クラスタリングにおけるパラメータ決定のための内的規準最適化

石川 智己 † 岡留 剛 ‡
Tomoki Ishikawa Takeshi Okadome

1. 概要

教師なし学習であるクラスタリングは実行時にパラメータを設定する必要があり，手動に頼ることが多い．このパラメータによりクラスタリングの精度は大きく変わるため，この決定問題は重要である．本研究ではクラスタリングアルゴリズムのパラメータを最適決定するために，新たな内的規準を構築することを目指す．クラスタリング結果の評価指標は外的規準と内的規準に大別され，外的規準は正解クラスタを必要とするため，教師なし学習であるクラスタリングの結果の評価としてはふさわしくない．内的規準は正解クラスタを必要とせず，データそのものと予測クラスタによって評価される．パラメータの最適化のためにクラスタリング結果を評価するには，内的規準を用いることが望ましい．

本研究では，目標となる外的規準を与え，その外的規準が算出する評価値に近い値を出す内的規準を構築する．具体的には既存の内的規準の重み付き線形和によって新たな内的規準を構築する．Monte Carlo study によって学習用人工データを生成し，lasso によって線形重みを決定した．ピアソンの相関係数により評価を行い，外的規準による値に対して，提案手法による内的規準による値が既存の内的規準値より相関が高くなることが示された．また，実データに対するパラメータの決定でも正しくパラメータが決定されることを確認した．

2. はじめに

データを効率的に処理するための一手法としてクラスタリングが挙げられる. クラスタリングはデータを分割し，グループごとの特徴を見いだす知識発見や情報の集約・圧縮などに用いられることが多い．これらは前処理として用いられることが多いが，深い知識を持たずに利用していると思われる事例が散見される．

人工知能学会全国大会過去 5 年（2013-2017）分の大会論文集 [1] を調査したところ，クラスタリングにあたってパラメータの設定を考慮せずに使用している事例が見られた．データの前処理という基礎段階で適切な評価をせずパラメータを決定することは大きな問題であると考えられる．

クラスタリング結果の評価規準には大きく分けて外的規準と内的規準がある．外的規準は正解クラスタと予測クラスタを比較することによって評価を行う．教師なし学習であるクラスタリングにおいて，いわゆる正解はないため外的規準による評価はふさわしくない．一方，内的規準はデータそのものと予測クラスタによって評価を行う．その意味で内的規準の方が，クラスタリング結果の評価規準として適切である．

そこで本研究では，目標とする外的規準を決め，真のクラスラベルを必要としない新たな内的規準を定める手法を提案する．本手法により生成された規準による評価値を最大化することでクラスタリングにおけるパラメータの最適決定が可能になる．

外的規準と内的規準間の関係を学習するために，人工データの生成が必要となる．確率的生成モデルを仮定し，人工データを生成した．Milligan[2] は外的規準を元として，良い内的規準を見つけるために人工的にデータを生成し，相関係数によって評価を行っている．しかし，データの生成がガウス分布に限定されており，外的規準と内的規準の関係を見るためのデータとしては十分とは言えない．そこで，本研究では Milligan のデータ生成法を基礎として拡張を試みた．複数の分布族の無限混合分布からサンプリングすることで人工データを生成する確率的生成モデルを仮定した．

また，新たな内的規準の構築は既存の内的規準の重み付き線形和とした．目標とする外的規準値との二乗和誤差を最小とする線形重み $\mathbf{w}$ の回帰問題として，lasso を使い決定した．使用する分布の数を変え，$\mathbf{w}$ の比較を行い，一定の頑健性が見られることを確認した．

新たに構築した内的規準の評価として 2 種類の実験を行った．1 つはピアソンの相関係数を使用した実験で，目標とする外的規準に対して，提案手法による内的規準が既存の内的規準より相関が高くなることが示された．2 つ目は，実データに対する評価実験で，アヤメの品種別特徴を示した Iris データセットに対して新たな内的規準による評価で最適なクラスタ数である 3 が導かれることを確認した．

† 関西学院大学大学院
‡ 関西学院大学

人工データの生成において一部でパラメータの手動設定が必要な箇所が残っており，自動化の余地が残されている．また，現在は k-means 法のみで実験を行っているが，他のクラスタリング手法でも実験を試みる．

3. 先行研究

Milligan[2] は外的規準を元にして，良い内的規準を見つけ出すために，相関係数を用いた評価実験を行った．3 つの “design factor” によって基礎となるデータを生成し，4 つのエラー条件に従ってデータを加工し，2 つの外的規準と 30 の内的規準間の相関係数を算出している．3 つの “design factor” とは，(1) クラス数 (2-5) ，(2) 埋め込み次元 (4,6,8)，(3) クラスタごとのデータ数のばらつき（均等もしくは非均等）である．4 つのノイズ条件とは，(1) ノイズなし，(2) ノイズ付加，(3) ノイズのある次元を追加，(4) 一様分布である．

図 1 は内的規準と外的規準の相関係数を調べた結果となっている．外的規準との相関係数の高かった上位 10 を抜粋している．

Pearson and Spearman Correlations Between the External Criteria and Internal Criteria

Internal Criterion	Reference	Pearson		Spearman	
		Rand	Jaccard	Rand	Jaccard
Gamma	Baker and Hubert [1975]	.91	.77	.89	.82
C Index	Hubert and Levin [1976]	.90	.75	.89	.82
Point-Biserial	Milligan [1980]	.89	.78	.88	.82
Tau	Rohlf [1974]	.87	.74	.84	.78
W/B	McClain and Rao [1975]	.82	.74	.85	.73
G(+)	Rohlf [1974]	.78	.73	.82	.81
Tau A	Hubert and Levin [1976]	.79	.68	.77	.72
Tau C	Hubert and Levin [1976]	.79	.68	.77	.72
Somer's Gamma	Hubert and Levin [1976]	.79	.68	.77	.72
Modified Ratio of Repetition	Hubert and Levin [1976]	.75	.58	.76	.58

図 1　ラプラス分布を除いて学習した際の重みの比較

一部の内的規準は真のクラスタ構造を回復することが示された．また，外的規準と内的規準の相関によって評価しているが，データの性質によって相関が最大となる内的規準が異なり，最も良い内的規準を 1 つに定めることが困難という問題がある．また，生成している人工データの分布がガウス分布に限定されている問題もある．

4. クラスタリング結果の評価

クラスタリングは教師なし学習の 1 つであり，正解のある問題ではない．クラスタリング結果を評価することで，パラメータの調整が可能になり性能の向上が期待される．評価の種類として，外的規準と内的規準に大別される．また，最大化すべき規準と，最小化すべき規準が混在していることにも注意が必要である．

4.1 外的規準

外的規準とは，真のクラスタ構造が分かっているもとで，真のクラスラベルと予測クラスラベルを比較することでクラスタリング結果を評価する指標である．ジャッカード係数やランド指数[4] が挙げられる．ランド指数は全ての点間のペアに対して，正解クラスタと予測クラスタで状態の比較し数え上げることで評価する．状態とはペアとなる 2 点が同じくラスタに属しているか，異なるクラスタに属しているかの 2 状態のことを示す．正解クラスタと予測クラスタで状態が同じであれば良い，異なれば悪いとし，その割合によって評価する指標である．教師なし学習であるクラスタリングにおいては真のクラスタ構造は一般に与えられないため，外的規準を用いた評価はふさわしくない．

4.2 内的規準

内的規準とは，真のクラスラベルを用いず，入力データそのものと予測ラベルを用いることでクラスタリング結果を評価する指標である．クラスタ数を考慮する規準と，考慮しない規準が存在する．前者は Ball-Hall Index や Det Ratio など，後者は C-Index や Calinski Harabasz などが挙げられる[3]．Ball-Hall index はクラスタ内の点間の平均 2 乗距離の重み付き平均となっている．Calinski Harabasz はクラスタ内分散とクラスタ間分散の割合を基本とし，点数やクラスタ数によって補正した指標となっている．内的規準は正解クラスタを必要としないため，教師なし学習であるクラスタリングの評価指標としては適切である．

5. 提案手法

前述のように，クラスタリングは教師なし学習であり，正解のある問題ではない．しかし，しばしば正解ラベルを用いた外的規準が用いられている．このことに着目し，正解ラベルを使用することなく，外的規準に近い規準を構築することを試みる．

5.1 問題の定式化

入力として目標とする外的規準を与える．そのもとで，以下の式を満たす新たな規準関数 f^* を出力として得る．

$$f^* = \arg\min_{f \in F} \sum_{\hat{\mathbf{x}}} |E(\hat{\mathbf{X}}, \tilde{\mathbf{y}}) - f(\mathbf{X}, \tilde{\mathbf{y}})|^2. \qquad (1)$$

ただし，E は入力として与える外的規準，f は提案手法によって生成される内的規準であり，以下の写像と

なる．

$$E : \{(\hat{\mathbf{X}}, \tilde{\mathbf{y}})\} \quad \rightarrow \quad [0,1],$$
$$f : \{(\mathbf{X}, \tilde{\mathbf{y}})\} \quad \rightarrow \quad [0,1].$$

また，各変数は以下の意味で使用している．

$\mathbf{X}$: 人工データ,
$\mathbf{y}$: 正解クラスラベル,
$\hat{\mathbf{X}} = \{\mathbf{X}, \mathbf{y}\}$,
$\tilde{\mathbf{y}} = g(\mathbf{X}|\theta)$: 予測クラスラベル,

ここで，データ $\mathbf{X}$ は人工的に生成され，それぞれクラスタリングアルゴリズムによってクラスタリングが行われる．g はデータ $\mathbf{X}$ とクラスタ数 k を入力とし，予測クラスラベルを出力する関数であり，何らかのクラスタリング手法である．なお，クラスタ数を自動決定するクラスタリング関数も存在する．新たな規準 f^* を使用して以下の式でクラスタリングのパラメータ θ を決定する．

$$\theta^* = \arg\max_{\theta} f^*(\mathbf{X}, \tilde{\mathbf{y}}),$$
$$\tilde{\mathbf{y}} = g(\mathbf{X}|\theta).$$

5.2 内的規準を組み合わせる探索

既存の内的規準を組み合わせることで，新たな内的規準 f^* を構築する．これは，式 (1) における f の探索範囲 F を既存の内的規準の一部とすることに相当する．一例として，以下の式で与えられる M 個の内的規準の重み付き和が挙げられる．

$$f = w_0 + w_1 C_1 + w_2 C_2 + \cdots + w_M C_M,$$
$$w_i \in \mathbb{R} \qquad (i = 0, \cdots, M).$$

ここで，w_0 はバイアス項であり，$m \geqq 1$ に対して w_m は m 番目の内的規準に対する重みであり，C_m は m 番目の内的規準である．決定するパラメータにクラスタ数もしくは，クラスタ数に影響を与えるものが含まれている場合には，クラスタ数を考慮した内的規準のみを選定する必要がある．生成するデータの数を N とし，$N < M$ のもとで，以下で示される $\mathbf{w}$ の回帰問題となり，lasso[8] によって解くことができる．lasso は寄与の少ない変数の重みが 0 になりやすい特徴を持っている．

$$\sum_{\hat{\mathbf{X}}} \sum_{i=1}^{n} (w_0 + w_1 c_{1i} + \cdots + w_M c_{Mi} - e_i) + \alpha ||\mathbf{w}||_{\mathbf{1}} \rightarrow \min.$$

ここで，$\mathbf{c}_m$ は m 番目の内的規準を使用した長さ N の評価値ベクトルであり，$\mathbf{e}$ は外的規準を使用した長さ N の評価値ベクトルである．第 2 項は正則化項であり，重みが小さくなり過学習を抑制する効果が期待される．α は正則化パラメータである．なお，内的規準による評価値はあらかじめ平均が 0, 分散が 1 となるように標準化している．

6. 学習用人工データの生成モデル

人工データ $\mathbf{X}$ について、以下の確率的生成モデルを仮定する．人工データの各点は特定の分布族の無限混合の和で表現されるとする．すなわち，

$$U = \sum_{i=1}^{M} \sum_{j=1}^{\infty} \pi_{i,j} P_{i,j}. \tag{2}$$

$P_{,.}$ は確率分布であり，1 つめの添え字は M 個中 i 番目の分布族であることを表しており，2 つめの添え字はパラメータを変える走査の j 回目であることを示している．$\pi_{,.}$ は分布の混合係数である．

混合係数 π は潜在変数 $\mathbf{z}$ によって決定される．また，潜在変数 $\mathbf{z} = \{z_1, z_2, \cdots, z_M\}^{\mathbf{T}}$ は各分布属が選択されるか否かを $\{0,1\}$ で表しており，M 次元ベクトルである．潜在変数の各要素が 1 となる確率は K/M とし，K はべき乗則に基づいて確率的に決められる．べき乗則とはある量 x と y が $y = bx^a$ の関係にあることである．ここで x はクラスタ数であり，y はデータの生成確率である．K は小さいほど確率が高くなり，大きくなるほど確率が小さくなるようにする．K が小さいときは潜在変数の 1 の数が少なくなり，したがって分布の混合数が少なくなる．すなわち，クラスタ数が小さくなる．一方，K が大きくなるほどクラスタ数は大きくなり，生成確率は小さくなるものとする．

各分布属のパラメータは平均は一様分布から，その他のパラメータは各分布属に決めた値を平均に取るガウス分布から生成されるとする．すなわち，

$$\mu_i \quad = uniform(-q, q),$$
$$\theta_i \quad = N(0, r_i).$$

7. 評価実験

7.1 実験データ

本節では，実験に使用する人工データの生成方法について説明する．まず，使用する確率分布を 1 つ決める．使用する確率分布は表 1 に示す．これは Python のモジュール Numpy にあるランダムサンプリング関数から連続的な分布を選択した．次に，クラスタ数を決定しクラスタごとにクラスタ中心を決定する．クラスタ中心

は -10 から 10 の範囲で一様分布とした．各クラスタ中心を平均として，さきほど決定した確率分布からサンプリングした点を採用する．計算上，サンプルした数値が $[-10^{10}, 10^{10}]$ の範囲から外れる場合には絶対値が 10^{10} となるように調整した．確率分布のパラメータは確率分布ごとに指定した値を分散に持つガウス分布からサンプルする．各クラスタの点数は均等とし，全体で 100 となるようにした．以上で 1 つのデータが生成される．これを様々なクラスタ数・確率分布について行う．

表 1　実験データ生成に使用した確率分布

名称	パラメータの分散
gumbel	5
laplace	4
logstic	3
standard_t	10
triangular	15
uniform	10
vonmises	20
chisquare	25
f	100
gamma	50
normal	7
rayleigh	10

クラスタ数ごとに生成するデータ数はべき乗則に基づいて決める．べき乗則とはある量 x と y が $y = bx^a$ の関係にあることである．ここでは y を生成するデータ数，$x = C$ とし，クラスタ数が多くなるほど生成するデータ数が小さくなるようにする．また b は生成される全体のデータ数に依存する変数であり，今回は 1000 とする．a は逓減率であり，-2 とする．また，C はクラスタ数である．

7.2　実験データの頑健性

前節で生成した実験データの頑健性について評価を試みた．表 1 に示した確率分布を全て使用して学習した重み $\mathbf{w}$ と，そこから 1 分布除いて学習した重み $\mathbf{w}$ を比較した．図 2 は laplace 分布を除いて比較した結果である．横軸が $\mathbf{w}$ の各要素であり，縦軸が重みである．青い点が全ての分布を使用して学習した重み，黄色の点が 1 分布を除いて学習した重みである．他の分布を除いた結果は末尾の付録に掲載している．ある程度の頑健性が見られており，未知の分布に対しても対応できると考えられる．

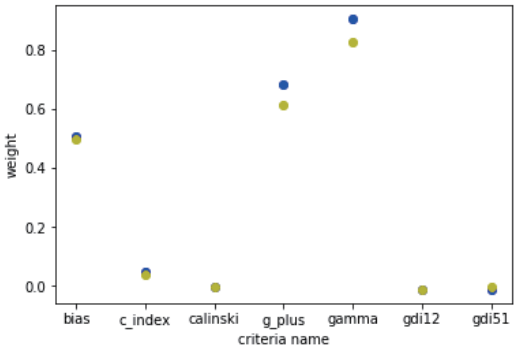


図 2　ラプラス分布を除いて学習した際の重みの比較

7.3　実験準備

実験を行うにあたり，決定すべき事項が 3 つ存在する．(1) 目標とする外的規準，(2) 使用するクラスタリングアルゴリズム，(3) 使用する内的規準である．

(1) 目標とする外的規準は最も一般的に使われていると考えられる ARI(adjusted rand index) を選択する．(2) 使用するクラスタリングアルゴリズムは k-means 法を選択する．(3) 使用する内的規準は表 2 に示す 6 規準を使用する．名称は R 言語のライブラリ “ClusterCrit”[3] に基づいている．k-means 法はクラスタ数をパラメータに持つクラスタリングアルゴリズムであるため，クラスタ数を考慮する内的規準を選択する必要がある．

表 2　クラスタ数を考慮する内的規準

名称	最適化方法
C_index	max
calinski_harabasz	min
g_plus	max
gamma	min
gdi12	min
gdi51	min

2 つの方法で評価を実施した．(1) 外的規準との相関係数による評価，(2) 実データに適用したパラメータの決定．

7.4　相関係数による評価実験

6.1 節の方法により生成した人工データセットを用いて学習を行った．具体的には，表 2 に示した 12 種類の分布から 1 つを抜きデータの生成・線形重み $\mathbf{w}$ の学習を行い，抜いた分布によるデータにより評価を行う．lasso による線形重み $\mathbf{w}$ の決定には Python のモジュール sklearn.linear_model.LassoCV を使用し，4-fold による交差検証を行った．図 3 は一様分布を除く 11 分布に

よるデータで学習を行い、一様分布によるデータでテストを行った結果である．縦軸は各内的規準値と外的規準値の相関係数の絶対値を表している．最も右の値は学習により決定された重みから生成された新たな内的規準と外的規準の相関係数である．内的規準単体より，重み付き線形和とした新たな内的規準の方が相関係数が高くなっていることが分かる．ここではピアソンの相関係数を用いた．

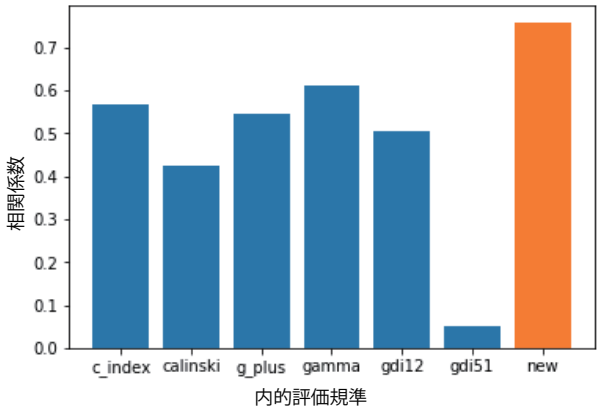


図 3　相関係数による評価

7.5　実データに対するパラメータ決定

また，実データに対してクラスタ数の決定も試みた．アヤメの品種別特徴を記録した Iris データセット [7] を使用した．学習データには 6.1 節の方法により生成した人工データセットを用いた．表 2 に示した 12 分布全てを使用した．図 4 は Iris データセットに k-means 法を適用し，本手法によって生成された内的規準による評価値を示したグラフである．横軸がクラスタ数，縦軸が提案手法による内的規準の評価値である．正しいクラスタ数である 3 で規準値が最大となり，最適なパラメータの決定が行えたことが分かる．

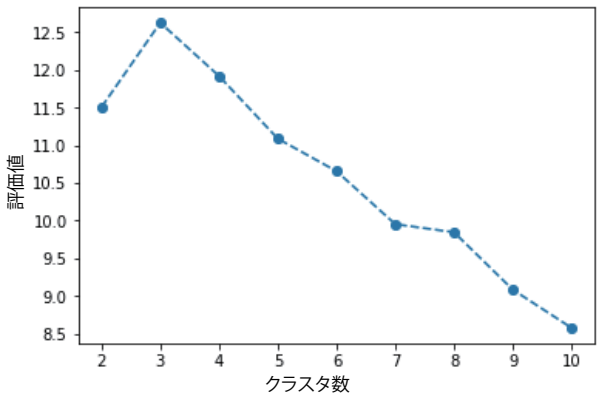


図 4　提案手法による最適パラメータの決定

8.　まとめと今後の課題

学習段階において，外的規準の値が一様になると学習がうまくいくことが経験的に分かっている．これを満たすために現在は確率分布のパラメータを手動で調整しているが，一様になるようにデータを自動的に棄却し，学習にかかる人手を減らすことを実現したい．

また，現在はクラスタ数をパラメータにもつ k-means 法で実験を行ったが，クラスタ数以外のパラメータをもつアルゴリズムに対してもパラメータの決定を試みる予定である．

また，現在は内的規準の重み付き線形和によって新たな内的規準を構築しているが，非線形な内的規準の組み合わせも試みる予定である．

さらに，今回の実験では提案した確率的生成モデルとは異なる方法により人工データの生成を行ったが，今後生成モデルに基づいた方法で人工データの生成を行う予定である．

参考文献

1) 人工知能学会全国大会　2013　大会論文集,http://2013.conf.ai-gakkai.or.jp/wp-content/uploads/2013/files/jsai2013.zip　**人工知能学会**, (2018 年 3 月 26 日確認).

2) Glenn Milligan (1981) . A Monte Carlo study of 30 internal criterion measures for cluster analysis *Psychometrika, 46(2),* 187-199.

3) Bernard Desgraupes (2013) . Clustering indices. *Bernard Desgraupes University Paris Ouest Lab Modal' X.*

4) William M. Rand (1971) . Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66 (336),* 846-850.

5) NumPy v1.14 Manual , https://docs.scipy.org/doc/numpy-1.14.0/reference/routines.random.html **SciPy.org**, (2018 年 7 月 23 日確認).

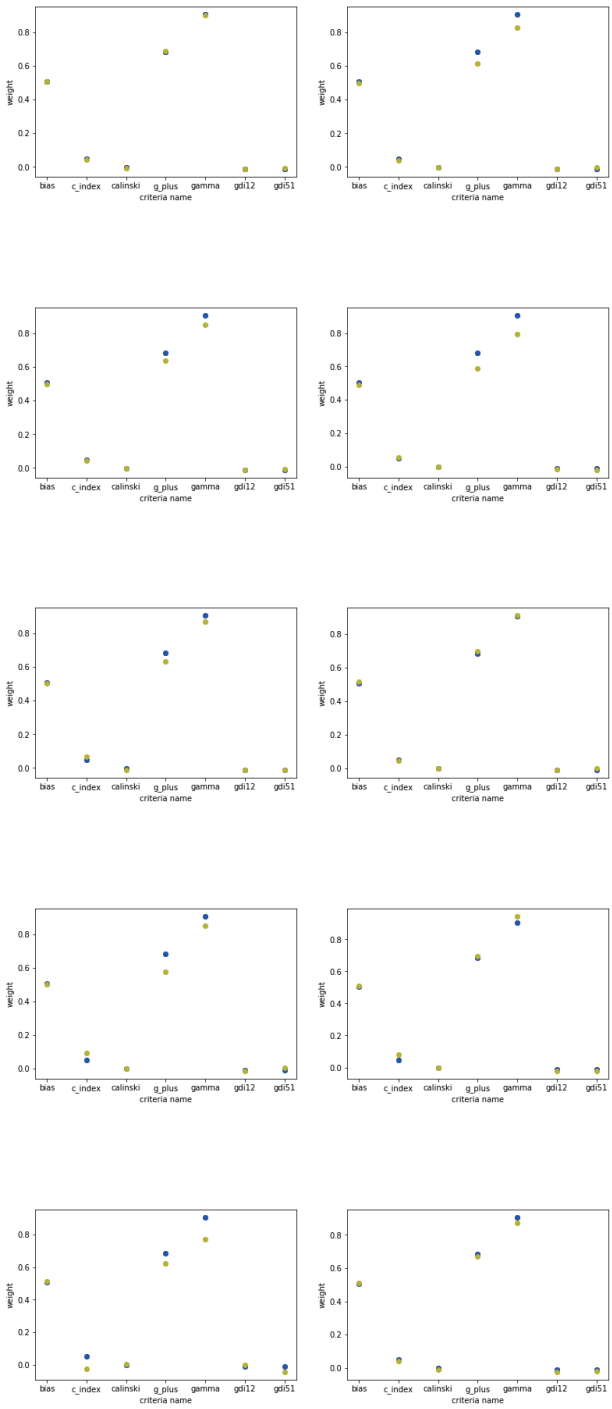
6) sklearn.linear_model.LassoCV , http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LassoCV.html **http://scikit-learn.org/**, (2018 年 7 月 23 日確認).

7) R.A. Fisher (1936). Iris Data Set. https://archive.ics.uci.edu/ml/datasets/Iris

8) C.M. Bishop (2012) . パターン認識と機械学習（上）丸善出版.

9. 付録

ここでは，7.2 節で示した重みの頑健性について，他の分布を除いて学習した際の重みの変化を図 5 に示す．横軸が $\mathbf{w}$ の各要素であり，縦軸が重みである．青い点が全ての分布を使用して学習した重み，黄色の点が 1 分布を除いて学習した重みである．左上から右下に表 1 の順で並んでいる．



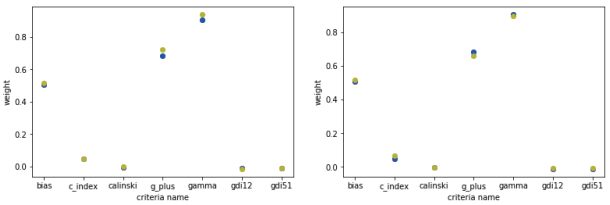


図 5　各分布を除いて学習した際の重みの比較