高信頼化モデル構築用サンプルを用いた GMDH による生理活性値推定

山森　一人[1)]・春日亀　裕也[2)]・吉原　郁夫[3)]

Physiological Activity Estimation by Group Method of Data Handling with Dependable Samples based on Statistical Analysis

Kunihito YAMAMORI, Yuya KASUGAME, and Ikuo YOSIHARA

Abstract

Recent years, keeping our health through daily meals is attracted because some foods have physiological activity to affect our biological activities. However, it is very difficult to evaluate physiological activities, and it is impossible to evaluate them for all kinds of foods. Therefore some researchers have been proposed methods to estimate physiological activities from protein expression levels. Physiological activities and protein expression levels are measured on living cells. These values include relatively large noise, and it leads low estimation accuracy. In this report, we propose to use only dependable samples based on statistical analysis to make an estimation model. Experimental results said our approach improved estimation accuracy.

Keywords: physiological activity, protein expression levels, GMDH, Smirnov-Grubbs test

1. はじめに

近年、食生活の欧米化や栄養バランスの偏りから、がんや肥満を始めとする生活習慣病が増加しており、これを食品がもつ機能により予防しようとする試みがなされている。このため、生体調節機能である食品の第三次機能に関心が寄せられている[1)]。第三次機能の評価指標として生理活性値があるが、生理活性値を調べるには多大な労力と時間が必要になる。そこで、生理活性値よりも測定が容易であるバイオマーカ発現量から生理活性値を推定する手法が提案されている。生理活性値の推定に用いる手法として、これまでにニューラルネットワーク[2)]や自己組織化マップ(Self-Organizing Map)[3)]、遺伝的プログラミング(Genetic Programming)[4)]などが提案されている。しかし、これまでの推定実験では十分な推定精度が得られていない。その原因として、モデル構築や学習に用いるサンプルの測定値に、測定環境や細胞状態の変化に由来する、測定誤差を超えるノイズが含まれており、それが推定精度向上の妨げになっていることが考えられる。

本研究では、生理活性値の推定精度向上を目的とし、測定値のばらつきが少ない測定対象物のみを推定モデル構築に用いること、及びスミルノフ・グラブス検定で大きな誤差を含むと考えられる外れ値を除外することによって、モデル構築用サンプルの信頼性を高めることを提案する。また、その効果を、Group Method of Data Handling(GMDH)を用いて検証した。

1)　情報システム工学科准教授
2)　情報システム工学科学部生
3)　情報システム工学科教授

2. 実験方法

1968 年に Ivakhnenko[5)]により考案された Group Method of Data Handling(GMDH)は、簡単な非線形式を組み合わせて複雑な非線形モデルを自己組織化的に構成していく手法であり、今日まで予測、モデリングなどに幅広く応用されている[6)7)]。

モデル構築のための基本関数は 2 変数の 2 次式であり、(1)式で表される。

$$G(x_i, x_j) = a_0 + a_1 x_i + a_2 x_j + a_3 x_i x_j + a_4 x_i^2 + a_5 x_j^2. \quad (1)$$

以下に、GMDH によるモデル構築アルゴリズムについて説明する。

システムへの入力変数を $x_1, x_2, \cdots, x_m$、出力変数を y とする。まず、m 個の入力変数の二つずつの組み合わせを考える。それぞれの組み合わせに対し y を最もよく近似するように、$G(x_i, x_j)$の係数 a_k (k =0,1,⋯,5)を決定する。入力変数を二つ組み合わせた $u_{ij} = G(x_i, x_j)$ $\{i, j$ =1,2, $\cdots, m; i<j\}$は ${}_mC_2$ 通りあるが、その中から、出力 y との二乗誤差が小さいものから順に p 個を選択する。以下では、選択された p 個を $u_1, u_2, \cdots, u_p$ と記す。選択した $u_1, u_2, \cdots, u_p$ を改めて入力変数とみなし、同様に、それ以上高次の組み合わせを作っても出力 y への近似度が上がらなくなるまで、同じ操作を繰り返す。

3. 測定データの前処理

推定に用いる測定データは、化合物と食品抽出物を測定対象とし、化合物30種類を各3種類の濃度で、食品抽出物21種類を各1～3種類の濃度でHepG2細胞に与えたときのバイオマーカ発現量と生理活性値である。測定対象物の総数は化合物が90種類、食品抽出物が40～41種類となっている。

3.1. 測定項目

バイオマーカ発現量はThioredoxin、Survivin、HSP70、XIAP、FADD、TXNRD1、HSP90、MxA、tNOX、NQO1、ERK2、p53、Bcl2の13種類で測定されている。

生理活性値はがん細胞増殖抑制活性、抗炎症活性、抗酸化ストレス活性、血管新生阻害活性、抗血管新生活性、抗転移活性の6種類が測定されている。

3.2. 測定値の相対化

推定に用いるバイオマーカ発現量、及び生理活性値は、各測定対象物において当該測定対象物を細胞に与えなかった時の測定値の平均値によって相対化した値を用いる。したがって、バイオマーカ発現量と生理活性値は、当該測定対象物を与えなかった時の値を1.0とした相対値となる。

3.3. 分散による測定対象物の選定

生理活性値の推定精度を向上させるには、測定誤差を超えると思われる測定値を含むと考えられる測定対象物を除外したうえでモデル構築を行う必要がある。そこで、同一対象物の同一濃度での複数個の測定値の分散が、あるしきい値を超える測定対象物データは信頼性が低いとみなし、その測定対象物はモデル構築用サンプルから除外した。しきい値は0.2、0.1、0.075、0.05の4つとし、分散がそのしきい値以下の測定対象物データをモデル構築用サンプルに用いた。

測定対象物は化合物が90種類と食品抽出物が40～41種類、各々にバイオマーカ発現量が13種類と生理活性値が6種類測定されているので、測定対象物データの総数は2,489個である。これら全てで分散を求め、各しきい値で測定対象物の選定を行った。

13種類のバイオマーカのうち、しきい値0.1と0.05でモデル構築に用いることができると判定されたデータ数を表1に示す。表1から、バイオマーカFADDのデータ数が他のマーカデータと比べ極端に少ないことがわかる。したがって、FADDは測定値のばらつきが非常に大きく、モデル構築用サンプルに用いるのに適していないと考えられる。そのため、本研究ではFADDを除く12種類のマーカデータを用いた。

同様に、しきい値0.1と0.05でモデル構築に用いることが可能と判定された生理活性値のデータ数を表2に示す。

表 1. 各しきい値でのマーカデータのデータ数.

マーカ名	データ数			
	化合物		食品抽出物	
	0.1	0.05	0.1	0.05
Thioredoxin	80	71	37	34
Survivin	85	66	38	34
HSP70	85	76	41	36
XIAP	83	75	41	39
FADD	35	18	23	18
TXNRD1	77	66	41	39
HSP90	77	56	38	32
MxA	84	76	40	35
tNOX	82	64	39	36
NQO1	74	65	34	28
ERK2	82	71	39	35
p53	84	71	39	37
Bcl2	85	73	40	39

表 2. 各しきい値での生理活性値のデータ数.

生理活性名	データ数			
	化合物		食品抽出物	
	0.1	0.05	0.1	0.05
がん細胞増殖抑制活性	90	90	41	41
抗炎症活性	68	55	40	40
抗酸化ストレス活性	71	52	35	25
血管新生阻害活性	90	90	41	41
抗血管新生活性	86	81	41	40
抗転移活性	90	90	41	41

3.4. スミルノフ・グラブス検定

モデル構築用サンプルの信頼性をさらに高めるため、スミルノフ・グラブス検定により測定値の中から外れ値を除外した。以下、スミルノフ・グラブス検定に関して簡単に述べる[8]。

測定データ $\{x_1, x_2, \cdots, x_N\}$ において、データ x_k だけが非常に外れているとする。

① 帰無仮説 H_0 と対立仮説 H_1 をたてる。ここで、帰無仮説 H_0 は「x_k は外れ値ではない」、対立仮説 H_1 は「x_k は外れ値である」である。

② この外れ値の検定統計量 $T(x_k)$ を計算する。検定統計量 $T(x_k)$ は、x_k が最大値の場合、

$$T(x_k) = \frac{x_k - \bar{x}}{s}$$

、x_k が最小値の場合、

$T(x_k)=\frac{\bar{x}-x_k}{s}$ となる。ただし、s は標準偏差である。

③ スミルノフ・グラブス検定の数表から有意点を求め、$T(x_k)$ が棄却域に入る場合に外れ値とし、モデル構築用サンプルから除外する。

3.5. モデル構築用サンプルの作成

検定後の測定値を用いてモデル構築用サンプルを作成する。バイオマーカ発現量、及び生理活性値は同一測定対象物の同一濃度であっても全て独立に測定されているため、表 3 のようにモデル構築用サンプルとして対応付けを行う必要がある。

本研究では、学習サンプルの作成を 2 通りの方法で行った。1 つ目は、同一測定対象物の同濃度ごとに複数回測定された値の平均値を用いる方法である。この場合、平均値 1 つを代表値として用いるため、バイオマーカ発現量と生理活性値の組み合わせは一意に決定できる。こうして作成したサンプルは平均値サンプルと呼ぶ。2 つ目は、バイオマーカ発現量と生理活性値を単回帰分析 [9]によって対応付ける方法である。単回帰分析を応用して対応付けたサンプルは単回帰サンプルと呼ぶ。

以上、2 つの方法をモデル構築用サンプルの作成方法とした。

表 3. モデル構築用サンプルの例.

測定対象物	バイオマーカ発現量				生理活性値
	Thioredoxin	Survivin	…	Bcl2	
Gaba-100	0.92	1.17	…	1.04	0.88
EGC-10	0.75	0.84		0.78	0.86
…	…	…	…	…	…
IFN-100	0.89	0.93	…	1.00	0.87

4. 生理活性推定実験

4.1. 実験条件

本研究では、Group Method of Data Handling(GMDH)を用いてバイオマーカ発現量から生理活性値の推定実験を行った。入力変数はバイオマーカ発現量が 12 個で、出力変数は生理活性値が 1 つである。

モデル構築用サンプルは化合物の測定データと食品抽出物の測定データの両方を含むものとし、検証用に抜き出すサンプルを変えて 5 種類作成した。

推定精度は、(2)式に示した生理活性値の実測値と GMDH による推定値間の平均二乗誤差 e により評価する。

$$e=\frac{1}{n}\sum_{i=1}^{n}(y_i-y_i')^2. \quad (2)$$

ここで、n は検証用データの個数、y_i は生理活性の実測値、y_i' は推定値を示している。

4.2. がん細胞増殖抑制活性推定実験結果と考察

平均値サンプルと単回帰サンプルを用いて、しきい値ごとにがん細胞増殖抑制活性の推定実験を行った場合の平均二乗誤差を図 1 に示す。図 1 は 5 種類のサンプルでの推定結果の平均値を示している。また、推定実験を行った際のサンプルデータ数を表 4 に示す。

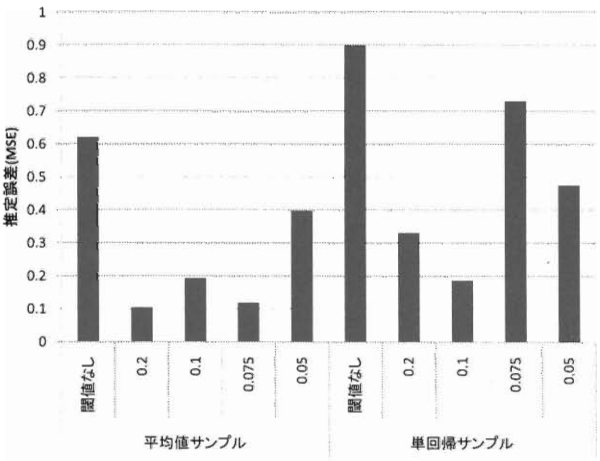

図 1. がん細胞増殖抑制活性推定結果（average）.

表 4. がん細胞増殖抑制活性推定用サンプルデータ数.

しきい値	平均値サンプル		単回帰サンプル	
	モデル構築用	検証用	モデル構築用	検証用
0.2	80	23	436	115
0.1	60	17	292	76
0.075	52	12	240	62
0.05	28	10	148	36

平均値サンプルに関しては、測定値のばらつきが大きい測定対象物と外れ値の除外により、推定精度の向上が得られた。ただし、平均値サンプルは、サンプルデータ数が少ないため、モデル構築用サンプルに含まれる測定対象物の組み合わせの影響を強く受ける。しきい値 0.1 の場合の平均値サンプルでの、測定対象物ごとの推定値と実測値を図 2 に示す。

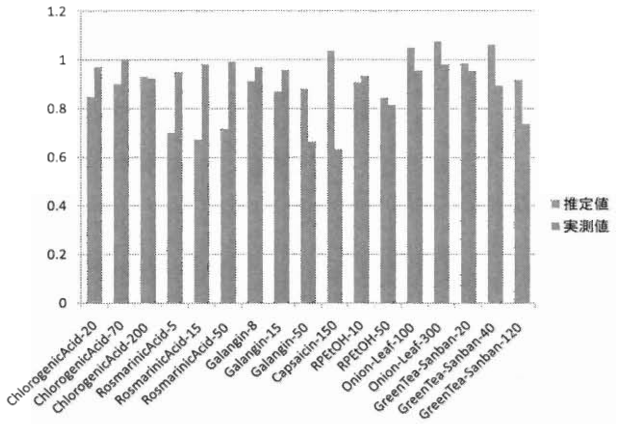

図 2. 平均値サンプル，しきい値 0.1 での細胞増殖抑制活性推定実験結果.

単回帰サンプルについては、図 1 を見てわかるように、測定対象物と外れ値の除外による推定精度の向上が見られたが、分散に対するしきい値を小さくとりサンプルを過度に除外すると推定精度の低下が起こっている。このことから、適切な基準でばらつきの大きな測定値を除外することで、推定精度向上につながるといえる。

4.3. 抗炎症活性推定実験結果と考察

平均値サンプルと単回帰サンプルを用いて、しきい値ごとに抗炎症活性の推定実験を行った場合の平均二乗誤差を図 3〜図 5 に示す。図 3 は 5 種類のサンプルの推定結果のうち、平均二乗誤差が最も小さかったもの、図 4 は平均二乗誤差が最も大きかったもの、図 5 は 5 種類のサンプルの推定結果の平均値を示している。また、推定実験を行った際のサンプルデータ数を表 5 に示す。

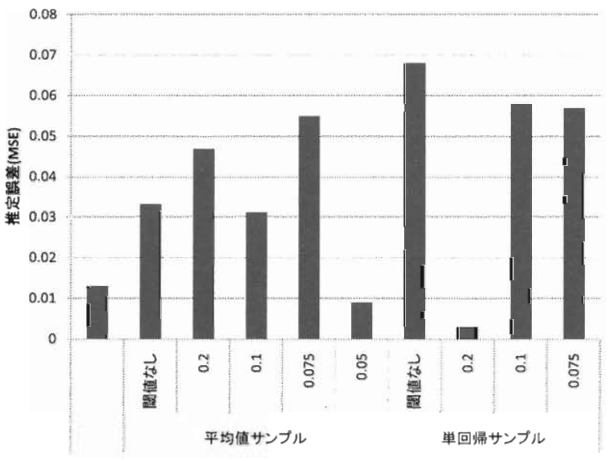

図 3. 抗炎症活性推定結果（best）.

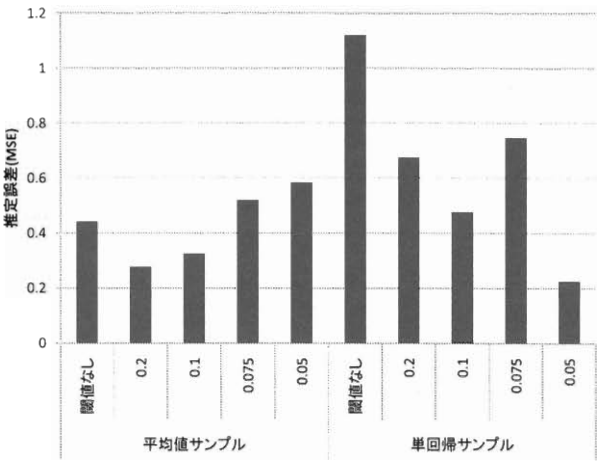

図 4. 抗炎症活性推定結果（worst）.

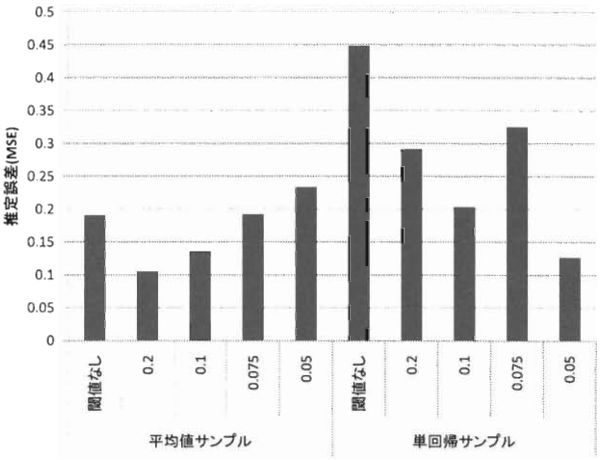

図 5. 抗炎症活性推定結果（average）.

表 5. 抗炎症活性推定用サンプルデータ数.

しきい値	平均値サンプル		単回帰サンプル	
	モデル構築用	検証用	モデル構築用	検証用
0.2	68	22	380	98
0.1	48	14	232	62
0.075	40	10	184	49
0.05	20	6	100	26

平均値サンプルに関しては、図 5 から、しきい値 0.2 と 0.1 の場合に推定精度の向上が見られた。測定対象物の選定と外れ値の除外を行った平均値サンプルの中で、最も推定精度の高かった図 3 のしきい値 0.05 の場合を例に、測定対象物ごとの推定値と実測値を図 6 に示す。

単回帰サンプルに関しては、図5から、測定対象物の選定と外れ値の除外を行ったことで推定精度が向上していることがわかる。しかし、しきい値0.075のとき、しきい値0.01に比べ平均二乗誤差が大きくなっている。そこで、しきい値0.075における平均二乗誤差の最良値と最悪値に着目する。図3と図4から、最良値はおよそ0.06であるのに対し、最悪値はおよそ0.75と、大きな差が見られる。ここで、表5から、しきい値0.075におけるモデル構築用データ数は184個であり、サンプルデータ数は十分と考えられる。このことから、モデル構築用サンプルに含まれる測定対象物の違いが推定精度に影響を与えていると考えられる。

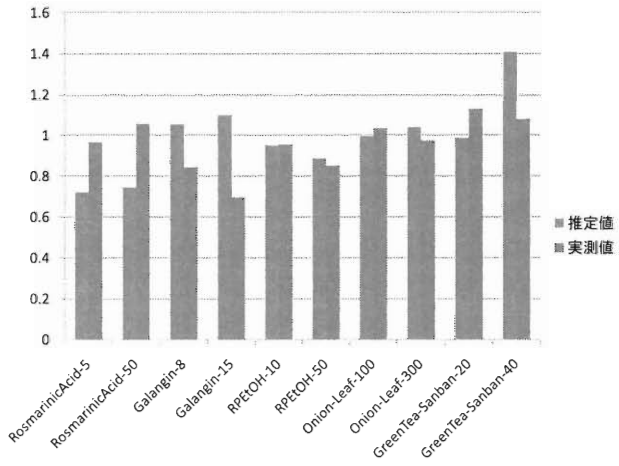

図 6. 平均値サンプル，しきい値0.075での抗炎症活性推定結果.

4.4. 抗酸化ストレス活性推定実験結果と考察

平均値サンプルと単回帰サンプルを用いて、しきい値ごとに抗酸化ストレス活性の推定実験を行った場合の平均二乗誤差を図7〜図9に示す。図7は5種類のサンプルの推定結果のうち、平均二乗誤差が最も小さかったもの、図8は平均二乗誤差が最も大きかったもの、図9は5種類のサンプルの推定結果の平均値を示している。また、推定実験を行った際のサンプルデータ数を表6に示す。

平均値サンプルに関しては、図9から、測定対象物の選定と外れ値の除外を行うことで、全てのしきい値で推定精度の向上が見られた。しかし、表6から分かるとおり、平均値サンプルではサンプル数が少ないため、モデル構築や検証に用いた測定対象物の組み合わせの影響を強く受ける。図7のしきい値0.1の場合を例に、測定対象物ごとの推定値と実測値を図10に示す。

単回帰サンプルを用いての推定結果は、図7〜図9の全てにおいて、しきい値0.2の場合に推定精度の向上が見られた。このことから、適切な基準でばらつきの大きな測定値を除外したことで、推定精度向上につながったと言える。単回帰サンプルの平均値では、図9からわかる通りしきい値0.2が最も精度が良く、これ以上小さいしきい値ではかえって精度が悪化する。このことから、モデル構築に用いるサンプルの質とともに、数も重要であることがわかる。

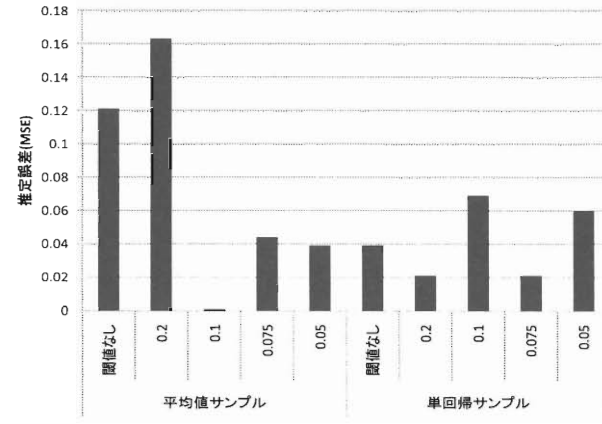

図 7. 抗酸化ストレス活性推定結果（best）.

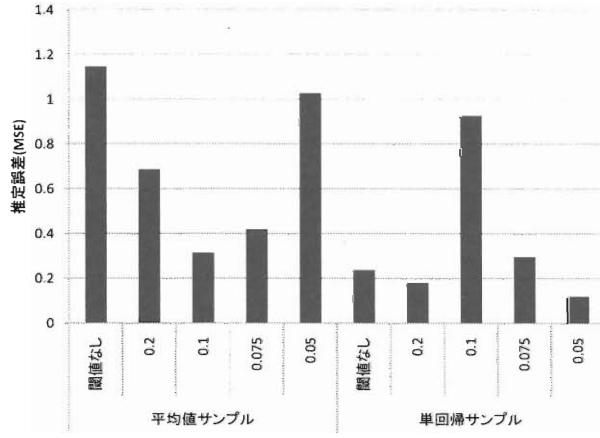

図 8. 抗酸化ストレス推定結果（worst）.

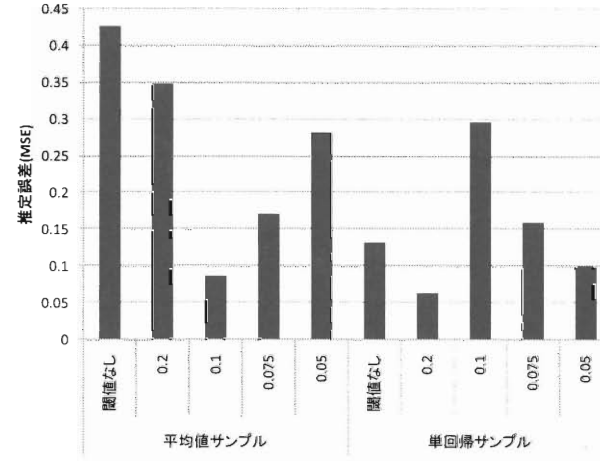

図 9. 抗酸化ストレス活性推定結果（average）.

表 6. 抗酸化ストレス活性推定用サンプルデータ数.

しきい値	平均値サンプル		単回帰サンプル	
	モデル構築用	検証用	モデル構築用	検証用
0.2	76	22	416	107
0.1	48	17	248	62
0.075	40	11	192	49
0.05	16	7	88	22

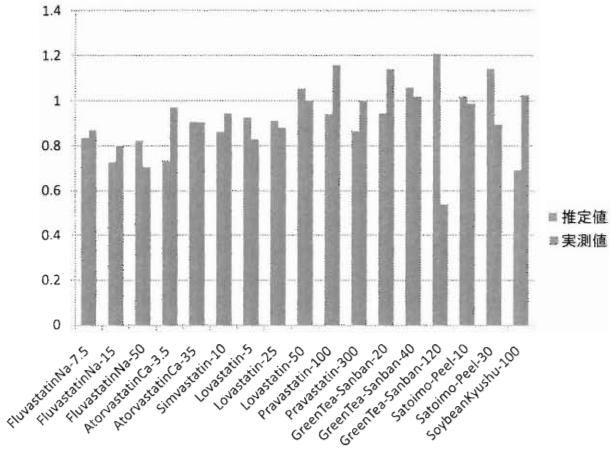

図 10. 平均値サンプル，しきい値 0.1 での抗酸化ストレス活性推定結果.

5. おわりに

本研究では、生理活性値の推定精度向上を目的とし、測定値のばらつきが大きい測定対象物をサンプルから除外し、さらにスミルノフ・グラブス検定により外れ値を除外することで、信頼性の高いサンプルを作成して推定に用いる手法を提案した。推定手法としては、Group Method of Data Handling(GMDH)法を用いた。測定データの信頼性向上処理として、はじめに分散があるしきい値を超える測定対象物の除外を行った。しきい値は 0.2、0.1、0.075、0.05 とし、分散がそのしきい値以下の測定対象物データをサンプルに用いた。また、スミルノフ・グラブス検定で外れ値を除外してサンプルを作成することにより、さらに測定データの信頼性を高めた。一方、これらの処理を一切行わない測定データを用いたサンプルを作成したうえで生理活性値推定実験を行い、推定結果を比較した。その結果、測定値のばらつきが大きな測定対象物を除外してモデル構築を行うことで、推定精度の向上が見られた。一方、どの測定対象物をモデル構築用に用いるかにより、同じ条件で選択したサンプルでもその推定精度が異なる場合があった。これは、モデル構築に用いるサンプルに推定精度が大きく依存することを意味する。このことから、どのサンプルがモデル構築の精度に影響を与えているか詳しく調査することが今後の課題として挙げられる。

謝辞

本研究は、独立行政法人科学技術振興機構・地域結集型共同研究事業「食の機能を中心としたがん予防基盤技術創出」の一部として行われ、バイオマーカ発現量や生理活性値は宮崎大学農学部及び宮崎県産業支援財団コア研究室にて測定されたものである。関係各位に感謝する。

参考文献

1) 青柳康夫編著, 有田政信[他]共著: 改訂 食品機能学, 建帛社, 2008.
2) Christopher M. Bishop: PATTERN RECOGNITION AND MACHINE LEARNING, Springer-Verlag, 2006.
3) マーク M. ヴァン・フッレ著, 徳高平蔵, 藤村喜久郎監訳: 自己組織化マップ - 理論・設計・応用, 海文堂出版, 2001.
4) 伊庭斉志著: 遺伝的プログラミング入門, 東京大学出版会, 2001.
5) A.IVAKHNENKO: Polynomial Theory of Complex Systems, IEEE Trans.Systems, Vol.SMC-1,No.4,pp.364-378, 1971.
6) 吉原郁夫, 佐藤周一: GA を用いた非線形モデル構築の最適化－GA と GMDH の融合－, 情報処理学会研究報告（人工知能研究会）, 96(78), 1-6, 1996.
7) 山森一人, 岩崎敬太, 吉原郁夫: GMDH によるタンパク質発現量からの食品生理活性値の推定, 宮崎大学工学部紀要, 38, 355-360, 2009.
8) 石村貞夫著: 入門はじめての統計解析, 東京図書, 2010.
9) 盛山和夫著: 統計学入門, 放送大学教育振興会, 2004.