Random Subspace によるロジスティック回帰モデルの構築

Building a Logistic Regression Model using Random Subspace

北原洋一 [1*]　折原良平 [1]

Youichi Kitahara　Ryohei Orihara

[1] 株式会社東芝 研究開発センター

[1]Corporate Research & Development Center, Toshiba Corporation

Abstract: A Logistic Regression Model using Random Subspace Method is investigated through experiments. Ensemble learning is known as one of better prediction methods. The framework makes it possible to improve precision of prediction, however interpretation of the model often gets difficult. Using random subspace method with logistic regression model, this paper tried to provide a solution to the problem. Precision improvement of the model is verified from a preliminary experiment. Furthermore the meaning of the combined model is easily understandable by means of coefficients of the prediction model.

1　はじめに

データマイニングのような機械学習的なアプローチの活躍が期待されている分野の一つとして医療分野がある．近年，医療分野においては EBM(Evidence Based Medicine)や EBH(Evidence Based Health)の推進にも見られるように，データに基づいた治療や健康指導に関心が高まっており，データマイニングを活用して興味深いルールやパターンを発見する試みもなされている．EBM や EBH に関するデータマイニングの適用事例としては，心不全の投薬内容と心機能検査成績に対する決定木解析[2]や，高血圧に関する系列マイニング[1]，パターンマイニングを利用した GHQ スコア分析[3]などがある．

一方，ルールやパターンの発見だけでなく，発症もしくは健康状態悪化の予測モデルについての関心も高い．

医療分野において，予測モデルを構築する目的は，主に二種類ある．一つは，ハイリスク者を抽出することである．発症するもしくは健康状態が悪化する確率の高い人を事前に特定することができれば，早期予防対策を施すことが可能になる．もう一つは，発症もしくは健康状態悪化に関わる要因を特定することである．影響の強い要因が明らかになれば，予防対策を施す際に，注力すべき内容を決めることができる．したがって，高精度かつ解釈の容易な予測モデルがもっとも望ましい．

*連絡先：(株)東芝研究開発センターシステム技術ラボラトリー
〒210-8582　神奈川県川崎市幸区小向東芝町 1 番地
E-mail: youichi.kitahara@toshiba.co.jp

しかしながら，説明変数によって構成される特徴量空間における予測対象の性質は単純ではないため，その双方の目的を同時に満たすのは困難である．決定木やニューラルネットをベースモデルとする集団学習や SVM(Support Vector Machine)を利用することで高精度な予測は可能になるものの，モデルの解釈は困難になってしまう．逆に，線形回帰モデルを利用する場合，解釈は容易になるけれども，精度は集団学習や SVM に及ばないことが多い．

本研究の目的は，高性能かつ解釈の容易な予測モデルを構築することである．予測モデル構築アルゴリズムとしては，Random Subspace に着目することにした．また，ベースモデルとしては，疫学調査などの医療分野でも利用されることの多いロジスティック回帰モデルを用いることにした．

ベースモデルにロジスティック回帰モデルを採用したのは，最終的に得られる予測モデルの解釈を容易にさせることを意図したためでもある．従来の Random Subspace を利用した研究では，決定木や回帰木を用いることが多い．しかしながら，決定木や回帰木をベースモデルにすると，最終的に複数のモデルが組み合わされることで得られるモデルの解釈は困難になってしまう．ロジスティック回帰モデルをベースモデルとした場合，線形結合させることで得られるモデルも当然のことながらロジスティック回帰モデルになるので，解釈が容易である．

また，Random Subspace に着目した理由は，次の二点にある．一つ目は，過去の研究事例が少なく，特性が十分解明されているとは言いがたいことである．例えば，ランダムに特徴量空間を縮小して生成

されたモデルを結合することで構築されるモデルの特性は，興味深い研究対象の一つであると思われる．二つ目は，モデル構築に利用される属性数が少ないため，多数の属性を有するデータについて高速な処理が期待できるという実用的利点があることである．これについては，後に述べる LogitBoost などの既存研究があるけれども，boosting とは異なった観点からの研究も必要であると思われる．

本稿は，6 章より構成される．2 章では，ロジスティック回帰モデルに関する簡単な説明を行う．3 章では，集団学習アルゴリズムについて述べる．4 章では，今回実験で利用したアルゴリズムについて説明する．5 章では，実験結果を紹介する．6 章では，まとめを行う．

2 ロジスティック回帰モデル

ロジスティック回帰モデルは，ロジスティック関数を用いて，以下のように表すことができる．

$$p = \frac{\exp(\beta_0 + \beta_1 x_1 + ,...,+\beta_i x_i)}{1 + \exp(\beta_0 + \beta_1 x_1 + ,...,+\beta_i x_i)} \quad \text{--- (1)}$$

ここで p は生起確率，x_k は k 番目の説明変，β_k は k 番目の説明変数の回帰係数を示す．式(1)は，p を算出するのには適しているが，回帰分析には適していない．そこで，回帰分析を行うときには，式(2)のように，左辺を p のロジットで表現し，右辺を説明変数の線形結合で表すようにする．

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + ,...,+\beta_i x_i \quad \text{--- (2)}$$

このモデルでは，交絡因子の調整が可能であり，交絡因子を除去したオッズ比を容易に算出できる．そのため，モデルの解釈が容易である．

3 集団学習アルゴリズム

3.1 Boosting

Boosting は，逐次レコードの重みを変化させながら複数のモデルを作成し，最終的に複数のモデルを結合することで高性能な予測モデルを生成するアルゴリズムである．多くの研究事例があるが，特に，AdaBoost は，理論的にも実用的にも優れたアルゴリズムとして知られている[6].

ロジスティック回帰を扱う Boosting アルゴリズムとしては，LogitBoost が有名である[7]．LogitBoost は，重み付け最小二乗法を行うことで生成される関数を逐次付加することで，ロジスティック回帰モデルを構築するアルゴリズムである．このアルゴリズムによって生成されるモデルは式(2)と同形のロジットモデルであるから，係数を調べることでモデルの解釈が可能である．

3.2 Bagging

Bagging は，ブートストラップ法を用いてレコードを復元抽出することで得られるレコード集合から複数のモデルを生成し，最終的に複数のモデルを結合することで高性能な予測モデルを生成するアルゴリズムである[4]．モデルの推定量の分散を減少させるため，ベースとなるモデルが不安定であっても，ロバストなモデルを構築することができる．

Boosting と比較すると，Bagging によって生成されるモデルはやや劣ることが多いことが報告されている．しかしながら，過去に生成したベースモデルに依存する Boosting と異なり，Bagging ではベースモデルを独立に生成するため，並列処理によって高速化が可能であるという利点もある．

3.2 Random Subspace

Bagging ではレコードを復元抽出していたが，Random Subspace では属性をランダムサンプリングすることでモデルを構築する[8]．ロジスティック回帰モデルの場合，利用する説明変数がランダムに選択されることになる．

ベースモデルが線形関数の場合，Random Subspace によって生成される各モデルを結合させる方法は，複数存在する．もっともシンプルなのは，係数を平均する方法であり，これによって生成されるモデルは BEM(Base Ensemble Method)として知られている．i 番目のモデルを f_i と表すと，N 個のモデルの結合によって生成されるモデル F は

$$F = \frac{\sum_{k=1}^{N} f_k}{N} \quad \text{--- (3)}$$

となる．さらに発展的な方法としては，適宜重み付けを行い結合する方法があり，これによって精度が向上することが実験的に確認されている[9]．しかしながら，本研究では，ランダムサンプリングによる効果に着目しているため BEM のみを用いた．重み付けをしたモデル結合方法の研究は今後の課題である．

3.4 Random Forest

Random Forest では，レコードと属性双方についてランダムサンプリングを行うことでモデルを構築する[5]. Random Forest を構成する木の剪定は行われないことと，ランダムサンプリングによって一回の学習で用いる訓練データ量が少ないことから，大きな

データセットを用いても比較的高速な処理が可能である．

4 Random Subspace を用いたロジスティック回帰モデル構築

4.1 結合モデルの回帰係数

BEM を構築する場合，複数のロジスティック回帰モデルを結合させる方法は二種類ある．一つ目は，生起確率のロジットを結合させる方法である．二つ目は，生起確率を直接結合させる方法である．

ロジット結合では，各モデルを結合するのは容易である．結合されるモデルの数を N，説明変数の種類数を M，j 番目に生成された i 番目の説明変数の係数を $\beta_{k,j}$，i 番目に生成されたモデルの生起確率を p_i とすると，結合モデルの生起確率 $\bar{p}$ のロジットは，

$$\log\left(\frac{\bar{p}}{1-\bar{p}}\right)=\frac{1}{N}\sum_{k=1}^{N}\log\left(\frac{p_k}{1-p_k}\right)$$
$$=\frac{1}{N}\left(\sum_{k=1}^{N}\beta_{k,0}+\sum_{k=1}^{N}\beta_{k,1}x_1+\cdots+\sum_{k=1}^{N}\beta_{k,M}x_M\right) \quad \text{--- (3)}$$
$$=\overline{\beta_0}+\overline{\beta_1}x_1+,...,+\overline{\beta_M}x_M$$

であるから，結合後のモデルに関する，各説明変数の係数および定数項 $\overline{\beta_j}$ は以下のようになる．

$$\overline{\beta_j}=\frac{\sum_{k=1}^{N}\beta_{k,j}}{N} \quad \text{--- (4)}$$

なお，Random Subspace では，各モデルはサンプリングにて選択された説明変数のみを利用して生成されるので，選択されていない説明変数の係数は 0 であることに注意されたい．

確率結合では，そのままでは明瞭な係数を得ることができない．そこで，結合モデルの生起確率は，ロジスティック関数によって近似できると仮定する．このとき，結合モデルの生起確率を $\bar{p}$ は

$$\bar{p}=\frac{\sum_{k}^{N}p_k}{N}$$
$$=\frac{\sum_{k}^{N}\frac{\exp(\beta_{k,0}+\beta_{k,1}x_1+,...,+\beta_{k,i}x_i)}{1+\exp(\beta_{k,0}+\beta_{k,1}x_1+,...,+\beta_{k,i}x_i)}}{N} \quad \text{--- (5)}$$
$$\approx\frac{\exp(\overline{\beta_0}+\overline{\beta_1}x_1+,...,+\overline{\beta_i}x_i)}{1+\exp(\overline{\beta_0}+\overline{\beta_1}x_1+,...,+\overline{\beta_i}x_i)}$$

と表現することができる．(5)の仮定が成立するならば，結合モデルの定数項 $\overline{\beta_0}$ および j 番目の説明変数の回帰係数 $\overline{\beta_j}$ は，

$$\overline{\beta_0}=\log\left(\frac{\sum_{k}^{N}\frac{\exp(\beta_{k,0})}{1+\exp(\beta_{k,0})}}{N-\sum_{k}^{N}\frac{\exp(\beta_{k,0})}{1+\exp(\beta_{k,0})}}\right)$$
$$\overline{\beta_j}=\log\left(\frac{\sum_{k}^{N}\frac{\exp(\beta_{k,0}+\beta_{k,j})}{1+\exp(\beta_{k,0}+\beta_{k,j})}}{N-\sum_{k}^{N}\frac{\exp(\beta_{k,0}+\beta_{k,j})}{1+\exp(\beta_{k,0}+\beta_{k,j})}}\right)-\overline{\beta_0} \quad \text{--(6)}$$

と算出される．なお，$\overline{\beta_j}$ は $\overline{\beta_0}$ に依存するので，$\overline{\beta_0}$ を算出した後 $\overline{\beta_j}$ を算出する必要があることに注意されたい．

4.2 アルゴリズム

結合方法ごとのアルゴリズムを，ロジット結合に関しては図１に，確率結合については図２示す．いずれも学習データおよびパラメータを入力とし，結合モデルの係数を出力としている．なお，本来の Random Subspace では，Bagging のようなレコードのサンプリングは行わないが，さらなる精度向上を期待して今回の実験では利用している．そのため，ロジスティック回帰を利用した Random Forest と同等のアルゴリズムになっている．

```
#DATA :学習データ
#BMN :ベースモデルの数
#VN :説明変数の数
#SR :レコードのサンプリング率
RS_LC(DATA, BMN, VN, SR)
{
  for (i in 1 : BMN) {
    #変数をランダムサンプリング
    D1 < - variable_sampling(DATA, VN);
    #レコードをランダムサンプリング
    D2 < - record_sampling(D1, SR);
    #ロジスティック回帰分析
    M < - GLM(D2);
    #   モデルの係数を加算
    β < - β + coef(M);
  }
  β < - β / BMN;
}
```

図 1　擬似コード（ロジット結合）

```
RS_APC(DATA, BMN, VN, SR)
{
  for (i in 1 : BMN) {
    #変数をランダムサンプリング
    D1 < - variable_sampling(DATA, VN);
    #レコードをランダムサンプリング
    D2 < - record_sampling(D1, SR);
    #ロジスティック回帰分析
    M < - GLM(D2);
    #   モデルの係数をVに保存
    V < - append(V, coef(M));
  }
  #定数項を算出
  for (i in 1 : BMN) {
    L < - L + exp(V[i,0]) / (1 + exp(V[i,0]));
  }
  β[0] < - log(L/(N - L));
  #説明変数の係数を算出
  for (i in 1 : BMN) {
    L < - L + exp(V[i, ] + V[i,0]) / (1 + exp(V[i, ]) + V[i,0]);
  }
  β < - log(L/(N - L)) - β[0];
}
```

図 2　擬似コード（確率結合）

5　実験

5.1　実験環境

実験データとして，UCI の機械学習リポジトリ(http://www.ics.uci.edu/~mlearn/databases)に含まれる breast-cancer と ionosphere を用いた．breast-cancer データでは，属性 irradiat の値 yes と no をそれぞれ 1 と 0 に変換し，目的変数とした．ionosphere では，最後の属性の値 b と g をそれぞれ 1 と 0 に変換し，目的変数として用いた．データのレコード数と属性数は以下の通りである．

データ名	レコード数	属性数
breast-cancer	286	9
ionosphere	351	33

図 3　データのレコード数と属性数

実験には，統計ソフト R を利用している．ベースモデルとなるロジスティック回帰モデルは MASS パッケージの glm()関数を利用し，サンプリングには sample()関数を使った．また，LogitBoost は，ada パッケージの ada()関数，Random Forest は randomForest パッケージの randomForest()関数を利用した．さらに，変数選択を利用したロジスティック回帰には，MASS パッケージの stepAIC()関数を利用した．以下の実験説明において特に記述がなければ，いずれの関数もデフォルトパラメータを利用している．

5.1　モデル性能

まず，Random Subspace によって構築されたモデルの性能を把握するために，他のアルゴリズムによって生成されたモデルとの性能比較を行った．

モデル性能の指標としては，ROC 曲線の AUC(Area Under Curve)を利用し，検証には 10 分割のクロスバリデーションを用いた．

Random Subspace によるモデル構築では，ロジット結合(RS_LC)，近似係数を利用した確率結合(RS_APC)，直接生起確率の平均を利用した確率結合(RS_DPC)の三種類を試みた．ランダム選択される説明変数の数は，breast-cancer データでは 2～8 個，ionosphere データでは，3～25 個を設定したものを試み，さらに，サンプリング率を 0.5～1 としたレコードのサンプリングを併用している．図 4 の値は，上記パラメータを設定して行った全ての実験において，ベースモデルが 40 から 50 までのときの値を平均したものである．

比較対象とするアルゴリズムは，LogitBoost(LB)，Random Forest(RF)，ロジスティック回帰(LR)，AIC によるモデル選択を行ったロジスティック回帰(LR_A)である．なお，LogitBoost は，サンプリング率を 0.5 から 1 まで 0.1 ずつ変化させてレコードのサンプリングを行っており，図 4 の値はそれらの平均値である．

	RS_LC	RS_APC	RS_DPC	LB	RF	LR	LR_A
breast-cancer	0.727	0.741	0.747	0.715	0.730	0.707	0.715
ionosphere	0.918	0.871	0.937	0.957	0.977	0.870	0.885

図 4　モデル性能一覧

breast-cancer データでは，Random Subspace は全般的に良好な結果が得られている．また，ionosphere データでも，LogitBoost や Random Forest にはやや及ばないものの，ロジット結合や直接確率結合では良好な結果が得られている．ionosphere データでは近似精度が悪いのか，近似確率結合の結果は通常のロジスティック回帰と同程度である．しかし，ロジット結合と直接確率結合では，通常のロジスティック回帰を上回る結果となっている．

次に，選択される説明変数の数が精度に及ぼす影響を見るために，説明変数の数と AUC の関係を図 5 と図 6 に示す．

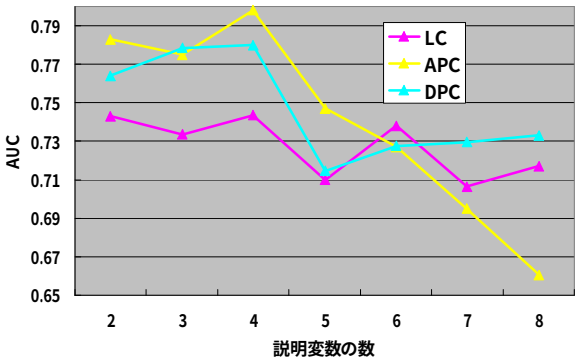

図 5　説明変数の数と AUC（breast-cancer）

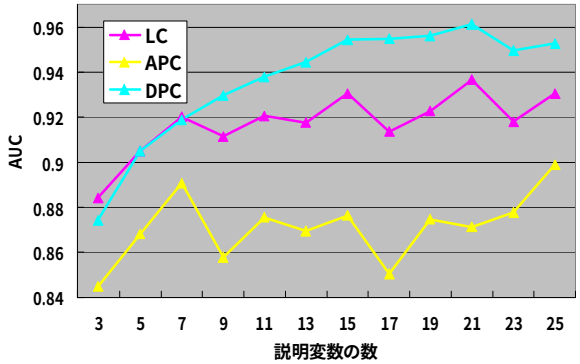

図 6　説明変数の数と AUC（ionosphere）

説明変数の数が増加するにしたがって，breast-cancer データでは性能が低下し，ionosphere データでは逆に向上している．このことは，モデル性能は説明変数の数に依存していることと，その特性はデータに依存していることを示している．

次に，サンプリング率が精度に及ぼす影響を見るために，サンプリング率と AUC の関係を図 7 と図 8 に示す．なお，参考のため LogitBoost の結果も示している．

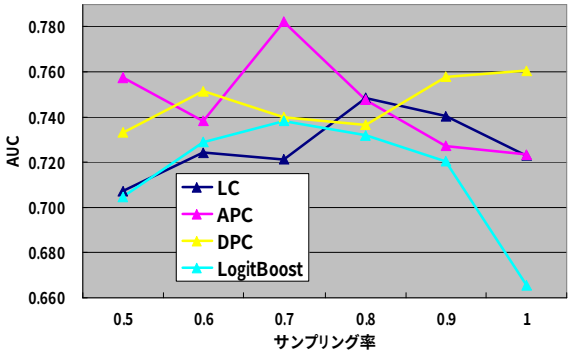

図 7　サンプリング率と AUC（breast-cancer）

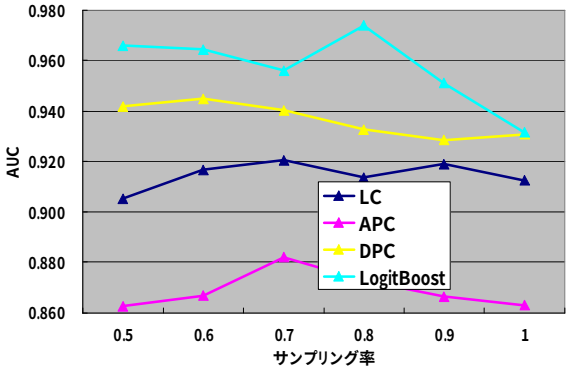

図 8　サンプリング率と AUC（ionosphere）

図 7 と図 8 より，サンプリング率の違いによってモデル性能はわずかに異なっているようであるが，明瞭な傾向などはないことがわかる．

5.2　モデル解釈

モデル解釈について検討するために，比較的良好な結果が得られている breast-cancer データの全レコードを学習データとし，説明変数の数を 3，レコードのサンプリング率を 0.8 としたときの結果を例として紹介する．これは，高性能なモデルが構築されているときの方が，安定していて解釈しやすいと考えたためである．また，説明変数の係数の多くが同様の傾向を示すので，ここでは係数の絶対値が最も大きい説明変数「inv.nodes12-14」を例に取り上げる．

まず，モデル性能の収束状況を見るために，ベースモデルの数と AUC の関係を図 9 に示す．

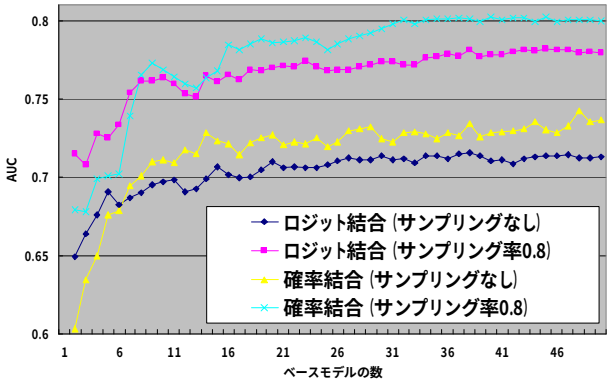

図 9　ベースモデルの数と AUC

図 9 より，サンプリング率 0.8 のロジット結合は収束状況が悪いものの，他のケースではベースモデルの数が 30 を超えたあたりからは性能がほぼ安定していることがわかる．属性数 9，説明変数の数 3 であるから，ベースモデルは 84 種類あり，この半数に達しないうちにモデル性能が安定していることがわかる．

次に，係数の変化と係数の分散の変化を見るために，ベースモデルの数と「inv.nodes12-14」変数の係数との関係を図 10 に示す．

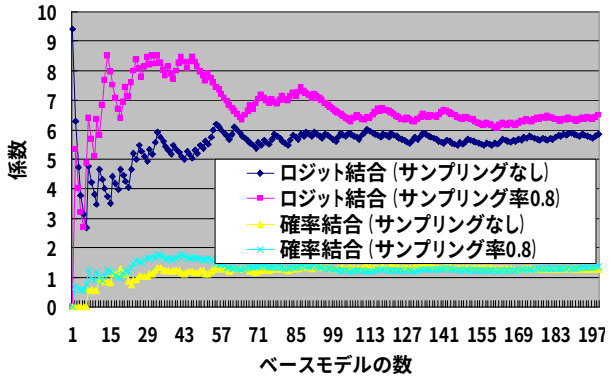

図 10　ベースモデルの数と係数(inv.nodes12-14)

図 11 より，ベースモデルの数が 30 より大きいときでも係数は大きく変動していることがわかる．AUC がほぼ安定していても，係数は安定していないようである．すべてのベースモデルが出現する前，つまり，ベースモデルの数が 84 に達する前にランダム化の効果がもっとも表れていると考えられるため，モデル解釈の観点からは，この範囲において係数が安定するのが望ましいけれども，そのようなケースは稀である．

また，同一の変数の係数であるにも関わらず，ロジット結合と確率結合とでは値が大きく異なっており，確率結合の係数はかなり小さい．この傾向は，他の説明変数の係数についても見られた．さらに，単一のロジスティック回帰モデルの係数と比較して，ロジット結合，確率結合とも係数が小さくなる傾向がある．

次に，説明変数の数が係数の値に与える影響を見るために「inv.nodes12-14」変数の係数と説明変数の数の関係を図 11 に示す．

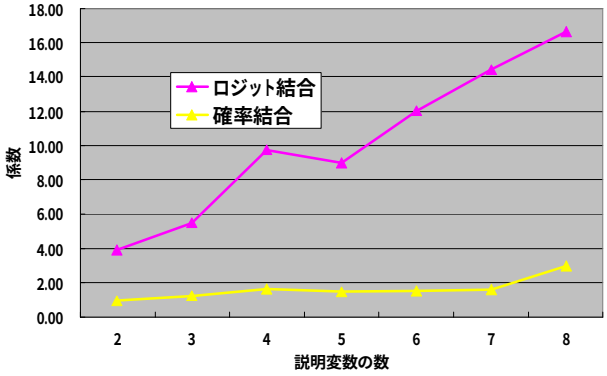

図 11　説明変数の数と係数(inv.nodes12-14)

図 11 より説明変数の数が増加するにつれて係数も大きくなっており，係数の大きさは説明変数の数に依存することがわかる．ベースモデルが多くなると，選択可能な説明変数のすべての組み合わせで生成したモデルを単純に結合したモデルに収束していくと考えられる．係数の大きさはこの最終的に収束するモデルに強く依存していることが予想され，同時に係数の取り得る値が制限されることから，モデル性能の限界についてもこの最終的に収束するモデルに強く依存していることが考えられる．

モデルを解釈する上では，ランダム化の効果によって，説明変数の数などのパラメータに影響されず，ある一定値に係数が収束してくれるのが望ましい．しかしながら，今回の実験の範囲では，そのような傾向は見られなかった．つまり，説明変数の数やベースモデルの数が異なる同様の性能を有するモデルが複数存在してしまう．現在のところ複数存在する高性能モデルから最適なモデルを選択する方法がないため，モデル性能を向上させることができ係数を得ることができても，モデルの解釈を行うには問題がある．解釈可能な高性能モデルを実現するためには，係数が安定しているかを判定する方法，さらには，安定した係数を得る方法を開発する必要がある．

最後に，確率結合の近似の有効性を見るために，近似を行わないときの AUC と近似を行ったときの AUC の差分を図 12 と図 13 に示す．

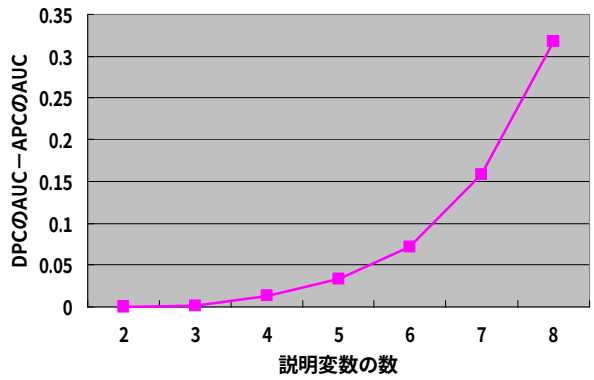

図 12　説明変数の数と AUC の差分（breast-cancer）

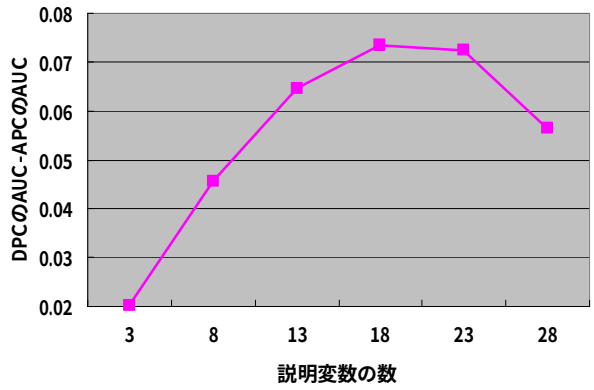

図 13　説明変数の数と AUC の差分（ionosphere）

breast-cancer データでは説明変数の数が増加するにつれて AUC の差分も増加し，ionosphere データでも説明変数の数が多いときに AUC の差分が大きい傾向にあることがわかる．理由はわからないが，確率結合の近似は，説明変数の数が多いと適切ではなくなるようである．また，AUC の差分の大きさは，データによって異なっていることから，近似の有効性はデータの特性にも依存するものと思われる．

6　まとめ

Random Subspace を用いてロジスティック回帰モデルを構築する試みを行った．Random Subspace によって構築されたモデルの性能は，適切なモデル結合方法を採用することによって，通常のロジスティック回帰の性能を上回ることが確認された．しかしながら，データによっては，Random Forest や LogitBoost の方が高性能モデルを構築しており，優劣はつけがたい結果となっていた．

モデルを結合する際に生起確率の和を用いた場合でも，係数を近似値として算出することで，結合モデルの回帰係数を知ることが可能になった．しかしながら，係数が安定せず，さらに説明変数やベースモデルの数に強く依存してしまうため，得られたモデルの係数をそのまま解釈してよいかには疑問が残る．今後は，他データでの実験や理論的考察を行うことで，係数の安定性を判定する方法や，最適な係数を決定する方法を開発していく必要がある．

参考文献

[1] 植野ら: 依存関係に着目した系列パターン再構成, 第 140 回知能と複雑系研究会, (2005)

[2] 金智隆ら，科学的根拠に基づく医療(EBM:Evidence-Based Medicine)におけるデータマイニングの適用事例と今後の展望, 課題について, 人工知能学会誌, 19 巻, 6 号, pp.710-711, (2004)

[3] 林ら: 蓄積された健康診断データからの知識発見 –GHQ スコアの変化と問診回答変化の関係-, 第 79 回日本産業衛生学会, (2006)

[4] Breiman, L.: Bagging Predictors, Machine Learning, Vol. 24, No. 2, pp.123-149, (1996)

[5] Breiman, L.: Random Forests, Machine Learning, Vol. 45, pp.5-32, (2001)

[6] Freund, Y. et. al.: Experiments with a New Boosting Algorithm, Proceedings 13th International Conference on Machine Learning, pp.148-156, (1996)

[7] Friedman, J. et. al.: Additive Logistic Regression: A Statistical View of Boosting, The Annals of Statistics, Vol. 28, No. 2, pp.337-407, (2000)

[8] Ho, T. K.: The Random Subspace Method for Constructiong Decision Forests, IEEE Trans Pattern Analysis and Machine Intelligence, Vol 20, No. 8, pp.832-844,(1998)

[9] Rooney, N. et. al.: Random Subspacing for regression ensembles, FLAIRS Conference 2004, (2004)

[10] Skurichina, M. et. al.: Bagging, Boosting and the Random Subspace Method for Linear Classifiers, Pattern Analysis & Applications, Vol. 5, pp.121-135, (2002)