

報告

GNU R を用いた判別分析と機械学習のための 統計処理プログラムの作成

王丸 愛子^{1) 2)}・小林 龍徳¹⁾・金 珍澤¹⁾

- 1) 純真学園大学 保健医療学部 放射線技術科学科
- 2) 大分県立看護科学大学大学院 健康科学専攻 放射線健康科学領域

Programming of a statistical analysis function for the discriminant analysis and machine learning by using GNU R

Authors: Aiko Ohmaru^{1) 2)}, Tatsunori Kobayashi¹⁾, Jintaek Kim¹⁾

- 1) Department of radiological science, Faculty of health science, JUNSHIN GAKUEN UNIVERSITY
- 2) Environmental Health Science Division, Department of Health Sciences, Graduate School of Oita University of Nursing and Health Sciences

要旨：近年、ビッグデータ解析や人工知能の研究において判別分析や機械学習などの統計学的手法が利用されている。本研究では、判別分析と機械学習を簡便に実行することを目的とし、GNU R を用いた統計処理プログラムを開発した。本プログラムは、GNU R に含まれる識別器（線形判別分析と二次判別分析、Random Forest）を簡単なコマンド入力ですぐに実行できる。本プログラムの動作確認のため、オンラインで公開されている乳がんデータベースの30の画像特徴量に基づいて、がんの良悪性の分類を行った。その結果、各識別器の正判別率は90%以上であり、分類結果は数秒で得られた。結論として、本プログラムは膨大な統計処理が求められる様々な領域において有用であると考えられ、一例として病理診断の補助的なツールとして応用できる可能性を示した。

キーワード：判別分析, 機械学習, Random Forest, R 言語プログラミング, 乳がん

Abstract: In recent years, statistical methods such as discriminant analysis and machine learning are used in the research of Big Data analysis and AI. In this study, we developed a statistical processing program using GNU R, which aims to perform discriminant analysis and machine learning efficiently. The program enables us to run classifiers (linear discriminant analysis, quadratic discriminant analysis, and Random Forest) contained in GNU R quickly with a simple command input. To confirm the operation of the program, we conducted a classification test of benign/malignant tumors based on the 30 feature values of images in the breast cancer database which is publicly available online. The test result was obtained within a few seconds, and the result showed that the percentage of correct discrimination at each classifier was more than 90%. In conclusion, the program was considered to be useful in various areas which require a significant amount of statistical processing, and as an example, we were able to demonstrate that this program can be applied as an additional useful tool to pathological diagnosis.

Keyword: discriminate analysis, machine learning, Random Forest, R programming language, breast cancer

1. はじめに：

近年、ビッグデータ解析や人工知能の研究が注目されている。これらの研究では、大規模な画像やテキストデータに対して、判別分析や機械学習を用いて、入力データの特徴に応じて、クラス分類や対象物を抽出する。判別分析や機械学習は、

事前に学習データで作成した判別式や判別基準に基づいて、未知の入力データを、統計的に判別し、分類するものである。この判別式や判別基準の作成には、多数の学習データとテスト用データが必要である。これらのデータは、入力データの特徴量と、クラスの情報から構成される。

判別分析法は、学習データとして与えられた、特徴量とクラスの情報から、判別式を作成する。そして、未知の入力データが与えられたときに、その特徴量を判別式に代入し、判別得点を計算し、その判別得点に基づいてクラス分類を行う¹⁾。機械学習は、人工知能に利用されている技術のひとつである。これは、多数の学習データから特徴抽出や人間の判断に近い処理をコンピュータにさせるものである。機械学習による判別法として、Random Forest (RF) がある^{2,3)}。RFは、2クラスに分類するための決定木を多数用いて判別する。RFの特徴として、RFの処理過程で利用した特徴量の変数重要度が計算される。この変数重要度に基づいて、判別に利用する特徴量について検討できる。

これらの判別分析や機械学習を行うためにはSPSS⁴⁾、MATLAB⁵⁾、GNU R⁶⁾などの、高機能な統計処理専用のソフトが利用される。本研究では、GNU R⁶⁾ (R) を利用した。Rは、R言語で記述されたオープンソースであり、マルチプラットフォームに対応しており、無料で利用できる。Rは、基本的な統計量の算出以外に、グラフの描画、テーブルの作成ができる。Rの使用方法については、多くの書籍が出されている⁷⁻¹⁰⁾。また、書籍だけではなく、インターネットにおいても、RユーザーがRの使い方やTipsについて多くの情報を載せており、最新の使い方などの情報も得やすい。Rは、基本パッケージ以外の高度な統計処理を実現するために、追加パッケージが公式ホームページで無料配布されている。Rの基本機能に、このパッケージを追加し、より高度な統計処理が可能となる。Rの欠点は、統計処理がコマンドの入力である点である。Rを使いこなすためには、専用のコマンドを覚える必要がある。このコマンドの入力や操作を簡便にするために、R-コマンド¹¹⁾ や EasyR¹²⁾ などの入力支援ソフトウェアが開発されている。また、各コマンドにはオプションがあり、Rのコマンド実行時に引数として与えて、統計解析の目的に合わせた処理が可能である。

統計処理では、複数のデータに対して同じ処理の繰り返し、複数の統計処理を組み合わせる場合が多い。この統計処理毎に、コマンドと引数の繰

り返し入力は、入力間違いや処理手順の間違いの原因となる可能性がある。そこで、本研究では、判別分析とRFを実行するためのコマンドをまとめて、これらを簡便に実行するための統計処理プログラムをR言語で作成した。そして、オンラインで公開されている機械学習用のデータベースに適用し、統計処理プログラムの有用性について述べる。

2. 方法：

2.1. 使用ソフトとパッケージ

判別分析とRFは、R (version 3.4.2) で実行した。使用したRのパッケージは、MASS¹³⁾ と randomForest¹⁴⁾ である。MASSは、多変量解析のためのプログラムがまとめられたパッケージである。random Forestは、RFを実行するために必要なパッケージである。

2.2. データベース

本研究で使用したデータベースは、カリフォルニア大学アーバイン校が公開している機械学習用のデータベース (UCI Machine Learning Repository) の Wisconsin Diagnostic Breast Cancer (WDBC)¹⁵⁻¹⁷⁾ である。WDBCは、乳がんを含む乳腺疾患のある569人の患者の、ID、良性 (Benignancy : B) ・悪性 (Malignancy : M) のクラス、穿刺吸引細胞診 (fine needle aspiration cytology : FNA) の標本画像の標本画像から検出された30特徴量からなる計32個の特徴量で構成される。FNAの標本画像から得た特徴量を Table. 1に示す。本データベースは、Table. 1に示した10種類の特徴量について、細胞核ごとに、平均値、標準誤差、最も悪い値が記録されている。この10種類の特徴量は、乳管癌の組織学的異型度の評価における核異型度の判定基準となる核の大小不同・形態不整を表すものである¹⁸⁾。本研究では、良性・悪性のクラス分類には、FNAの標本画像から得られた30個の特徴量を用いて行った。

2.3. データベースの整形

WDBCのデータを学習用とテスト用のデータに分けるために、まず、データベースをBクラス、Mクラスの順番でソートした。次に、学習用デー

タとテスト用データのデータセットが B クラスと M クラスをほぼ同数にするために、B クラスと M クラスのデータ数を数えた。そして、各クラスの半分を学習用データ、残り半分をテスト用データとした。B のデータ数が奇数個であったため、学習用データを1例多くした。本研究で用いた学習用データは (n=285, (B=179, M=106)), テスト用データは (n=284, (B=178, M=106)) である。

Table. 1 WDBC に記載されている10種類の特徴量

1.	radius (mean of distances from center to points on the perimeter)
2.	texture (standard deviation of gray-scale values)
3.	perimeter
4.	area
5.	smoothness (local variation in radius lengths)
6.	compactness (perimeter ² / area - 1.0)
7.	concavity (severity of concave portions of the contour)
8.	concave points (number of concave portions of the contour)
9.	symmetry
10.	fractal dimension (“coastline approximation” -1)

2.4. 判別器の作成¹⁻³⁾

2.4.1 入力データと評価

本研究では、Table1に示した10種類の特徴量を持つ学習用データから判別式および RF を用いた判別器をそれぞれ作成した。そして、学習用データとテスト用データをそれぞれの判別器に入力し、その正判別率を評価した。

2.4.2. 判別分析

本研究では、判別分析として、線形判別分析 (linear discriminant analysis : LDA) 法と2次曲線判別分析 (quadratic discriminant analysis : QDA) 法を用いた。一般に、線形判別関数は、式 (1) で表される。

$$f(x) = w^T x + w_0 \quad (1)$$

ここで、 w は係数ベクトル、 x は入力ベクトル、 w_0 はバイアス項である。判別分析では、この判別関数に基づいて、クラス分類のための判別式を作成する。判別式は、判別得点を計算し、クラス分

類を行う。 x を2クラスに分類したとき、各クラスの平均値は、式 (2) で表される。

$$\mu_k = \frac{1}{N_k} \sum_{i \in C_k} x_i \quad (k=1, 2) \quad (2)$$

この平均値を式 (3) のように写像する。

$$m_k = w^T \mu_k \quad (3)$$

ここで、写像後の m_1 と m_2 のデータ分布の各平均値を求めたとき、2クラスの平均値の差 ($m_1 - m_2$) が大きいほど、クラスの分離は良くなる。また、この平均値の差の2乗がクラス間変動 (S_1) となる。ここで、各クラスの分散が小さいほうが、データの分布が小さいためクラス間の重なりが小さくなる。ここでクラス内変動は、式 (4) で定義され、全クラス変動は、 $S_1^2 + S_2^2$ で表される。

$$S_k^2 = \sum_{i \in C_k} (f(x) - m_k)^2 \quad (4)$$

ここで、クラス間変動とクラス内変動の比であるフィッシャー基準 $J(w)$ を式 (5) から求め、このクラス内変動が最大となる w を求める。

$$J(w) = \frac{(m_1 - m_2)^2}{S_1^2 + S_2^2} \quad (5)$$

R では、MASS パッケージに含まれている `lda` 関数と `qda` 関数を用いて、LDA と QDA が利用できる。これらの統計処理は、Fig. 1 に示すコマンド例で実行できる。そして、これらの関数で得られた判別結果は、変数 `Zlda` と `Zqda` にそれぞれ代入した。その判別式の判別能力は、`predict` 関数を利用して、テスト用データを判別した。テスト用データの判別結果は、`Ylda` と `Yqda` にそれぞれに代入した。これらの変数の値や結果は、`print` 関数により計算結果を表示できる。

```
library(MASS)
#線形判別分析：
Zlda = lda(学習用データ, クラス)
Ylda = predict(Zlda, テスト用データ)
print(Zlda)
print(Ylda)

#2次判別分析：
Zqda = qda(学習用データ, クラス)
Yqda = predict(Zqda, テスト用データ)
print(Zqda)
print(Yqda)
```

Fig. 1 判別分析を実行するコマンド例

2.4.3. Random Forest (RF)

RF は, Fig. 2 に示すように, 学習データから重複を許した複数回のランダムサンプリングを行うブートストラップサンプリングにより, 多数の決定木を作成する. そして, その決定木の多数決により, 判別器を作成する. RF では, 判別器の作成において, 多数の決定木を利用するため Random Forest (乱数の森) といわれる.

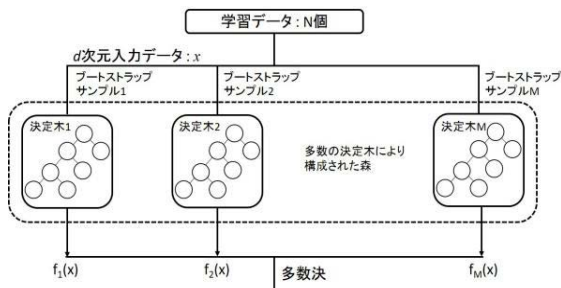


Fig. 2 RF の判別器の作成法

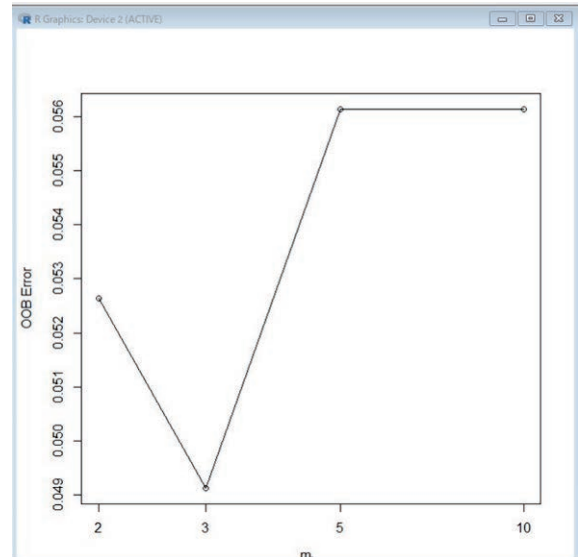
2.5. 簡便化のための R プログラムの作成

ここまでの処理を簡便に実行するために, 処理内容とコマンドの引数をまとめた統計処理プログラムを R 言語で作成した. この統計処理プログラムの名前は, Multiple classification analysis and Random forest Program の頭文字を取り mrp.r とした. mrp.r は, 処理対象のデータが入っているフォルダ内に mrp.r を置く. まず, Fig. 3 (a) に示すように source 関数で本プログラムを R に読み込む. 次に, mrp (“入力ファイル名”)としてコマンドを打てば実行できる. 本プログラムの実行後, random Forest パッケージ内のパラメータ調整のための tuneRF 関数から Fig. 3 (b) に示すグラフが出力される. このグラフは, 縦軸がパラメータ調整における特徴量数による誤差率, 横軸が特徴量数を示す. 表示される誤差率が最小となる特徴量数を Fig. 3 (c) に示すコンソール画面に入力する. このグラフにおいて, 特徴量数が3のときに誤差率が最小となっているため, コンソール画面にて3を入力している. この mrp.r は, 出力結果をテキストファイルとして保存する. また, 判別分析と RF の処理過程で, 分析に必要なグラフ等を PNG 形式の画像として保存する.

(a) mrp.r を実行するコマンド例

```
source("mrp.r")#自作プログラムの読み込み
mrp("WDBC.csv")#自作プログラムの実行
```

(b) tuneRF 関数から出力されたグラフ



(c) コンソール画面のスクリーンショット

```
R Console
R version 3.4.2 (2017-09-28) -- "Short Summer"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R は、自由なソフトウェアであり、「完全に無保証」です。
一定の条件に従えば、自由にこれを再配布することができます。
配布条件の詳細に関しては、'license()' または 'licence()' と入力してください

R は多くの貢献者による共同プロジェクトです。
詳しくは 'contributors()' と入力してください。
また、R や R のパッケージを出版物で引用する際の形式については
'citation()' と入力してください。

'demo()' と入力すればデモをみることもできます。
'help()' とすればオンラインヘルプが出ます。
'help.start()' で HTML プラッファによるヘルプがみられます。
'q()' と入力すれば R を終了します。

[以前にセーブされたワークスペースを復元します]

> source("mrp.r")
> mrp("WDBC.csv")
randomForest 4.6-12
Type rfNews() to see new features/changes/bug fixes.
3
> |
```

Fig. 3 統計処理プログラム mrp.r の実行コマンド例

3. 結果 :

3.1. 判別分析と RF による判別結果

Table. 2-4 に学習データから作成した LDA, QDA, RF に, 学習データとテスト用のデータを入力としたときの, それぞれの判別結果を示す. また, 3つの判別器の正判別率を Fig. 4 に示す. 正判別率 D は, 式 (6) より計算した.

$$D = \frac{n_c}{N} \quad (6)$$

ここで, n_c は, クラスに分類されたデータ数, N は学習用データまたはテスト用データのデータ数である. Fig. 4 における各判別式における左側が学習用データに対する判別結果, 右側がテスト用データに対する判別結果である.

LDA は, B クラスの判別では, 学習データとテストデータともに, 98.8 % であり, 本研究で用いた判別器の中で最も高い正判別率であった. M クラスの判別では, 学習用データとテスト用のデータにおいて 90 % 以上の正判別率であったが, 学習用データにおける誤判別率が 9.4 % であった.

QDA は, B クラスの判別では, 学習データでは 98.8 % であり, LDA と同様に, 高い判別率であった. しかし, テスト用データに関しては, 93.3 % まで減少した. M クラスの判別では, 学習用データが 95.3 %, テスト用データが 98.1 % であった.

RF は, B クラスの判別では, 学習データとテスト用データともに約 95 % であった. M クラスの判別では, 学習用データが 92.5 % であり, テスト用データで 96.2 % まで上昇した. また, RF は, 判別器を作成する前に tuneRF 関数を利用して, 有効な特徴量の数を検討できる. RF は, 判別分析のように全ての特徴量を判別式の作成に利用しないため, 判別に有効な特徴量数で判別式を作成できる. また, RF の計算過程で作成される変数重要度 (Decrease Mean Gini) のグラフを Fig. 5 に示す. この変数重要度は, その特徴量を除いたときに, 判別精度への影響の大きさを表している.

Table. 2 LDA の判別結果

	B		M	
	学習 (179)	テスト (178)	学習 (106)	テスト (106)
B	177	176	2	2
M	10	9	96	97

Table. 3 QDA の判別結果

	B		M	
	学習 (179)	テスト (178)	学習 (106)	テスト (106)
B	177	166	2	12
M	5	2	101	104

Table. 4 RF の判別結果

	B		M	
	学習 (179)	テスト (178)	学習 (106)	テスト (106)
B	172	171	7	7
M	6	4	100	102

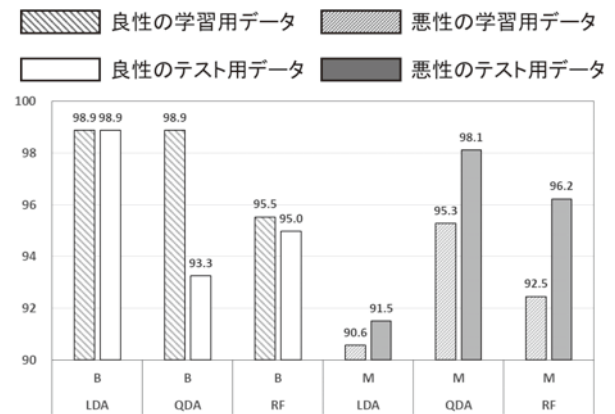


Fig. 4 判別分析と RF の正判別率

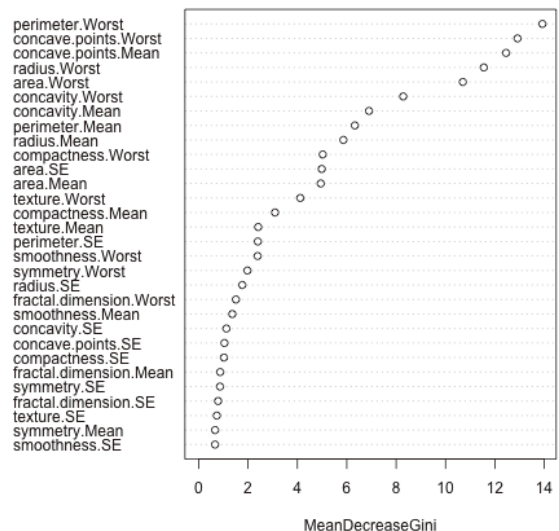


Fig. 5 RF で得られた変数重要係数の例

3.2. 統計処理プログラムの利用

mrp.rにまとめた判別分析と機械学習の実行時に、コマンドや引数を手入力した場合、1回の処理過程を終えるまでに、30分から1時間程度の処理時間を要した。これは、コマンド入力、引数の確認、オプションの入力に時間を要していたためである。mrp.rは、Fig. 3に示した2行のコマンド入力で行うことができる。このmrp.rの利用により、本研究の統計処理は、数秒で結果が得られた。

4. 考察：

本研究では、ビッグデータ解析や人工知能の研究で用いられている判別分析と機械学習を簡便に行うための統計処理プログラムをR言語で作成した。

本研究で用いたLDA、QDA、RFの正判別率は、学習用データとテスト用データにおいて90%以上の正判別率であった。本研究で得られた結果について、先行の臨床研究と比較した。山口らの報告では、本邦におけるFNAの診断能は感度96.7%、特異度84.3%と報告されている¹⁹⁾。また29の文献を比較したFNAの診断能についての検討²⁰⁾では、感度65~98%、特異度34~100%と報告により大きな差があった。本研究における判別分析とRFの正判別率は全て90%以上であることから、FNA検体画像を検出して得られた特徴量に基づいて、良・悪性の判別に有効であると示唆できた。また、FNAの診断精度は、細胞診断医の経験・能力および検体を採取する術者の技量に左右され²¹⁾、異型度の低い癌の鑑別が難しいことはFNAの問題点とされている²²⁾。本研究では検体画像から特徴量を機械的に検出したWDBCのデータを用いて判別を行っており、経験・能力に左右されずに良悪性の鑑別が可能である。

Fig. 5に示したグラフは、変数重要度であり、判別に有効な特徴量ほど、高い値をとっている。変数重要度が上位の特徴量は、Perimeter.Worst、Concave.Points.Worst、Concave.Points.Meanであった。悪性度が高いがん細胞では、核異型すなわち核の大小不同と核形不整がみられ、核クロマチンは細顆粒状で密に充満する。変数重要度が上位の特徴量は、核の大きさが大きく、核の凹部数が多い分葉状であることを示す¹⁸⁾。癌の細胞所見と変

数重要度の傾向は、一致しており、RFの変数重要度の結果の妥当性が確認できた。よって、この変数重要度から、多数の特徴量から構成されたデータベースを利用するときに、判別に有効な特徴量に絞って解析ができる。

これらの結果から、本研究で作成した統計処理プログラムは、簡便に判別分析とRFを実行できるため、有用であると考えられる。本mrp.rは、2行のコマンド入力で行う・出力結果が得られるため、利用者のコマンド操作による負担を軽減できる。そして、数秒で処理結果が得られるため、統計処理の効率化と高速化が可能となる。

判別分析法の判別精度は、判別式の作成に利用する学習用データに依存する。そのため、学習用データとテスト用データに含まれている特徴量の分布に偏りがあった場合、判別精度が低下してしまう。本研究において使用したデータベースは機械学習用に作成されており、本来FNAの検体には、検体不適正や偽陰性、偽陽性を呈し得る鑑別困難な病変が存在する²¹⁾。また、本研究では30個の特徴量を全て利用して、判別式を作成している。これらの特徴量の中には、BクラスとMクラスの判別に、有効性が低い特徴量が含まれている可能性がある。より判別精度の高い判別式を構築するためには、有効な特徴量で判別式を構築する必要がある。特徴量を削減した判別式の作成方法としてStepwise変数選択法がある²³⁾。Stepwise変数選択法は、赤池情報量基準(Akaike's Information Criterion : AIC)に基づいて変数を増加または減少させ、判別精度の高い判別式を作成できる。また、判別式の有効性の検討法のひとつとして用いられる交差検証(Cross Validation : CV)は、mrp.rが実装できていないため、今後、mrp.rにCVの導入について検討する。

判別分析やRFは、画像解析の領域だけではなく、Bioinformaticsの領域でも利用されている^{24,25)}。現在、画像解析の領域においても、画像特徴と遺伝子情報を組み合わせた、radiogenomics^{26,27)}が注目されている。遺伝子解析の結果で得られた特徴量と、画像から得られた特徴量から、病変のクラス分類や、予後予測などの利用に向けて研究が行われている。

5. まとめ：

本研究では、機械学習用のデータベースとしてオンラインで公開されている WDBC に判別分析と RF を適用し、それらの判別率を評価した。そして、今後の研究において、各判別器の利用方法に注意すべき点や特性などを述べた。本研究で作成した判別器を作成するプログラムにおいて、学習用データとテストデータに対しての正判別率は臨床研究と同等であった。加えて、RF は、特徴量の変数重要度が計算できるため、多数の特徴量の中から、分類に重要な特徴量の検討に適用できると考えられる。

参考文献：

1. 辻谷 将明, 竹澤 邦夫, 金 明哲. “マシンラーニング (R で学ぶデータサイエンス6)”, 共立出版, 東京, 2009.
2. 平井有三. “はじめてのパターン認識”, 森北出版, 東京, 2012.
3. Breiman L. Random Forests, *Machine Learning*, 45 (1), pp 5-32, 2001.
4. SPSS. <https://www.ibm.com/analytics/jp/ja/technology/spss/>
5. MATLAB. <https://jp.mathworks.com/products/matlab.html>
6. R Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
7. 青木繁信. “R による統計解析”, オーム社, 東京, 2009.
8. 金明哲. “R によるデータサイエンス データ解析の基礎から最新手法まで”, 森北出版, 東京, 2009.
9. 金 明哲, 中村 永友. “多次元データ解析法 (R で学ぶデータサイエンス 2)”, 共立出版, 東京, 2009.
10. 金森 敬文, 竹之内 高志, 村田 昇, et al. “パターン認識 (R で学ぶデータサイエンス 5)”, 共立出版, 東京, 2009.
11. R Commander. <http://www.rcommander.com/>
12. EasyR. <http://www.jichi.ac.jp/saitama-sct/SaitamaHP.files/statmed.html>
13. Ripley B, Venables B, M. Bates D et al. <https://cran.r-project.org/web/packages/MASS/MASS.pdf>
14. Breiman L, Cutler A, Liaw A, et al. <https://cran.r-project.org/web/packages/randomForest/randomForest.pdf>
15. Breast Cancer Wisconsin (Diagnostic) Data Set, <https://archive.ics.uci.edu/ml/datasets/>
16. W.N. Street, W.H. Wolberg and Mangasarian O.L. Nuclear feature extraction for breast tumor diagnosis. *IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology*, 1905, 861-870, 1993.
17. Wolberg W.H, Street W.N., and Mangasarian O.L. Machine learning techniques to diagnose breast cancer from fine-needle aspirates. *Cancer Letters*, 77, 163-171, 1994.
18. 小池 盛雄, 真鍋 俊明, 秋山 太ほか. 乳腺生検診断 - 進め方・考え方 (生検鑑別診断アトラス), 文光堂, 東京, 13-17, 1997.
19. Yamaguchi R, Tsuchiya SI, Koshikawa T, Ishihara A et al. Diagnostic accuracy of fine-needle aspiration cytology of the breast in Japan: report from the Working Group on the Accuracy of Breast Fine-Needle Aspiration Cytology of the Japanese Society of Clinical Cytology. *Oncol Rep*, 28 (5), 1606-12, 2012.
20. Giard RW, Hermans J. The value of aspiration cytologic examination of the breast. A statistical review of the medical literature. *Cancer*, 69 (8), 2104-10, 1992
21. 日本乳癌学会 (編). 2015年 Web 版ガイドライン 科学的根拠に基づく乳癌診療ガイドライン. <http://jbc.gr.jp/guideline/guideline/g6/g61860/>
22. 住吉 一浩, 野原 丈裕, 岩本充彦ほか. 乳腺診断における穿刺吸引細胞診の有用性に関する検討. *日本臨床外科学会*. 69 (12), 3028-3037, 2008
23. R Documentation. <https://www.rdocumentation.org/packages/RcmdrMisc/versions/1.0-6/topics/stepwise>
24. Wu CM, Zhang L, Wang Z, Christiani CD et al. Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection, *Bioinformatics*. 25 (9), 1145-1151, 2009.
25. Chen X, Ishwaran H. Random forests for genomic data analysis, *Genomics*, 99, 323-329, 2012.
26. Medical note. 「Radiomics」そして「Radiogenomics」. <https://medicalnote.jp/contents/160317-001-HV>
27. Mazurowski MA. Radiogenomics: what it is and why it is important, *J Am Coll Radiol*, 12 (8), 862-6, 2015.