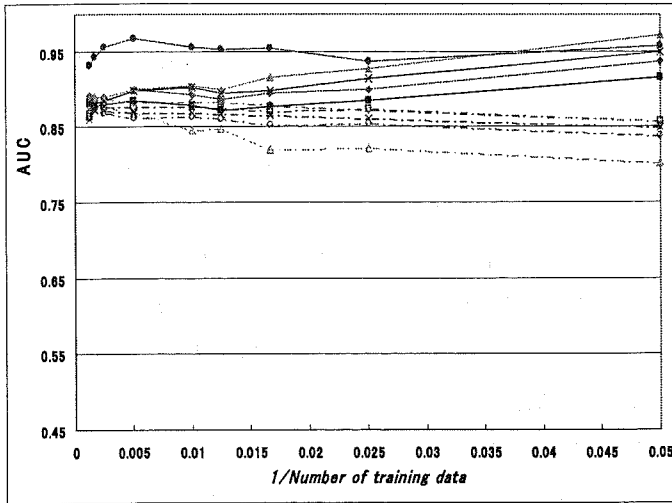
3.2 Case 2: 平均ベクトルが等しく, 共分散行列が異なる分布

平均ベクトルが等しく, 共分散行列が異なる分布での 3 次元データの AUC のグラフを図 3 (c) に示す。また, 15 次元データの AUC のグラフを図 3 (d) に示す。

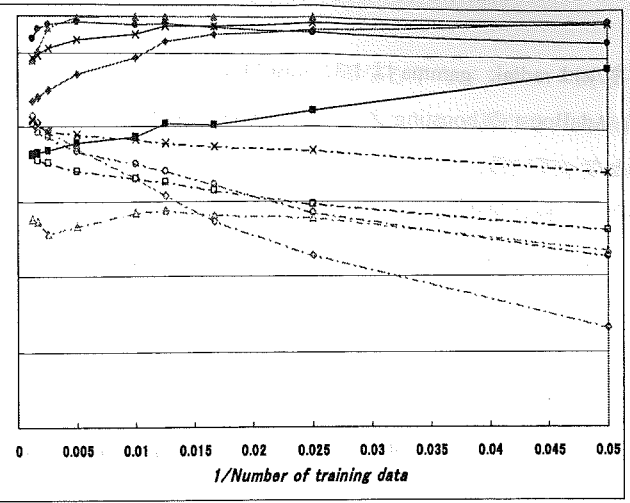
この分布では, Case 1 の分布の結果とは異なり, 全体的に AUC 値が低くなっている。この分布は Class A の分布範囲内に Class B が完全に含まれているため非線形な分類が求められる。5 つの識別器の中でも LDA の AUC 値の平均は 0.487 であり, 実用のレベルではないといえる。

一方, QDA と SVM の AUC 値の平均は, 0.849 と 0.847 となり, この分布の特徴量に対しては, これらは有効な識別器であるといえる。

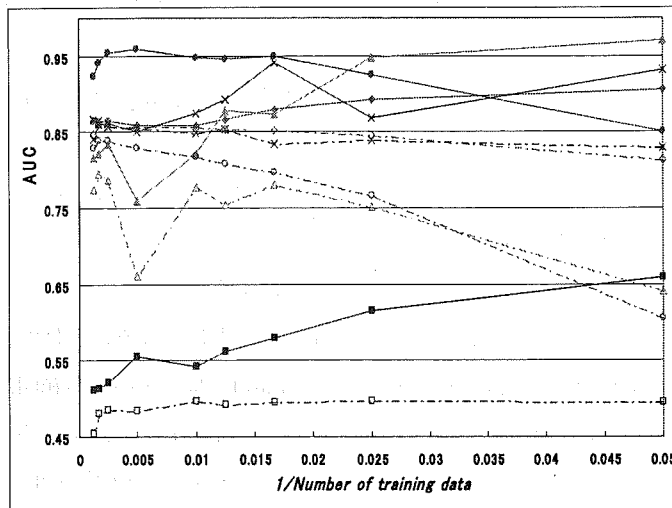
また, 次元数が増加したときの例として, 15 次元空間での結果を示す。トレーニングのデータ数の違いによる AUC 値の変化は, 3 次元の結果とは大きく異なった。トレーニングのデータ数を 1600 にした結果は, QDA が 0.857, SVM が 0.848, AdaBoost が 0.842 とこれらの識別器は高い値を示しているが, トレーニングのデータ数が最も少ない 40 での AUC 値は, QDA が 0.570, AdaBoost が 0.532 と低い値を示している。一方で, その状態での SVM は 0.787 と他の識別器と比べ高い値を示しており, AUC 値の平均も唯一 0.8 以上を示している。よって, 次元数, トレーニングのデータ数の影響を考えると, この分布では SVM が最も汎用性があり, 有効な識別器と考えられる。



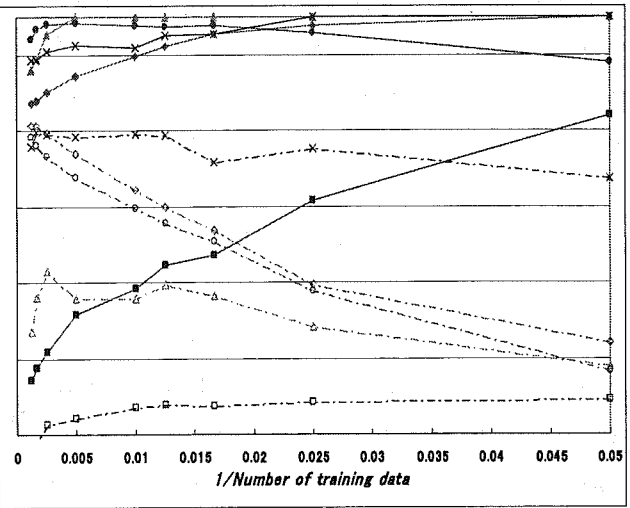
(a)



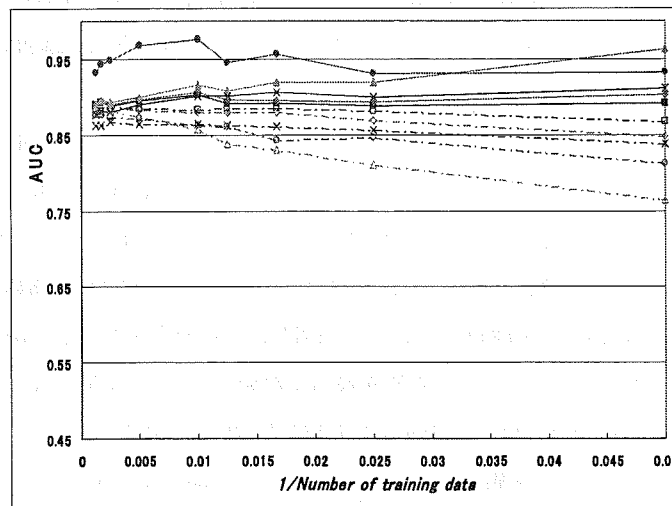
(b)



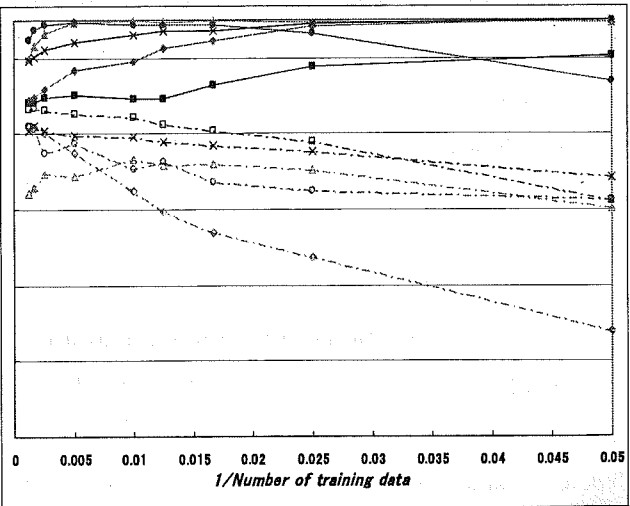
(c)



(d)

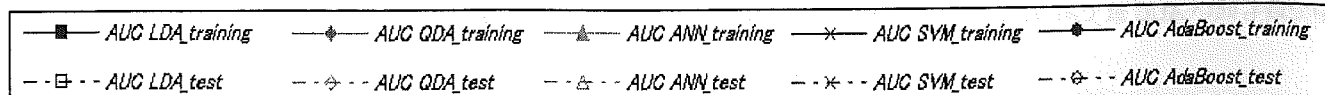


(e)



(f)

図3 識別器ごとのトレーニングとテストのAUC値
Case 1の3次元(a)と15次元(b), Case 2の3次元(c)と15次元(d), Case 3の3次元(e)と15次元(f)



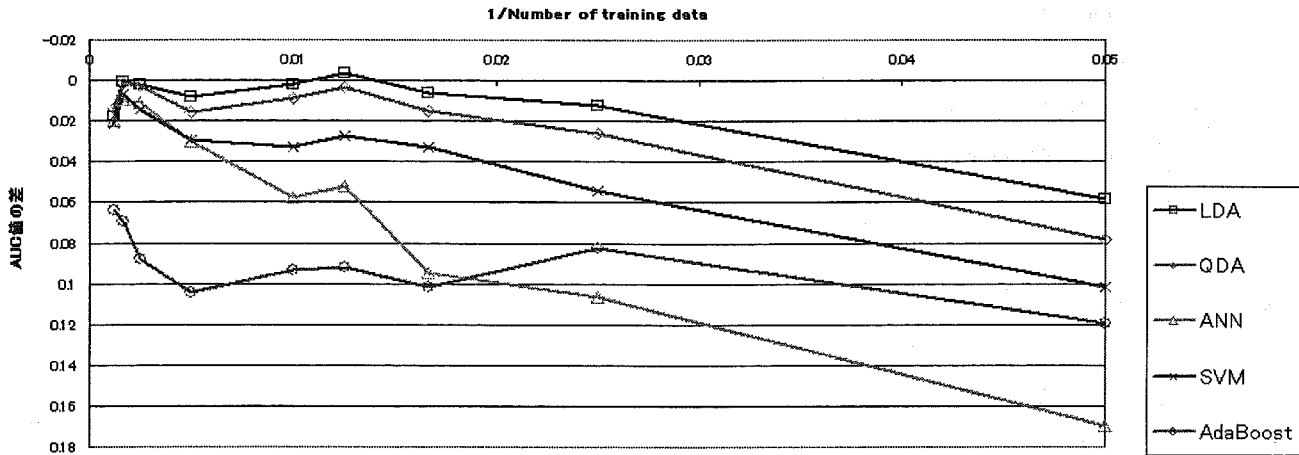


図4 5つの識別器に対するトレーニングとテストのAUC値の差

3.3 Case 3: 平均ベクトルが異なり, 共分散行列が等しい分布

最後に, 平均ベクトルが異なり, 共分散行列が等しい分布の結果を示す. 3次元データのAUCのグラフを図3(e)に示す. また, 15次元データのAUCのグラフを図3(f)に示す. この分布では, トレーニングのデータ数が1600でのAUC値は, すべての識別器で0.8以上という高いAUC値を示している.

3次元空間では, どの識別器も高いAUC値を示しているが, その中でも, LDAとQDAはAUC値の平均が0.882と0.877と特に高い値を示している. また, トレーニングのデータ数が最も少ない40のときでも, LDAとQDAは0.868と0.848とトレーニングのデータ数が少ない状態でも良好な数値を示している.

また, 次元数が増えた例として, 15次元の結果を見てみると, 3次元で有効な識別能を示したQDAは, トレーニングのデータ数が多い場合は高いAUC値を示すが, トレーニングのデータ数が40になると, 0.590と一番小さい値を示した. しかし, LDAは, トレーニングのデータ数が40の状態でも0.762であり, AUC値の平均も0.857と一番高い数値を示しており, この特徴量ではLDAが最良と考えられる. なお, SVMとAdaBoostも15次元空間でのAUC値の平均は0.8以上を示しており, 有効であると考えられる.

3.4 3つの分布の結果の比較

今回, 用いた3つのデータの分布に対して, QDAは次元数が少ない場合は, トレーニングのデータ数の大小に関わらず,

高い性能を示した. しかし, 次元数の増加に伴って, 多くのトレーニングデータ数を必要とし, トレーニングデータ数が少ないときに良好な性能をあげることはできなかった. また, AdaBoostはトレーニングでの性能が一番良い精度であったが, テストでは, トレーニングほどの性能をあげることはできなかった. このことから, AdaBoostは他の識別器よりも, 過学習に注意する必要があるといえる. また, トレーニングのデータに関してはQDAほどではないが, 同様に, 次元数の増加に伴って, 多くのトレーニングのデータ数を必要とする傾向が見られた. SVMは5つの識別器の中では, トレーニングのデータ数に最も影響を受けず, 次元数の増加に対しても安定しており, 良好な結果を得ることが確認された. LDAに関しては, 明らかに2クラスを区別しやすい分布に対しては良好に作用するが, 非線形な分布では, 実用に耐えうる識別性能は得られなかった.

また, トレーニングのAUC値を基準としたトレーニングとテストのAUC値の差を図4に示す. 識別器ごとのデータ数によるトレーニング結果とテストの結果の差が明らかとなれば, トレーニングの結果からテストの結果を予測することが可能となる.

4. まとめと今後の展望

本研究では, データの分布や次元数, トレーニングのデータ数が識別器に与える影響を明らかにするために, 人工的にデータを作成し, 5種類の識別器の性能をAUCにより評価した.

今回、用いた3つの分布のデータでは、QDAとSVMが有効であった。QDAに関しては、次元数が小さいときは、比較的トレーニングのデータ数に影響を受けることなく高い性能を示した。しかし、次元数の増加に従い、トレーニングのデータ数を多くする必要があった。そのため、QDAでは、少数の有効な特徴量を用いると、良い結果を得やすいと考えられる。それに対して、SVMは、次元数が増加しても、トレーニングのデータ数の影響が少なかった。よって、次元数やデータ数の大小に関わらず、安定した高い性能が得られると考えられる。

今回は単峰分布に基づくデータのみでのシミュレーション実験を行ったが、これらを多峰分布など異なる条件での実験に発展させ、さらには臨床データへの適用を検討する予定である。また、分類の性能の結果は、どのような特徴量が有益かという特徴量の選択の基準となり得ることが期待できる。

文 献

- [1] 永田靖, 棟近雅彦, “多変量解析法入門”, pp. 99-131, サイエンス社, 2001.
- [2] Heang-Ping Chan, Berkman Sahiner, Robert F. Wagner, and Nicholas Petrick, “Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers”, *Med.Phys.*, 26, pp.2654-2668, 1999.
- [3] 小野田崇, “サポートベクターマシン (知の科学)”, オーム社, 2007.