

多層パーセプトロンによる可読性と汎化性に優れた回帰分析に関する研究

著者	棚橋 裕輔
学位名	博士(工学)
学位授与番号	13903甲第685号
学位授与年月日	2009-03-23
URL	http://id.nii.ac.jp/1476/00002888/

博士論文

多層パーセプトロンによる
可読性と汎化性に優れた回帰分析に関する研究

2009年

棚橋 裕輔

論文要旨

いくつかの変数に基づいて別の変数の実数値を予測する手法である回帰分析は、大量のデータから有用な情報を取り出すデータマイニング技術の一つとして、様々な分野で実用化されている。回帰分析においては、可読性と汎化性が非常に重要である。可読性とは、得られた回帰式をどのように解釈可能かどうかを、汎化性とは、未知のデータに対していかに精度良い予測値を推定するかを意味する。重回帰や数量化理論一類などによる線形回帰モデルでは、回帰式が線形となるため、可読性は非常に良い。しかしながら、線形であるために、非線形の関係を持つデータに対して高い汎化性能を期待することはできない。一方、シグモイド型の多層パーセプトロンやサポートベクトルを用いた非線形回帰モデルでは、非線形であるために多くのデータに対して高い汎化性能を期待することができる。しかしながら、入出力関係がブラックボックスとなっているために、回帰式の可読性は悪く、推定した結果の解釈が難しい。

可読性と汎化性の両方に優れたモデルの一つとして多変量多項式回帰モデルがある。多変量多項式回帰モデルでは出力式が多項式として得られるため、可読性は優れている。また、非線形モデルであるため高い汎化性も期待できる。入出力データが与えられた上で多項式回帰モデルを推定する問題は、これまでにも研究されてきた。Langley が提案した BACON システムは、試行錯誤的な組合せ探索で適切な多項式を探すが、変量の数が多くなると組合せ爆発が生じる問題があった。その問題を克服するために、組合せ論的アプローチの代わりに数値的アプローチを採用した RF5 法が提案された。RF5 法では、多層パーセプトロンの学習によって組合せ爆発を回避して数値的に解を求めることが可能である。その後、量的変数と質的変数が混在するデータに対して質的条件付き多変量多項式回帰を実現する RF6.3 法も提案された。質的条件付き多変量多項式回帰では、質的変数を用いて自動的かつ適切に部分空間を構成し、各部分空間の中ではそれにフィットする多変量多項式を量的変数を用いて推定する。RF6.3 法では、3 層パーセプトロンの学習とベクトル量子化によるルール復元によってそれを実現している。

本研究では、多層パーセプトロンによる可読性と汎化性の両面に優れた回帰分析を行う 3 種の新しいモデルを提案し、その学習法を考案工夫し、有効性を計算機実験により評価した。1 つめの提案は、RF6.3 法を改良した RF6.4 法で、4 層パーセプトロンの

学習によって、RF6.3 法に比べてより柔軟に条件部を表現することが可能となる。2 つめの提案は、量的変数と質的変数の両方を含むデータに対する質的条件付き重回帰を実現する nc-MR 法である。質的条件付き重回帰法は、質的条件付き多変量多項式回帰法の各ルールでの回帰式を線形としたモデルで、より可読性を重視したモデルである。

更に、パーセプトロンの重みの取りうる値の種類を限定する、重み共有法の一つとして提案された BCW 法によって、RF5 法及び RF6.4 法の可読性と汎化性の更なる向上を目指す。3 つめの提案として、本研究ではオリジナルの BCW 法に改良を加え、RF5 法と RF6.4 法に適用した RF5+BCW 法及び RF6.4+BCW 法を提案する。

以下に本論文の構成を示す。第 1 章では、本研究の背景及び目的について述べる。第 2 章では、RF5 法の基本枠組みと学習法、モデル選択の方法、及び汎化性能向上法の 1 つである正則化法について説明する。第 3 章では、RF6.3 法の基本枠組みとその問題点、及びそれを改良した RF6.4 法について説明する。また、ルール復元法や、最適ルール数の決定方法についても説明する。第 4 章では、nc-MR 法について説明する。また、各回帰ルールの近さを視覚的に捕らえることを目的として、Spring Model を用いて回帰ルールの位置関係をグラフとして表示する方法についても説明する。第 5 章では、多層パーセプトロンの重み共有法の一つである BCW 法の基本枠組み、及び本研究における改良点を説明する。また、RF5 法及び RF6.4 法に適用した場合の処理手順について説明する。第 6 章では、人工データに対して RF5 法、RF6.4 法、nc-MR 法、及び重み共有法を適用した RF5+BCW 法、RF6.4+BCW 法を適用し、モデル選択によって正しいモデルが選択できるかどうか、また提案モデルが元の回帰式及び回帰ルール集合をうまく復元できるかどうかを検証した実験について述べる。第 7 章では、様々な現実データに対して提案モデルを適用し、現実データに対しても提案モデルが有効に働くかどうかを検証した実験について述べる。第 8 章では、本研究における結論をまとめ、今後の課題を示す。

目次

1	はじめに	1
1.1	データマイニングと回帰分析	1
1.2	パーセプトロンによる回帰分析	3
1.3	本研究の目的	3
1.4	本論文の構成	5
2	多変量多項式回帰法	7
2.1	多項式回帰法の研究について	7
2.2	3層パーセプトロンを用いた多変量多項式回帰法：RF5法	8
2.3	多層パーセプトロンの学習	10
2.4	最適中間ユニット数の決定	11
2.4.1	交差検証法	12
2.4.2	BIC	12
2.5	正則化法	14
3	質的条件付き多変量多項式回帰法	17
3.1	質的変数を含むデータ	17
3.2	3層パーセプトロンを用いた質的条件付き多変量多項式回帰法：RF6.3法	18
3.2.1	RF6.3法の基本枠組み	18
3.2.2	条件部の数値表現	19
3.3	4層パーセプトロンへの拡張：RF6.4法	20
3.3.1	RF6.3法の問題点	20
3.3.2	RF6.4法の基本枠組み	21
3.3.3	指数部分が異なる多項式群の発見	22
3.4	ルール復元法	23
3.5	モデル選択	24
4	質的条件付き重回帰法	27
4.1	重回帰分析と数量化理論一類	27
4.2	複数の関数を割り当てるモデル	28
4.3	4層パーセプトロンを用いた質的条件付き重回帰法：nc-MR法	30
4.4	回帰ルール復元における工夫	32
4.5	モデル選択	33
4.6	回帰ルールの位置関係	33
4.6.1	回帰ルールの近さ	33
4.6.2	Spring Model	35
4.7	nc-MR法の処理手順	37
5	多層パーセプトロンの重み共有	39
5.1	共有重み構造の多層パーセプトロン	39
5.2	重み双方向クラスタリング学習法：BCW法	41
5.2.1	重み共有法の基本枠組み	41
5.2.2	重みボトムアップクラスタリング	43
5.2.3	重みトップダウンクラスタリング	44
5.2.4	重み双方向クラスタリング	47

5.3 多変量多項式回帰法への適用	49
5.3.1 共有重み構造での勾配の計算	49
5.3.2 RF5 法への適用	49
5.3.3 RF6.4 法への適用	49
6 人工データを用いた実験による検証	51
6.1 多変量多項式回帰法 (RF5 法 , RF5+BCW 法)	51
6.1.1 実験における諸設定	51
6.1.2 実験結果	52
6.1.3 考察	55
6.2 質的条件付き多変量多項式回帰法 (RF6.4 法 , RF6.4+BCW 法)	56
6.2.1 実験における諸設定	56
6.2.2 実験結果	57
6.2.3 考察	64
6.3 質的条件付き重回帰法 (nc-MR 法)	65
6.3.1 実験における諸設定	65
6.3.2 実験結果	66
6.3.3 考察	69
6.4 まとめ	70
7 現実データを用いた実験による検証	71
7.1 現実データへの適用	71
7.2 ダイヤモンドの価格データ	72
7.2.1 実験における諸設定	72
7.2.2 実験結果	74
7.2.3 考察	81
7.3 プロ野球選手の年俸データ	83
7.3.1 実験における諸設定	83
7.3.2 実験結果	83
7.3.3 考察	90
7.4 様々なデータによる汎化性能の比較	91
7.4.1 実験における諸設定	91
7.4.2 実験結果	92
7.4.3 考察	100
7.5 まとめ	102
8 あとがき	105
謝辞	109
参考文献	111
研究業績	115
付録	117
A. BPQ 法における最適探索幅の計算方法	117
B. RF5 法における微分値の計算	118
C. RF6.4 法における微分値の計算	119
D. nc-MR 法における微分値の計算	120
E. Spring Model における微分値の計算	121

図目次

2.1	RF5 法を実現する 3 層パーセプトロン	10
2.2	交差検証法	13
2.3	オーバーフィットした例	14
2.4	正則化した例	14
3.1	RF6.3 法を実現する 3 層パーセプトロン	20
3.2	RF6.4 法を実現する 4 層パーセプトロン	22
4.1	重回帰分析を実現する 単層パーセプトロン	28
4.2	nc-MR 法を実現する 4 層パーセプトロン	32
4.3	式 (4.12) における各回帰ルールの位置関係	35
5.1	共有重み構造を持つパーセプトロンの例	42
5.2	BCW 法の動作の様子	48
6.1	誤差の変化の様子 (RF5+BCW 法, 人工データ, $nstd = 0.3$)	54
6.2	誤差の変化の様子 (RF6.4+BCW 法, 人工データ, $nstd = 0.3$)	60
7.1	ダイヤモンドの価格データ	73
7.2	回帰式のグラフ (重回帰分析, ダイヤモンドデータ)	78
7.3	回帰式のグラフ (RF5 法, ダイヤモンドデータ)	79
7.4	回帰式のグラフ (RF6.4 法, ダイヤモンドデータ)	79
7.5	回帰式のグラフ (MLPx, ダイヤモンドデータ)	80
7.6	回帰式のグラフ (nc-MR 法, ダイヤモンドデータ)	80
7.7	交差検証誤差の比較 (ダイヤモンドデータ)	81
7.8	各回帰ルールの位置関係 (nc-MR 法, 野球データ)	89
7.9	交差検証誤差の比較 (野球データ)	89
7.10	交差検証誤差の比較 (automobile データ)	93
7.11	交差検証誤差の比較 (b-carotene データ)	93
7.12	交差検証誤差の比較 (ball-bearing データ)	94
7.13	交差検証誤差の比較 (bb-america データ)	94
7.14	交差検証誤差の比較 (boston データ)	95
7.15	交差検証誤差の比較 (cholesterol データ)	95
7.16	交差検証誤差の比較 (college データ)	96
7.17	交差検証誤差の比較 (cpu データ)	96
7.18	交差検証誤差の比較 (fish データ)	97
7.19	交差検証誤差の比較 (iq-brain データ)	97
7.20	交差検証誤差の比較 (lung-cancer データ)	98
7.21	交差検証誤差の比較 (mpg データ)	98
7.22	交差検証誤差の比較 (philadelphia データ)	99
7.23	交差検証誤差の比較 (wage データ)	99
7.24	交差検証誤差の比較 (yokohama データ)	100

表目次

1.1 様々な回帰モデルの分類	2
6.1 BIC(J) が最小となった回数 (RF5 法, 人工データ)	52
6.2 BIC(J) の最小値 (RF5 法, 人工データ)	52
6.3 MSE _{CV} (J) の最小値 (RF5 法, 人工データ)	52
6.4 BIC(G) が最小となった回数 (RF5+BCW 法, 人工データ)	53
6.5 BIC(G) の最小値 (RF5+BCW 法, 人工データ)	53
6.6 MSE _{CV} (G) の最小値 (RF5+BCW 法, 人工データ)	53
6.7 BIC(J, R) が最小となった回数 (RF6.4 法, 人工データ, $nstd = 0.1$)	57
6.8 BIC(J, R) が最小となった回数 (RF6.4 法, 人工データ, $nstd = 0.3$)	57
6.9 BIC(J, R) が最小となった回数 (RF6.4 法, 人工データ, $nstd = 0.5$)	57
6.10 BIC(J, R) の最小値 (RF6.4 法, 人工データ, $nstd = 0.1$)	57
6.11 BIC(J, R) の最小値 (RF6.4 法, 人工データ, $nstd = 0.3$)	58
6.12 BIC(J, R) の最小値 (RF6.4 法, 人工データ, $nstd = 0.5$)	58
6.13 MSE _{CV} (J, R) の最小値 (RF6.4 法, 人工データ, $nstd = 0.1$)	58
6.14 MSE _{CV} (J, R) の最小値 (RF6.4 法, 人工データ, $nstd = 0.3$)	58
6.15 MSE _{CV} (J, R) の最小値 (RF6.4 法, 人工データ, $nstd = 0.5$)	59
6.16 BIC(G) が最小となった回数 (RF6.4+BCW 法, 人工データ)	59
6.17 BIC(G) の最小値 (RF6.4+BCW 法, 人工データ)	59
6.18 MSE _{CV} (G) の最小値 (RF6.4+BCW 法, 人工データ)	60
6.19 BIC(I) が最小となった回数 (RF6.4 法, 人工データ)	61
6.20 BIC(I) が最小となった回数 (RF6.4+BCW 法, 人工データ)	61
6.21 BIC(I) の最小値 (RF6.4 法, 人工データ)	61
6.22 BIC(I) の最小値 (RF6.4+BCW 法, 人工データ)	62
6.23 MSE _{CV} (I) の最小値 (RF6.4 法, 人工データ)	62
6.24 MSE _{CV} (I) の最小値 (RF6.4+BCW 法, 人工データ)	62
6.25 BIC(R) が最小となった回数 (nc-MR 法, 人工データ)	66
6.26 BIC(R) の最小値 (nc-MR 法, 人工データ)	66
6.27 MSE _{CV} (R) の最小値 (nc-MR 法, 人工データ)	67
6.28 BIC(I) が最小となった回数 (nc-MR 法, 人工データ)	67
6.29 BIC(I) の最小値 (nc-MR 法, 人工データ)	67
6.30 MSE _{CV} (I) の最小値 (nc-MR 法, 人工データ)	68
7.1 BIC(J) が最小となった回数 (RF5 法・MLP・nc-MR 法, ダイヤモンドデータ)	74
7.2 BIC(J) の最小値 (RF5 法・MLP・nc-MR 法, ダイヤモンドデータ)	74
7.3 MSE _{CV} (J) ($\times 10^6$) の最小値 (RF5 法・MLP・nc-MR 法, ダイヤモンドデータ)	75
7.4 BIC(J, R) が最小となった回数 (RF6.4 法, ダイヤモンドデータ)	75
7.5 BIC(J, R) の最小値 (RF6.4 法, ダイヤモンドデータ)	75
7.6 MSE _{CV} (J, R) ($\times 10^6$) の最小値 (RF6.4 法, ダイヤモンドデータ)	75

7.7	BIC(I) が最小となった回数 (RF6.4 法・HME・nc-MR 法, ダイヤモンドデータ)	76
7.8	BIC(I) の最小値 (RF6.4 法・HME・nc-MR 法, ダイヤモンドデータ)	76
7.9	MSE _{CV} (I) ($\times 10^6$) の最小値 (RF6.4 法・HME・nc-MR 法, ダイヤモンドデータ)	76
7.10	回帰ルールの条件部 (RF6.4 法, ダイヤモンドデータ)	77
7.11	回帰ルールの条件部 (nc-MR 法, ダイヤモンドデータ)	78
7.12	BIC(J) が最小となった回数 (RF5 法・MLP・nc-MR 法, 野球データ)	83
7.13	BIC(J) の最小値 (RF5 法・MLP・nc-MR 法, 野球データ)	84
7.14	MSE _{CV} (J) ($\times 10^7$) の最小値 (RF5 法・MLP・nc-MR 法, 野球データ)	84
7.15	BIC(J, R) が最小となった回数 (RF6.4 法, 野球データ)	84
7.16	BIC(J, R) の最小値 (RF6.4 法, 野球データ)	84
7.17	MSE _{CV} (J, R) ($\times 10^7$) の最小値 (RF6.4 法, 野球データ)	84
7.18	BIC(G) が最小となった回数 (RF5 法・RF6.4 法・MLP, 野球データ)	85
7.19	BIC(G) の最小値 (RF5 法・RF6.4 法・MLP, 野球データ)	85
7.20	MSE _{CV} (G) ($\times 10^7$) の最小値 (RF5 法・RF6.4 法・MLP, 野球データ)	85
7.21	BIC(I) が最小となった回数 (RF6.4+BCW 法・HME・nc-MR 法, 野球データ)	86
7.22	BIC(I) の最小値 (RF6.4+BCW 法・HME・nc-MR 法, 野球データ)	86
7.23	MSE _{CV} (I) ($\times 10^7$) の最小値 (RF6.4+BCW 法・HME・nc-MR 法, 野球データ)	86
7.24	回帰ルールの条件部 (RF6.4+BCW 法, 野球データ)	88
7.25	回帰ルールの条件部 (nc-MR 法, 野球データ)	88
7.26	実験に用いた現実データ	91
7.27	各回帰モデルにおける J^*, R^*, G^*, I^* (現実データ)	92
7.28	CPU 時間の比較 (秒)	101

第 1 章

はじめに

1.1 データマイニングと回帰分析

近年，コンピューターの普及によって社会にはあらゆる分野において大量の情報が溢れるようになった。それに伴い，大量のデータから有用な情報を取り出すデータマイニングの技術が注目されている。データマイニングとは，統計学，パターン認識，人工知能等のデータ解析の技法を用いて，データの中から価値ある情報を見つける知識発見の技術であり，データの特徴や目的などによって様々な手法が使い分けられる。その中でも，いくつかの変数に基づいて別の変数の実数值を予測する回帰分析は，その歴史も古く，データマイニング技術の中で最も重要な手法の一つであるといえる。回帰分析とは，与えられた有限のデータからその母集団における説明変数と目的変数の関係を推定する，帰納的モデリングの一種である。

しかし，一口に回帰と言っても様々なモデルがあり，手段や目的，データの種類などに応じて使い分けられる。主な回帰モデルの出力の形と説明変数の種類による分類を表 1.1 に示す。

最も単純な回帰モデルの一つである重回帰では，目的変数が説明変数に対する線形関数で表現されると仮定して，その線形関数を与えられたデータに基づいて予測する。重回帰では量的説明変数のみを用いるのに対して，説明変数に質的変数も含まれる場合は数量化理論一類と呼ばれる。また，通常回帰分析では目的変数は量的変数であり，その分布が正規分布に従うものと仮定することが多いが，目的変数が質的変数である場合には，ベルヌーイ分布に従うと仮定したロジスティック回帰モデルを適用する。これらの線形回帰モデルは，誤差分布を指數型分布族に拡張したモデルの総称として，一般化線形モデル^[1]と呼ばれる。

上述の線形回帰モデルに対して，非線形回帰モデルでは，説明変数と目的変

表 1.1: 様々な回帰モデルの分類

出力の形	説明変数	
	量的変数のみ	量的・質的変数の両方
線形	重回帰	数量化理論一類
		質的条件付き重回帰
多項式	多変量多項式回帰	質的条件付き 多変量多項式回帰
その他の非線形	サポートベクトル回帰	ニューラルネット回帰
木構造	回帰木	
その他	自己回帰	KNN
	ロジスティック回帰	加法回帰

数の関係をより柔軟に表現することができる。マージン最大化という規範から導出されるサポートベクトル回帰[2]や多層パーセプトロンによる回帰(ニューラルネット回帰)は非線形回帰モデルの一種である。また、多変量多項式回帰や質的条件付き多変量多項式回帰も非線形回帰モデルである。多変量多項式回帰についての詳細は後に述べる。

これらの線形回帰モデル、非線形回帰モデルとは別に、木構造の出力を得る回帰木モデルがある。回帰木モデルでは、入力空間を階層的に分割して木構造で表し、その葉の部分に出力値を格納する。葉の部分に各領域内の近似値を单一の値として格納するCART[3]や、各領域に線形回帰式を割り当てるよう拡張した線形回帰木[4]などが提案されている。

線形回帰モデル、非線形回帰モデル、回帰木モデル以外にも、様々なモデルがある。加法回帰モデルは、いくつかの説明変数についてはノンパラメトリック関数を、他の説明変数については線形関数をあてはめ、それらの加法性を仮定したモデルである。従って、加法回帰モデルは、線形回帰モデルと非線形回帰モデルとの中間的な位置付けのモデルであるといえる。また、一般化線形モデルを加法モデル化した一般化加法モデル[5]も提案されている。その他、目的変数の分布の分散も説明変数に依存して変化すると仮定した分散変動モデル、時系列性を持つデータに対する自己回帰モデル、特徴空間における最も近い訓練サンプルに基づいて、局所的な関数近似を行う k 近傍法(KNN)による回帰モデル[6]などがある。

1.2 パーセプトロンによる回帰分析

人間を含めた動物の脳を構成する最も基本的な素子であるニューロンの生物学的振舞いが明らかになるに連れ、その情報処理的側面を単純化したニューロンモデルが、McCulloch と Pitts によって提案された。その後、数十から数百のニューロンが繋がってできる一つのネットワークを考え、その情報処理的性質や数理的性質が論じられた。それがニューラルネットである。

パーセプトロン [7] の拡張としての多層パーセプトロンは階層型ニューラルネットの一つであり、Rumelhart らが誤差逆伝搬学習法 [8] と呼ばれるネットワークのパラメータを推定するためのアルゴリズムを提案して以来、パターン認識や制御などの様々な問題に適用された [9]。また、多層パーセプトロンにおけるネットワークの能力や学習アルゴリズムの高速化、多変量データ解析との関係、汎化能力の高いネットワークを構成するための方法などが研究された [10]。そして、回帰分析に多層パーセプトロンを適用することにより、高い汎化性能を得られることが示された。

1.3 本研究の目的

回帰分析を行う上で重要な要素として、計算量、回帰式の可読性、回帰式の汎化性が挙げられる。回帰式の可読性とは、得られた回帰式をどのように解釈可能かどうかを、汎化性とは、未知のデータに対していかに精度良い予測値を推定するかを意味する。もちろん、計算量は少ない方が良いが、それよりも可読性と汎化性の方がより重要である。なぜならば、回帰分析においてはリアルタイム処理が必要となる場合はほとんどないと思われるからである。リアルタイム処理が必要な場合があるとしても、それは推定した回帰式を用いて、新たなデータに対する予測値を計算する場合のみであり、回帰式の推定にある程度の計算量が必要であっても、実用上は問題ない場合が多い。それよりも、いかに精度良く予測値を推定し、得られた回帰式をどのように解釈してデータを解析するかということが重要である。

線形回帰モデルである重回帰や数量化理論一類では、回帰式が線形であるためその可読性は非常に良いが、線形であるために、非線形の関係を持つデータに対して高い汎化性能を期待することはできない。一方、サポートベクトル回帰やニューラルネット回帰などの非線形モデルでは、非線形であるために多くのデータに対して高い汎化性能を期待することができるが、入出力関係がブラックボッ

クスとなっているために、回帰式の可読性は悪く、推定した結果の解釈が難しい。このように、回帰分析においては、可読性と汎化性はトレードオフの関係にある場合が多い。

多変量多項式回帰、回帰木、加法回帰は他の回帰モデルと比較して可読性と汎化性の両面に優れたモデルであるといえる。本研究では、この中でも特に、人間に馴染み深い多項式という形で出力を得られる多変量多項式回帰に着目した。このモデルでは、出力式が多項式として得られるため可読性に優れており、また、非線形であるため高い汎化性も期待できる。入出力データが与えられたもとで多項式回帰モデルを推定する問題は、古くから研究されてきた。その先駆的な研究として提案された BACON システム [11] では、試行錯誤的な組合せ探索で適切な多項式を探す。しかし、変量の数が多くなると組合せ爆発が生じる問題があった。その問題を克服するために、組合せ論的アプローチの代わりに数値的アプローチを採用した RF5 法 [12] が提案された。RF5 法では、多層パーセプトロンの学習によって、組合せ爆発を回避して数値的に解を求めることが可能である。その後、量的変数と質的変数が混在するデータに対して質的条件付き多変量多項式回帰を実現する RF6.3 法 [13] も提案された。質的条件付き多変量多項式回帰では、質的変数を用いて自動的かつ適切に部分空間を構成し、各部分空間の中ではそれぞれにフィットする多変量多項式を量的変数を用いて推定する。RF6.3 法でも、この問題を解くために 3 層パーセプトロンの学習を行う。

しかし、RF6.3 法の研究を進めたところ、RF6.3 法ではパーセプトロンの構造上、回帰ルールの条件部の表現能力に限界があることが分かった。そこで、本研究の 1 つめの提案モデルとして、RF6.3 法の 3 層パーセプトロンを 4 層に拡張した RF6.4 法を提案する。RF6.4 法では、パーセプトロンの質的変数部分の層を増やすことにより、条件部の表現能力が強化され、それによって汎化性の向上が期待できる。

次に、新たなモデルとして、質的条件付き重回帰モデルについて考える。このモデルでは、質的条件付き多変量多項式回帰モデルにおいて、各部分空間にフィットする回帰式を線形に制限することによって、汎化性よりも可読性をより重視した出力を得ることを目的とする。本研究では、この質的条件付き重回帰を 4 層パーセプトロンの学習によって実現する、nc-MR 法を提案する。数量化理論一類は、説明変数に量的変数と質的変数の両方を用いる線形モデルの 1 つであり、各ルールの回帰式が質的変数の値によって定数部分の値のみ異なるモデルとなっている。その一方、質的変数の値の全ての組合せごとに個別な部分空間を考え、それぞれ

の中で量的変数を用いて通常の重回帰を行うといった線形モデルも考えられる。nc-MR 法は、これら 2 つのモデルの中間的な位置付けのモデルとなっている。

更に、多変量多項式回帰モデルの可読性と汎化性を向上させるための方法として、パーセプトロンの重みの取りうる値の種類を限定する重み共有法に着目した。多変量多項式回帰モデルの欠点の 1 つとして、変数の種類が多い場合に、予測に不要なパラメータが多くなり、多項式が非常に複雑になってしまうことが挙げられる。しかし、重み共有によって、不要なパラメータを楽に枝刈りでき、多項式の可読性の向上を期待できる。共有重み構造のパーセプトロンを学習する方法としては、BCW 法[14] が提案されている。BCW 法では、クラスタの併合と分割を双方に繰り返すことにより、大域的な解を探索する。本研究では、この BCW 法の初期クラスタの決定方法を変え、新たなクラスタの分割方法を用いることによって、より効率良く共有重み構造を学習できるように改良した。この改良した BCW 法を RF5 法と RF6.4 法に適用した RF5+BCW 法と RF6.4+BCW 法を提案する。

本研究では、可読性と汎化性に優れた新しい回帰モデルとして以上の 3 種を提案し、その学習法を考案工夫する。また、人工データと現実データを用いた実験により、有効性を評価する。

1.4 本論文の構成

まず 1 章では、本研究の背景及び目的について述べた。

2 章では、説明変数として量的変数のみを用いる、多変量多項式回帰法について述べ、それを実現するモデルの 1 つとして提案された RF5 法について説明する。RF5 法では、3 層パーセプトロンの学習によって、数値的アプローチでデータにフィットする多変量多項式を推定する。その詳細と、交差検証法及び BIC[15] による最適中間ユニット数の選択について詳しく説明し、また、汎化性能向上法の 1 つである正則化法について説明する。

3 章では、説明変数として量的変数と質的変数の両方を用いる質的条件付き多変量多項式回帰法について述べ、それを実現するモデルの 1 つとして提案された RF6.3 法について説明する。また、RF6.3 法の問題点、及びその問題点を克服するために提案した RF6.4 法について説明する。RF6.3 法及び RF6.4 法では、質的条件付き多変量多項式回帰問題を、3 層または 4 層のパーセプトロンの学習と係数ベクトルの量子化によるルール復元によって解く。その詳細について述べる。

4 章では、説明変数として量的変数と質的変数の両方を用いる質的条件付き重

回帰法について述べ，それを実現するモデルの1つとして提案した nc-MR 法について説明する．また，各回帰ルールの位置関係をグラフとして表示し，各回帰ルールの位置関係を視覚的に解釈するための方法についても述べる．

5章では，多層パーセプトロンの重み共有について述べ，共有重み構造パーセプトロンの学習法として提案された BCW 法について説明する．本研究で BCW 法に加えた改良について説明し，それを RF5 法と RF6.4 法に適用した RF5+BCW 法，RF6.4+BCW 法の処理手順を説明する．

6章では，人工データを用いた実験によって，提案モデルが元の回帰式または回帰ルール集合を正しく復元できるかどうかを検証する．また，交差検証法と BIC によって，正しい中間ユニット数，重み共有のクラスタ数，及びルール数を選択できるかどうかを検証する．

7章では，様々な現実データを用いた実験によって，提案モデルが現実のデータに対しても高い可読性と汎化性を示せるかどうかを検証する．特に，提案モデルを用いて，ダイヤモンドの価格データとプロ野球選手の年俸データについて詳しく解析する．また，15種類のデータを用いた実験によって，他の既存モデルと提案モデルの汎化性能を比較する．

最後に8章で，本研究成果のまとめと将来の課題について述べる．

第 2 章

多変量多項式回帰法

2.1 多項式回帰法の研究について

これまで線形回帰や非線形回帰に関して多くの研究がなされてきたものの、多変量データから多項式を発見する研究は十分に確立したとはいえないかった。昔から一般的に用いられてきた多項式回帰というのは、説明変数が単変量であり、かつ指数部分が整数に限定されたもので、出力は以下のような形で表される。

$$y = w_0 + w_1x^1 + w_2x^2 + \cdots + w_Jx^J = \sum_{j=0}^J w_jx^j \quad (2.1)$$

ただし、 w_j は多項式のそれぞれの項の係数値を表す。このような形の多項式をデータから発見するのは容易で、説明変数 x を 2 乗、3 乗、… したものをおも用意して入力データとして用いれば、係数値パラメータ w_j に関しては線形となっているので、単に重回帰分析を適用するだけで最適な係数値 w_j を決定することが可能である。ただし、この場合には多項式の項数 J をなんらかの方法によって決定する必要があり、これをモデル選択という。モデル選択についての詳細は後に述べる。

上記の方法は、全データにフィットする多項式をただ 1 つだけ求めるための方法であるが、データを説明変数の値によっていくつかの区間に分けて、それぞれの区間にフィットする多項式を求める、区分的多項式回帰についても古くから研究されてきた。その方法としては、スプライン補間 [16] が有名である。スプライン補間では、区間ごとの多項式曲線を、区間の境界で滑らかに接続する。

データから多項式を発見した実例として有名なものに、ケプラーの第 3 法則がある。ケプラーの第 3 法則は、惑星と太陽の距離 r と惑星の公転周期 T との関係を表すもので、ケプラーは観測データから手計算によって $T = kr^{3/2}$ という関係を発見した。ただし、 k は定数を表す。これは説明変数として r のみを含む单変量の問題であるが、指数が $3/2$ と、整数になっていない。従って、重回帰分析のような

線形回帰ではこのような式を発見することはできない。また、もしデータに他の不要変数が含まれていれば、データは多変量となり、上記の方法は適用できない。

多変量の多項式回帰法については、BACONシステム [11] での先駆的な研究の後、同様のアイデアに基づきさまざまな方法が提案されたが、それらの基本的な探索戦略はほとんど同じである。その探索戦略は、乗算、除算、または、あらかじめ定義した関数を用いて、2つの既存変数を組み合わせ、新たな変数を再帰的に生成するというものである。しかし、これらの既存法には共通に以下の問題を指摘できる。まず第1に、2つの変数を順番に組み合わせて新たな変数を作るので、多くの変数から成るデータにおいて複雑な法則を探索すれば、容易に組合せ爆発が起き、また、探索パラメータが適切でなければ、望ましい回帰式を発見できないことが予想される。第2に、回帰式に現れる指数の値が、ケプラーの第3法則のように整数でないとき、適当な関数 (e.g., $r^{3/2}$) をあらかじめ定義しなければ、回帰式の発見が困難になる。しかし、多くの場合、それを可能とする事前知識はない。そして第3に、現実の観測データはノイズを含むが、既存法は比較的ノイズに弱いということである [17]。

これらの問題を解決するために、組合せ論的アプローチの代わりに数値的アプローチを採用した RF5 法 (Rule extraction method from Facts version 5) [12] が提案された。RF5 法では、3層パーセプトロンの学習によって、組合せ爆発を回避して数値的に解を求めることができるのである。また、指数部分が整数でない多項式を求めることも可能であり、ケプラーの第3法則のような多項式も、事前知識無しで容易に発見することが可能である。

次節では、数値変数だけから成る多変量データから多変量多項式を求める RF5 法について説明する。

2.2 3層パーセプトロンを用いた多変量多項式回帰法：RF5 法

3層パーセプトロンを用いた多変量多項式回帰法である RF5 法を定式化する。いま、 K 個の説明変数を $x = (x_1, x_2, \dots, x_K)$ 、目的変数を y として、以下の多変量多項式回帰モデルを考える。

$$y = f(\mathbf{x}; \boldsymbol{\theta}) + \epsilon, \quad f(\mathbf{x}; \boldsymbol{\theta}) = w_0 + \sum_{j=1}^J w_j \prod_{k=1}^K x_k^{w_{jk}} \quad (2.2)$$

ただし、定数項 w_0 、係数 w_j や指数 w_{jk} は未知の実数パラメータ、項数 J は未知の整数パラメータ、 ϵ は誤差項である。このとき、データから求めるべき数法則は項

数が未知で指数が実数の多項式である。対象とする法則が周期関数や不連続関数からなる場合、式(2.2)では厳密に対処できない。しかし、このような関数でも、入力ベクトル x のレンジが限られている場合には、有限項数の多項式を用いて、ある程度の精度で近似可能であり、さらに、実数指数の多項式ならば、各項の表現能力が向上するので、少ない項数での近似が期待できる。

いま、実数パラメータ群 $w_0, \{w_j\}, \{w_{jk}\}$ をまとめて単に θ と書く。パラメータ θ の次元を $M (= KJ + J + 1)$ とする。いま、 $x_k > 0$ を仮定すると、式(2.2)の $f(x; \theta)$ は以下のように変形できる。まず、 $\prod_{k=1}^K x_k^{w_{jk}}$ に対して対数をとり、その上に指数をとる形にする。

$$f(\mathbf{x}; \boldsymbol{\theta}) = w_0 + \sum_{j=1}^J w_j \exp \left(\ln \left(\prod_{k=1}^K x_k^{w_{jk}} \right) \right) \quad (2.3)$$

さらにこれを次のように変形する。

$$f(\mathbf{x}; \boldsymbol{\theta}) = w_0 + \sum_{j=1}^J w_j \exp \left(\sum_{k=1}^K w_{jk} \ln x_k \right) \quad (2.4)$$

これは、図 2.1 に示す 3 層パーセプトロンの前向き伝播にほかならない。ただし、隠れユニットにはバイアスはなく、活性化関数は指数関数であるので、隠れユニットの出力は以下となる。

$$m_j = \exp \left(\sum_{k=1}^K w_{jk} \ln x_k \right) \quad (2.5)$$

出力ユニットは線形で、バイアスを持つので、その出力は以下となる。

$$f(\mathbf{x}; \boldsymbol{\theta}) = w_0 + \sum_{j=1}^J w_j m_j \quad (2.6)$$

このような 3 層パーセプトロンを多項式ネットと仮称する。なお、中間ユニット数は多項式の定数項を除いた項数 J に対応することがわかる。 J の最適な値は未知のため、なんらかの方法を用いて決定する必要がある。これについては後に詳細を述べる。

いま、データとして N 個のサンプル $\{(x^\mu, y^\mu) : \mu = 1, \dots, N\}$ が与えられているとする。最適な J をなんらかのモデル選択法によって決定すれば、データを最も良く説明するパラメータベクトル θ は以下の 2 乗和誤差を最小にするものとして決まる。これは 3 層パーセプトロンの学習問題である。

$$E(\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{2} \sum_{\mu=1}^N \left(f(\mathbf{x}^\mu; \boldsymbol{\theta}) - y^\mu \right)^2 \quad (2.7)$$

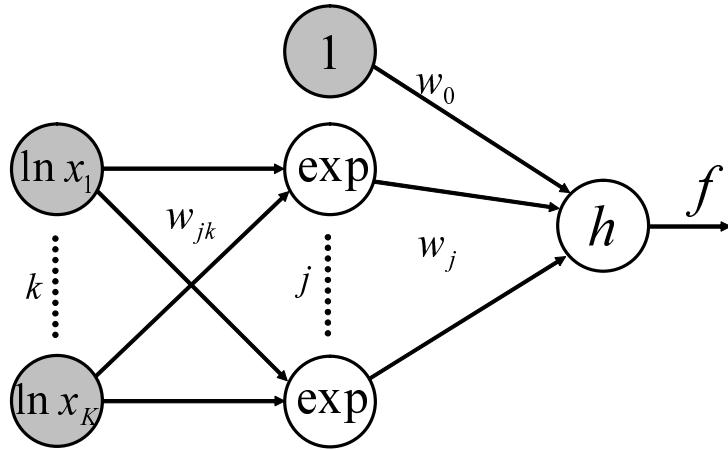


図 2.1: RF5 法を実現する 3 層パーセプトロン

2.3 多層パーセプトロンの学習

式 (2.2) で表される多項式は表現能力が高く、非線形なデータにも柔軟に対応できるが、多くの局所解を含むため、そのパラメータの学習は比較的困難である。従って、標準的な最急降下法では望ましい結果を得ることはできず、1次収束であるため、学習が収束するのも遅い。そこで、つねに効率良く望ましい結果を得るために、RF5 法では、2次の学習アルゴリズム BPQ 法 (Back Propagation based on Quasi-Newton)[18] を用いて学習を行う。BPQ 法は準ニュートン法を基本枠組みとし、最適ステップ幅を2次近似の最小点として計算する。準ニュートン法は、探索の過程で反復により2次微分の逆行列 $(\nabla^2 E(x; \theta))^{-1}$ の近似値 (H) を各ステップで求めることを特徴とする。準ニュートン法の基本アルゴリズムは以下である。

step1 初期化

θ_1 を初期化し、 $H_1 = I$, $k = 1$ とする。

step2 探索方向の計算

探索方向 $\Delta\theta_k = -H_k \nabla E(x; \theta_k)$ を計算する。

停止条件を満たせば、反復を終了させる。

step3 最適探索幅の計算

$E(\mathbf{x}; \boldsymbol{\theta}_k + \alpha_k \Delta \boldsymbol{\theta}_k)$ を最小にする α_k を求める .

step4 結合重みの修正

結合重みを $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha_k \Delta \boldsymbol{\theta}_k$ に修正する .

step5 \mathbf{H} の更新

$k \equiv 0 \pmod{M}$, (M :全パラメータ数) ならば $\mathbf{H}_{k+1} = \mathbf{I}$ とし, さもなければ, \mathbf{H}_{k+1} を更新する .

$k \leftarrow k + 1$ として step2 へ戻る .

step5 での \mathbf{H}_{k+1} の計算法にはいくつか方法があるが, 制約のない非線形最適化では, BFGS(Broydon-Fletcher-Goldfarb-Shanno) 公式が最も優れた方法であると指摘される. BFGS 公式は以下で与えられる .

$$\begin{aligned}\mathbf{H}_{k+1} &= \mathbf{H}_k + \left(1 + \frac{\mathbf{q}^T \mathbf{H}_k \mathbf{q}}{\mathbf{p}^T \mathbf{q}}\right) \frac{\mathbf{p} \mathbf{p}^T}{\mathbf{p}^T \mathbf{q}} - \frac{\mathbf{p} \mathbf{q}^T \mathbf{H}_k + \mathbf{H}_k \mathbf{q} \mathbf{p}^T}{\mathbf{p}^T \mathbf{q}}, \\ \mathbf{p} &= \lambda_k \Delta \boldsymbol{\theta}_k, \\ \mathbf{q} &= \nabla E(\mathbf{x}; \boldsymbol{\theta}_{k+1}) - \nabla E(\mathbf{x}; \boldsymbol{\theta}_k)\end{aligned}\tag{2.8}$$

なお, 最適探索幅の計算方法および勾配の計算式については, 付録に掲載した .

多層パーセプトロンの学習には, BPQ 法以外の方法を用いることも考えられる. 準ニュートン法は, ヘッセ行列を求めるための目的関数の 2 階微分の計算を避けるために, ヘッセ行列の逆行列を近似的に推定しながらニュートン法を行う方法であるが, それ以外にも, 最小二乗法の目的関数におけるヘッセ行列を近似的に求めてニュートン法を行うガウス・ニュートン法や, その拡張であるレーベンバーグ・マーカート法, 逆行列の計算を避ける代表的な手法である共役勾配法などがある [19] . これらの方法を用いて学習をすることも可能であるが, 収束の速度や性能に大差はないので, 本研究では主に BPQ 法を用いることにする .

2.4 最適中間ユニット数の決定

式 (2.2) における多項式の項数は J であるが, これは図 2.1 の 3 層パーセプトロンの中間ユニット数に相当する. 一般に, 与えられたデータ集合に対して, 最適な中間ユニット数をあらかじめ知ることはできないので, 中間ユニット数を変えて発見した候補を適切に評価するための尺度が必要となる. なお, 中間ユニット数が増えるほどパーセプトロンの表現能力は向上するので, 式 (2.7) で表される訓

練誤差は小さくなり、データに含まれるノイズにオーバーフィットしてしまうので、式(2.7)を中間ユニット数を決定するための尺度にすることはできない。本研究では、中間ユニット数を決定するためのモデル評価尺度として、交差検証法またはBICを用いる。

2.4.1 交差検証法

予測の観点によるモデル評価とは、観測データに基づいて1つのモデルを構築したとき、そのモデルの良さをモデル構築に用いたデータとは独立に得られたデータ（テストデータ）によって評価することである。しかし現実には、別個にテストデータが得られる状況は想定しがたく、また、実際に得られる場合には、そのデータもモデル構築に用いる方が良いモデルが求まる。そこで、予測の観点からの評価を観測データのみに基づいて実行しながら、なるべくパラメータ推定の精度を上げるための工夫を施した方法が交差検証法(cross-validation)である。交差検証法はニューラルネットなどの学習機械の汎化性能を評価するのにしばしば利用される[9]。

S -fold 交差検証法では、与えられたデータ D をランダムに S 個のセグメントに分割する ($D_s, s = 1, \dots, S$)。そして、 $S - 1$ 個のセグメントを学習に使用し、残りの 1 個をテスト用としてその誤差を求める。この手順を S 回繰り返して、最終的に式(2.9)の平均2乗誤差 MSE_{CV} を求める。

$$MSE_{CV} = \frac{1}{N} \sum_{s=1}^S \sum_{\mu \in D_s} \left(f(\mathbf{x}^\mu; \hat{\boldsymbol{\theta}}_s) - y^\mu \right)^2 \quad (2.9)$$

ただし、 $\hat{\boldsymbol{\theta}}_s$ は $S - 1$ 個のセグメントで学習されたときの最適な重みである。モデル選択を行う場合、この MSE_{CV} を最小とするものをモデルとして選択する。交差検証法は、特にセグメント数とサンプル数が等しい場合 ($S = N$) は、一つ抜き法(Leave one out)と呼ばれ、少数のサンプルの問題に対してよく使用される。

2.4.2 BIC

交差検証法は非常に有用な方法であるが、モデルの評価にはセグメントの分割ごとに学習が必要となり、計算量が増大するという問題もある。そこで、交差検証法のような繰返しの学習を必要としないモデル選択法として BIC(Bayesian Information Criterion)[15] を適用することを考える。BIC では、ある J の値に対して 1 回学習をすれば済むので、少ない計算量で最適中間ユニット数 J^* を決定でき

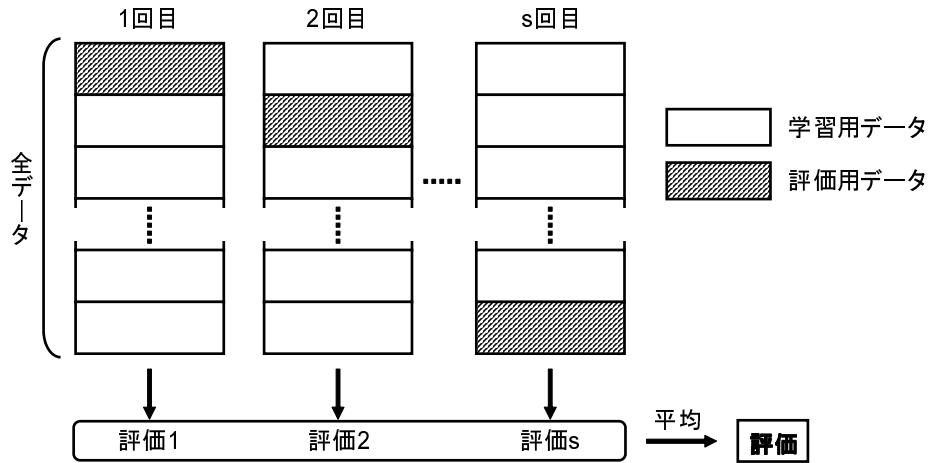


図 2.2: 交差検証法

る。BIC は以下の式で計算される。

$$\text{BIC} = -L(\hat{\boldsymbol{\theta}}) + \frac{M}{2} \log N \quad (2.10)$$

ただし、 $L(\hat{\boldsymbol{\theta}})$ は最大対数尤度、 M は総パラメータ数、 N は総サンプル数を表す。

いま、目的変数の誤差 ϵ が平均 0、標準偏差 σ の正規分布に従うとする。

$$p(y|\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(f(\mathbf{x}; \boldsymbol{\theta}) - y)^2}{2\sigma^2}\right) \quad (2.11)$$

学習によって得られたパラメータベクトルを $\hat{\boldsymbol{\theta}}$ とすると、最大対数尤度は以下となる。

$$\begin{aligned} L(\hat{\boldsymbol{\theta}}) &= \log \prod_{\mu=1}^N p(y^\mu | \mathbf{x}^\mu; \hat{\boldsymbol{\theta}}) \\ &= \sum_{\mu=1}^N \log p(y^\mu | \mathbf{x}^\mu; \hat{\boldsymbol{\theta}}) \\ &= -\sum_{\mu=1}^N \frac{(f(\mathbf{x}^\mu; \hat{\boldsymbol{\theta}}) - y^\mu)^2}{2\sigma^2} - N \log(\sqrt{2\pi}\sigma) \\ &= -\frac{1}{\sigma^2} \sum_{\mu=1}^N \frac{(f(\mathbf{x}^\mu; \hat{\boldsymbol{\theta}}) - y^\mu)^2}{2} - \frac{N}{2} \log(\sigma^2) - N \log(\sqrt{2\pi}) \end{aligned} \quad (2.12)$$

よって、式(2.12)を最大化する σ^2 を求めるとななる。

$$\begin{aligned} \frac{\partial L(\hat{\theta})}{\partial \sigma^2} &= \frac{1}{(\sigma^2)^2} \sum_{\mu=1}^N \frac{(f(\mathbf{x}^\mu; \hat{\theta}) - y^\mu)^2}{2} - \frac{N}{2\sigma^2} = 0 \\ \iff \hat{\sigma}^2 &= \frac{1}{N} \sum_{\mu=1}^N (f(\mathbf{x}^\mu; \hat{\theta}) - y^\mu)^2 \end{aligned} \quad (2.13)$$

これを式(2.12)に代入することによって $L(\hat{\theta})$ を計算できる。モデル選択に無関係の項を除外すると、BIC(J)は以下となる。

$$\text{BIC}(J) = \frac{N}{2} \log \left(\frac{1}{N} \sum_{\mu=1}^N (f(\mathbf{x}^\mu; \hat{\theta}_J) - y^\mu)^2 \right) + \frac{M}{2} \log N \quad (2.14)$$

$J = 1, 2, \dots$ と変えてパーセプトロンの学習を行い、BIC(J)を最小にする J を J^* として選択する。ただし、 $\hat{\theta}_J$ は J 個の中間ユニットのパーセプトロンの学習後に得られたパラメータベクトルを表す。

2.5 正則化法

何の制約も課さないままで、多変量多項式回帰のような強力な学習モデルを学習させると、図2.3のように訓練データのノイズに過剰に適合してしまい、汎化誤差の増大を招く。

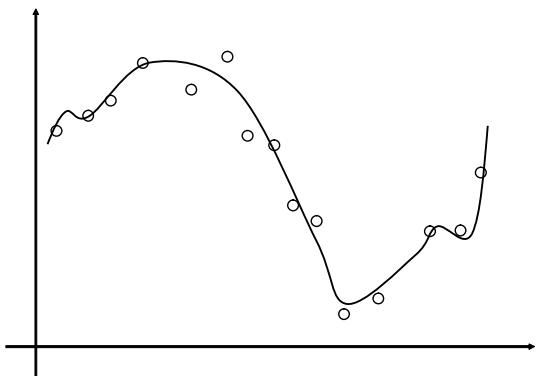


図 2.3: オーバーフィットした例

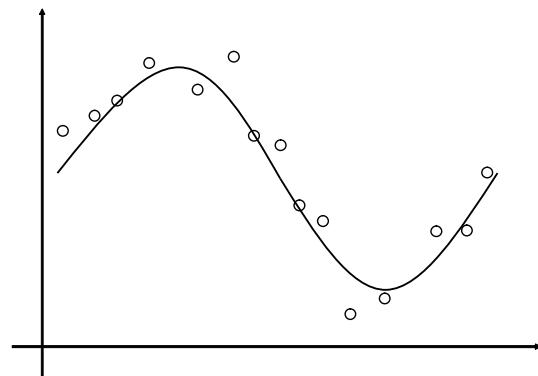


図 2.4: 正則化した例

そこで、全く自由に学習させるのではなく、ある種の拘束を与えて学習させ、汎化性能の向上を図る。その方法の1つが正則化である。正則化では、各重みの絶対値をできるだけ小さくする制約を加えるために、目的関数に正則化項(ペナルティ項)を加える。特に、重みの平方和を正則化項とする方法を重み減衰法、ある

いはガウス正則化と呼ぶ。重み減衰法では2乗和誤差 E に正則化項を加えた $\mathcal{E}(x; \theta)$ を新たな目的関数として学習を行う。

$$\mathcal{E}(x; \theta) = E(x; \theta) + \lambda \left(\frac{1}{2} \sum_m w_m^2 \right) \quad (2.15)$$

ここで、 λ は非負で、正則化係数と呼ばれる。 λ を大きくすると、2乗和誤差に対して正則化項の割合が大きくなり、各重みは 0 に近づく。逆に、 λ を小さくしすぎると、2乗和誤差に比して正則化項が効かなくなる。適度な大きさの λ において、汎化性能の改善が得られる。

一般に、重みには重要な働きをするものと不要なものがあるが、重み減衰法では、どの重みにも同じ正則化係数 λ を乗じて小さくしようとする。本来は、重要な重みの正則化係数は小さくして自由度を与え、不要な重みの正則化係数は大きくして 0 に近づけるのが望ましい。そこで、各重みごとに個別の正則化係数を用意する方法が考えられる。この場合の目的関数は以下となる。

$$J(\theta) = E(\theta) + \frac{1}{2} \theta^T \Lambda \theta \quad (2.16)$$

ここで、 M 次元ベクトルを $\lambda = (\lambda_1, \dots, \lambda_M)^T$ で定義し、 M 次元対角行列 Λ を次のように定義する。

$$\Lambda = \begin{pmatrix} \exp(\lambda_1) & & & \\ & \ddots & & 0 \\ & 0 & \ddots & \\ & & & \exp(\lambda_M) \end{pmatrix} \quad (2.17)$$

Λ の対角要素として λ の代わりに $\exp(\lambda)$ を用いるのは、ペナルティ係数は非負であり、常に正の数とするためである。

正則化係数は、パラメータを学習する環境を規定するので、一般に、ハイパー・パラメータと呼ばれる。ハイパー・パラメータの値は、学習結果の汎化性能に重大な影響を与える。上記の目的関数における個別の正則化係数の値を設定する方法としては、交差検証誤差を最小にするように正則化係数を学習する MCV(Minimum Cross-Validation error) 正則化法 [20] が提案された。MCV 正則化の学習アルゴリズムを以下に示す。

step1 $\theta^{(0)}$ と $\lambda^{(0)}$ を初期化し、 $r = 1$ とする。

step2 式 (2.16) を最小化する重み $\hat{\theta}^{(r)}$ を求める。

step3 式(2.18)を最小化する重み $\hat{\theta}_s^{(r)}$ を求める。

$$J_s(\boldsymbol{\theta}_s) = \frac{1}{2} \sum_{\mu \notin D_s} (y^\mu - f(\mathbf{x}^\mu; \hat{\boldsymbol{\theta}}_s))^2 + \frac{1}{2} \boldsymbol{\theta}_s^T \boldsymbol{\Lambda} \boldsymbol{\theta}_s \quad (2.18)$$

step4 式(2.19)を用いて探索方向 $\Delta \lambda$ を求める。

$$\frac{\partial \text{MSE}_{\text{cv}}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = -\frac{2}{N} \sum_{s=1}^S \sum_{\nu \in D_s} (y^\nu - f(\mathbf{x}^\nu; \hat{\boldsymbol{\theta}}_s)) \frac{\partial f(\mathbf{x}^\nu; \hat{\boldsymbol{\theta}}_s)}{\partial \boldsymbol{\lambda}} \quad (2.19)$$

終了条件を満たせば、反復を終了する。

step5 式(2.20)を用いて探索幅 α を求める。

$$\alpha = \frac{\sum_{s=1}^S \sum_{\nu \in D_s} (y^\nu - f(\mathbf{x}^\nu; \hat{\boldsymbol{\theta}}_s)) \frac{\partial f(\mathbf{x}^\nu; \hat{\boldsymbol{\theta}}_s)}{\partial \boldsymbol{\lambda}^T} \Delta \boldsymbol{\lambda}}{\sum_{s=1}^S \sum_{\nu \in D_s} \left(\frac{\partial f(\mathbf{x}^\nu; \hat{\boldsymbol{\theta}}_s)}{\partial \boldsymbol{\lambda}^T} \Delta \boldsymbol{\lambda} \right)^2} \quad (2.20)$$

step6 $\boldsymbol{\lambda}^{(r)} = \boldsymbol{\lambda}^{(r-1)} + \alpha \Delta \boldsymbol{\lambda}$ とする。

$r = r + 1$ とし、step2に戻る。

Step2とStep3で用いる重みの初期値は、それぞれ、 $\hat{\boldsymbol{\theta}}^{(r-1)}$ (または $\boldsymbol{\theta}^{(0)}$)と $\hat{\boldsymbol{\theta}}^{(r)}$ に設定する。これは初期値を解の近くに設定し、学習の効率化を図るためにある。Step4では、勾配ベクトルからBFGS公式を用いて、探索方向を計算する。Step4での終了条件としては、ペナルティ係数の勾配ベクトルが十分小さいときなどが設定される。

MCV正則化法の短所としては、やはり、交差検証法を用いるために繰返しの学習が必要となり、大きな計算コストが必要となる点が挙げられる。特に、モデル選択の方法としても交差検証法を用いる場合、交差検証法が入れ子の構造となり、莫大な計算時間がかかるてしまう。MCV正則化の計算時間増大を回避するための方法として、MCV正則化を単純化したMV(Minimum Validation error)正則化[21]が提案された。MV正則化は、正則化係数の更新に交差検証誤差を用いず、単に検証誤差を用いる方法で、その有効性も示されている。

第 3 章

質的条件付き多変量多項式回帰法

3.1 質的変数を含むデータ

RF5 法では、量的変数のみを含むデータから多変量多項式を発見する方法を考えたが、現実のデータには、量的変数と共に質的変数も含むデータが多く存在する。そのようなデータに対しては、例えば質的変数を無視し、量的変数のみを用いて回帰を行うことが可能である。しかし、多くの場合、量的変数と共に質的変数も用いた方がより精度のよい推定を行うことが可能である。

この章では、量的変数と質的変数が混在した多変量データを対象にした多変量多項式回帰を考える。データに質的変数を含む場合といつても様々な場合があり、例えば目的変数が質的変数の場合や、説明変数が質的変数のみから成る場合なども考えられる。そのような場合に対する有効な手法として様々なものが研究されているが、ここでは、説明変数に量的変数と質的変数の両方を含み、目的変数が量的変数である場合のみを考える。

説明変数に量的変数と質的変数の両方を含む場合における多変量多項式回帰として有名なものに、クーロンの法則がある。クーロンの法則は、二つの電荷を帯びた粒子（荷電粒子）間に働く力の大きさに関する法則で、距離 r だけ離れた電荷 q_1 と q_2 の間に働く力 F の大きさが $F = 4\pi\epsilon q_1 q_2 / r^2$ で表されると規定する。この法則は媒体の誘電率 ϵ を比例定数とし、たとえば水の中では $F = 8897.352 q_1 q_2 / r^2$ 、空気中では $F = 111.280 q_1 q_2 / r^2$ のようになる。従ってこのような場合には、「水中」や「空気中」などといった質的変数を用いて適切に部分空間を構成しつつ、各部分空間の中ではそれぞれに適合する多変量多項式を量的変数を用いて推定する、質的条件付き多変量多項式問題を解くことになる。

質的条件付き多変量多項式回帰が有効だと考えられる問題としては、例えばプロ野球選手の年俸推定問題がある。プロ野球選手の年俸を安打数、打率、勝利数、

防御率などの多項式として推定する問題においては、同じ働きをしても球団Gの選手と球団Cの選手では年俸が異なると思われる所以、別の多項式として推定したい。また、守備位置などによっても、やはり別の多項式として推定すべきである。従って、この年俸推定問題は、質的変数（球団や守備位置）を用いて部分空間を構成し、各部分空間の中ではそれぞれにフィットする多変量多項式を量的変数（安打数、打率、勝利数、防御率など）を用いて推定するのが有効であるといえる。

このような質的条件付き多変量多項式回帰の研究は、多変量多項式回帰の研究と同様に、主に機械学習の分野で進められ、ABACUS[22] や IDS[23] などが提案された。しかし、いずれも組合せ探索のため、やはり実用的な解法とは言えない。そこで、前述のRF5法を拡張して質的条件付き多変量多項式回帰問題を解くRF6法[24] が提案された。さらに、RF6法に自動的なルール復元とモデル選択の改良を加えたRF6.3法[13] が提案された。RF6.3法では、質的変数を扱うためにRF5法で用いた多項式ネットを拡張した上で、ベクトル量子化を用いたルール復元法などを用い、効率良く質的条件付き多変量多項式回帰問題を解く。

3.2 3層パーセプトロンを用いた質的条件付き多変量多項式回帰法： RF6.3法

3.2.1 RF6.3法の基本枠組み

解析対象とするデータは2次元の多変量データとし、各サンプルが変数群 $(q_1, \dots, q_{K_1}, x_1, \dots, x_{K_2}, y)$ で与えられたとする。また、これを (q, x, y) と表す。ただし、 q は質的説明変数、 x は量的説明変数、 y は量的目的変数を表す。また、 L_k を質的変数 q_k がとり得るカテゴリーの数とする。 q_k には、次のように q_{kl} で表現する2進のダミー変数を取り入れる。

$$q_{kl} = \begin{cases} 1 & l \text{番目のカテゴリーにマッチするとき} \\ 0 & \text{それ以外のとき} \end{cases}$$

説明変数を用いて目的変数の挙動を説明する回帰モデルとして、以下のような回帰ルール集合から成る質的条件付き多変量多項式回帰モデルを考える。

$$\text{if } \bigwedge_k \bigvee_{q_{kl} \in Q_k^i} q_{kl} \text{ then } y = \phi(\mathbf{x}; \boldsymbol{\theta}^i) + \epsilon, \quad i = 1, \dots, I \quad (3.1)$$

ここで、 Q_k^i は i 番目のルールで使われるダミー変数の集合、 $\boldsymbol{\theta}^i$ は i 番目のルールの回帰式に現れるパラメータベクトル、 I はルール数である。 $\phi(\mathbf{x}; \boldsymbol{\theta})$ としては、RF5

法で求められるのと同様な指数が整数に限定されない一般化多項式(下式)を考える。

$$\phi(\mathbf{x}; \boldsymbol{\theta}) = w_0 + \sum_{j=1}^{J^*} w_j \prod_{k=1}^{K_2} x_k^{w_{jk}} \quad (3.2)$$

ただし、 w_j, w_{jk} は未知の実数パラメータ、 J は項数に相当する整数パラメータで、 $\boldsymbol{\theta}$ は w_0, w_j, w_{jk} から成るベクトルである。これは前章で説明した3層パーセプトロンの前向き処理である。

3.2.2 条件部の数値表現

次に、回帰ルール集合(3.1)が単一のパーセプトロンで表現できることを示す。まず、条件部を数値表現するために、以下のような関数 c を導入する。

$$c(\mathbf{q}; \mathbf{v}) = g \left(\sum_{k=1}^{K_1} \sum_{l=1}^{L_k} v_{kl} q_{kl} \right) \quad (3.3)$$

ここで、 v は v_{kl} から成るベクトルで、活性化関数 g は指數関数($g(h) = \exp(h)$)またはシグモイド関数($g(h) = \sigma(h)$)とする。 $g(h) = \exp(h)$ に対しては、パラメータ値を以下のように考える。

$$v_{kl}^i = \begin{cases} 0 & \text{if } q_{kl} \in Q_k^i, \\ -\beta & \text{if } q_{kl} \notin Q_k^i, q_{kl'} \in Q_k^i \text{ for some } l' \neq l, \\ 0 & \text{if } q_{kl'} \notin Q_k^i \text{ for any } l' \end{cases} \quad (3.4)$$

ただし、 $\beta \gg 0$ である。このようにパラメータ値を設定すれば、 $c(\mathbf{q}; \mathbf{v}^i)$ は第*i*ルールの条件部の真偽値の近似表現となる。また、 $g(h) = \sigma(h)$ に対しては、パラメータ値を以下のように考える。

$$v_{kl}^i = \begin{cases} \beta_2 & \text{if } q_{kl} \in Q_k^i, \\ -\beta_1 & \text{if } q_{kl} \notin Q_k^i, q_{kl'} \in Q_k^i \text{ for some } l' \neq l, \\ 0 & \text{if } q_{kl'} \notin Q_k^i \text{ for any } l' \end{cases} \quad (3.5)$$

ただし、 $\beta_1 \gg \beta_2 \gg 0$ である。このようにパラメータ値を設定すれば、やはり $c(\mathbf{q}; \mathbf{v}^i)$ は第*i*ルールの条件部の真偽値の近似表現となる。よって、式(3.1)で定義されたルール集合からの最終出力に対して、下式は十分な精度の近似となる。

$$F(\mathbf{q}, \mathbf{x}; \mathbf{v}^1, \dots, \mathbf{v}^I, \boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^I) = \sum_{i=1}^I c(\mathbf{q}; \mathbf{v}^i) \phi(\mathbf{x}; \boldsymbol{\theta}^i) \quad (3.6)$$

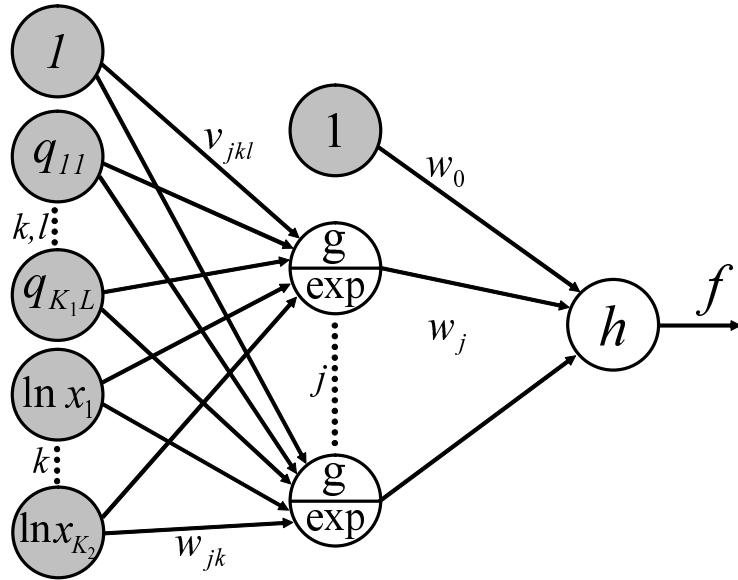


図 3.1: RF6.3 法を実現する 3 層パーセプトロン

一方，適当な J の値が分かれば，次式は式 (3.6) を完全に表現することができる。
ただし， Θ は $w_0, w_j, v_{jkl}, w_{jk}$ から成るベクトルである。

$$f(\mathbf{q}, \mathbf{x}; \Theta) = w_0 + \sum_{j=1}^J w_j g \left(\sum_{k=1}^{K_1} \sum_{l=1}^{L_k} v_{jkl} q_{kl} \right) \exp \left(\sum_{k=1}^{K_2} w_{jk} \ln x_k \right) \quad (3.7)$$

従って，式 (3.7) で複数のルールを同時に学習できることがわかる。ただし，式 (3.7) で式 (3.1) を近似するとき，重み v は式 (3.4) や (3.5) で定義した重みに限定されない。なお，式 (3.7) も図 3.1 のような 3 層パーセプトロンの前向き処理と見なすことができる。

3.3 4 層パーセプトロンへの拡張：RF6.4 法

3.3.1 RF6.3 法の問題点

RF6.3 では質的変数の部分の重みの表現が線形であるため，発見できる多項式群にはかなりの制限がある。例えば，以下のような形の多項式群の発見について考える。

$$\begin{cases} \text{if } q_{11} \wedge q_{21} \text{ then } y = w_0 + p_1 x_1 x_2 \\ \text{if } q_{11} \wedge q_{22} \text{ then } y = w_0 + p_2 x_1 x_2 \\ \text{if } q_{12} \text{ then } y = w_0 + p_3 x_1 x_2 \end{cases} \quad (3.8)$$

ただし，量的説明変数は x_1, x_2 ，質的説明変数は q_1, q_2 で，質的変数のカテゴリ数

は $L_1 = L_2 = 2$ とする。式(3.7)より、係数部分 p_1, p_2, p_3 は以下のように表される。

$$p_1 = w_1 g(v_{111} + v_{121}) \quad (3.9)$$

$$p_2 = w_1 g(v_{111} + v_{122}) \quad (3.10)$$

$$p_3 = w_1 g(v_{112} + v_{121}) = w_1 g(v_{112} + v_{122}) \quad (3.11)$$

式(3.11)より $v_{121} = v_{122}$ となり、式(3.9), (3.10)より $p_1 = p_2$ となってしまう。質的条件付き多変量多項式回帰では当然 $p_1 \neq p_2 \neq p_3$ であることを想定しているので、これでは具合が悪い。このように、ルール条件部の構成によっては RF6.3 法では表現できないような場合が存在する。これは、質的変数部分のパーセプトロンの構造が单層で線形となっており、非線形な判別ができないのが原因である。これに対処するために、質的変数部分のパーセプトロンの構造を 2 層にして、複雑な条件部を表現できるように改良した RF6.4 法を提案する。本節では、RF6.4 法について詳しく説明する。

3.3.2 RF6.4 法の基本枠組み

RF6.4 法では質的変数部分を 2 層の非線形のネットワークで表現して、回帰ルール条件部の表現能力を向上させている。さらに、RF6.3 法では各部分空間で定数項の値が同じであったのに対し、RF6.4 法では部分空間ごとに定数項の値も変わることができるようになっている。質的変数部分を 2 層にしたため、パーセプトロン全体では RF6.3 法が 3 層だったのに対し、RF6.4 法では 4 層となっている。RF6.4 法での 4 層パーセプトロンは図 3.2 のように表され、その出力の式は以下となる。

$$\begin{aligned} f(\mathbf{q}, \mathbf{x}; \Theta) &= c_0 + \sum_{j=1}^J c_j m_j, \\ c_0 &= \sum_{r=1}^R v_{0r} \sigma_r, \quad c_j = \sum_{r=1}^R v_{jr} \sigma_r, \\ \sigma_r &= \sigma \left(\sum_{k=1}^{K_1} \sum_{l=1}^{L_k} v_{rkl} q_{kl} \right), \quad m_j = \exp \left(\sum_{k=1}^{K_2} w_{jk} \ln x_k \right) \end{aligned} \quad (3.12)$$

ここで、 R, J は共に隠れユニット数であり、特に J は多項式の項数にあたる。パーセプトロンの学習には、RF5 法と同様に BPQ 法を用いる。なお、学習の際に必要な微分値の計算については、付録に掲載した。

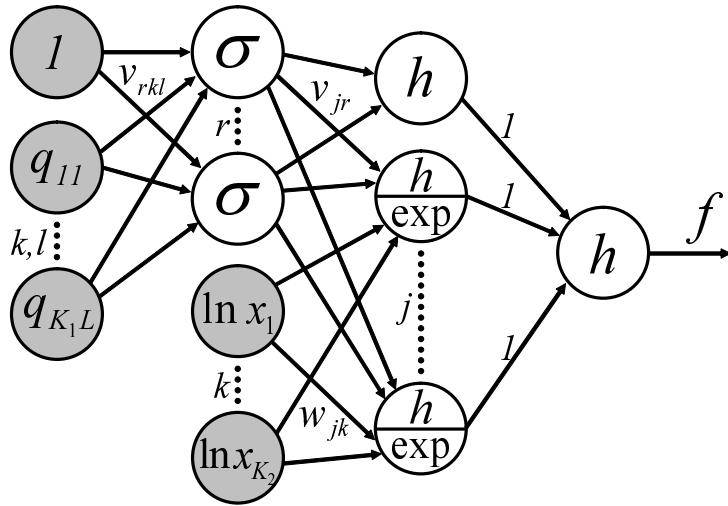


図 3.2: RF6.4 法を実現する 4 層パーセプトロン

3.3.3 指数部分が異なる多項式群の発見

RF6.4 法では、例えば以下のような形の回帰ルール集合をデータから発見することが可能である。

$$\left\{ \begin{array}{lll} \text{if } & q_{11} \wedge q_{21} & \text{then } y = 2 + 2x_1^2 x_2^{1/3} - 3x_2^{-1} x_3^{1/2} \\ \text{if } & q_{11} \wedge q_{22} & \text{then } y = 1 + 3x_1^2 x_2^{1/3} + 4x_2^{-1} x_3^{1/2} \\ \text{if } & q_{12} & \text{then } y = 3 - 4x_1^2 x_2^{1/3} + 2x_2^{-1} x_3^{1/2} \end{array} \right. \quad (3.13)$$

RF6.4 法では、図 3.2 のように、質的変数によって定数項部分の値も変わることができるようになっているので、式 (3.13) のように、回帰ルールごとに異なる定数項の値を取ることができる。

RF6.4 法では、基本的には各回帰ルールにおいて、式 (3.13) のように、多項式の係数値だけが異なり、指数部分の値は全ルールにおいて共通の値になるような回帰ルール集合を発見する。もし回帰ルールごとに指数部分も異なってしまうと、その最終出力は非常に複雑なものとなり、解釈困難な結果となる可能性がある。多項式の表現能力は強力であるので、指数部分の値が各回帰ルールで共通であるという制約の元でも、十分に汎化能力の高い出力を得ることはできると思われる。

しかし、RF6.4 法では RF6.3 法に比べて表現能力が向上しているため、指数部分が異なる回帰ルール集合を発見することも可能である。例えば、以下のような

回帰ルール集合の発見について考える。

$$\begin{cases} \text{if } q_{11} \wedge q_{21} \text{ then } y = 2 + 2x_1^2 x_2^{-1/3} x_3^{-2} \\ \text{if } q_{11} \wedge q_{22} \text{ then } y = 1 + 3x_1^{1/2} x_2^{1/3} x_3^3 \\ \text{if } q_{12} \text{ then } y = 3 - 4x_1^3 x_2^1 x_3^{-1/2} \end{cases} \quad (3.14)$$

式(3.14)は式(3.13)とは異なり、指数部分の値が回帰ルールごとに異なった値になっている。しかし、式(3.14)は以下のように解釈することが可能である。

$$\begin{cases} \text{if } q_{11} \wedge q_{21} \text{ then } y = 2 + 2x_1^2 x_2^{-1/3} x_3^{-2} + 0x_1^{1/2} x_2^{1/3} x_3^3 + 0x_1^3 x_2^1 x_3^{-1/2} \\ \text{if } q_{11} \wedge q_{22} \text{ then } y = 1 + 0x_1^2 x_2^{-1/3} x_3^{-2} + 3x_1^{1/2} x_2^{1/3} x_3^3 + 0x_1^3 x_2^1 x_3^{-1/2} \\ \text{if } q_{12} \text{ then } y = 3 + 0x_1^2 x_2^{-1/3} x_3^{-2} + 0x_1^{1/2} x_2^{1/3} x_3^3 - 4x_1^3 x_2^1 x_3^{-1/2} \end{cases} \quad (3.15)$$

このように、不要な項の係数値が0となることによって、回帰ルールごとに異なる指数値となる回帰ルール集合を発見することも可能である。

3.4 ルール復元法

次に、学習済みのパーセプトロンから式(3.1)の形の回帰ルール集合を取り出す回帰ルール復元法について説明する。

式(3.1)を用いれば、質的変数の各カテゴリーの全組合せに対応して回帰ルール集合を作ることができるが、回帰ルール数が多くなるだけでなく、各回帰ルールは質的変数の値が全く同じである少数のサンプルしかカバーしておらず、特化しすぎているので好ましくない。そこで、サンプル分布の代表点を求めることにより、より一般化された回帰ルール集合を得ることを考える。対象データを用いて学習した後に得られる多層パーセプトロンのパラメータを \hat{v}_{jr} と \hat{v}_{rkl} とすると、多項式の第 j 項の係数のサンプルごとの値は以下で表される。

$$c_j^\mu = \sum_{r=1}^R \hat{v}_{jr} \hat{\sigma}_r^\mu, \quad \hat{\sigma}_r^\mu = \sigma \left(\sum_{k=1}^{K_1} \sum_{l=1}^{L_k} \hat{v}_{rkl} q_{kl}^\mu \right) \quad (3.16)$$

ここで、 N 個の係数値ベクトル $\{c^\mu = (c_0^\mu, c_1^\mu, \dots, c_{J^*}^\mu) : \mu = 1, \dots, N\}$ を I 個の代表点 $\{a^i = (a_0^i, a_1^i, \dots, a_{J^*}^i) : i = 1, \dots, I\}$ にベクトル量子化する。ベクトル量子化の方法には、アルゴリズムの簡潔性を考慮して、K平均法を採用する。K平均法では、すべての係数値ベクトルを最も近い代表点へ割当てた後、各代表点をそのクラスタの重心へ移動という操作を収束するまで繰り返す。その結果、すべての係数値ベクトルが I 個の互いに素なクラスタ集合 $\{G_i, i = 1, \dots, I\}$ に振り分けられる。このとき、K平均法では以下の二乗ひずみが最小（極小）になる。

$$d_{VQ} = \sum_{i=1}^I \sum_{\mu \in G_i} \|c^\mu - a^i\|^2, \quad a^i = \frac{1}{N_i} \sum_{\mu \in G_i} c^\mu \quad (3.17)$$

ただし， N_i はクラスタ G_i に含まれるサンプルの個数である。ここで，サンプル μ の係数値ベクトルに最も近い代表点の番号を返す関数を以下で定義する。

$$i(\mathbf{q}^\mu) = \arg \min_i \|\mathbf{c}^\mu - \mathbf{a}^i\|^2 \quad (3.18)$$

この $i(\mathbf{q}^\mu)$ を回帰ルールの条件部に用いれば，代表点の値を係数値とする多項式から成る以下の回帰ルール集合が得られる。

$$\text{if } i(\mathbf{q}) = i \text{ then } \hat{f} = a_0^i + \sum_{j=1}^J a_j^i \prod_{k=1}^{K_2} x_k^{\hat{w}_{jk}}, \quad i = 1, \dots, I \quad (3.19)$$

サンプルが与えられると， $i(\mathbf{q})$ はどれか一つの回帰ルールの条件部と合致し，その実行部の回帰式のみが適用される。

最後に，関数 $i(\mathbf{q})$ を式(3.1)で定義したような連言標準形の条件部に変換する必要がある。そのためには，訓練データを $\{(\mathbf{q}^\mu, i(\mathbf{q}^\mu)), \mu = 1, \dots, N\}$ とする分類問題を解けばよい。このとき $i(\mathbf{q})$ はサンプル μ のクラスレベルを表す。ここでは，分類問題を解くために，機械学習分野で知られる c4.5 決定木生成プログラム [25] を用いる。生成された決定木を条件部とすれば，式(3.1)で定義したような回帰ルール集合が得られる。

3.5 モデル選択

RF6.4 法では，モデル選択によって決定すべきパラメータが 3 つある。それは，最適中間ユニット数 J^* , R^* 及び最適ルール数 I^* である。なお， J^* は多項式の項数となるパラメータだが， R^* は RF6.4 法での最終的な出力に陽には現れない。

まずは 4 層パーセプトロンの中間ユニット数 J^* 及び R^* を決定する。モデル選択の方法としては，RF5 法と同様に，交差検証法または BIC を用いることを考える。交差検証法を用いる場合は J, R を変えてパーセプトロンを学習し，式(3.20)が，最小となったときの J, R をそれぞれ J^*, R^* とすればよい。

$$\text{MSE}_{\text{CV}} = \frac{1}{N} \sum_{s=1}^S \sum_{\mu \in G_s} \left(f(\mathbf{q}, \mathbf{x}; \Theta) - y^\mu \right)^2 \quad (3.20)$$

BIC を用いる場合は式(3.21)を評価尺度として用いる。

$$\text{BIC}(J, R) = \frac{N}{2} \log \left(\frac{1}{N} \sum_{\mu=1}^N \left(f(\mathbf{x}^\mu; \widehat{\Theta}_{J,R}) - y^\mu \right)^2 \right) + \frac{M}{2} \log N \quad (3.21)$$

ただし， $\widehat{\Theta}_{J,R}$ は中間ユニット数が J, R のパーセプトロンの学習後に得られたパラメータベクトルを表す。

J^*, R^* が決まつたら，そのときの係数値ベクトルをベクトル量子化して，回帰ルールを復元する．ここで，最適ルール数 I^* を決定する必要がある． I^* の決定には， J, R のときと同様に交差検証法を用いることは可能であるが，(3.21) のように出力と目的変数との平均自乗誤差を尤度として用いて BIC を適用することは望ましくない．なぜなら，K 平均法によるベクトル量子化では，クラスタとそのクラスタの代表点との距離が極小になるように学習が進むのであって，出力と目的変数との誤差は極小にはならないので，その誤差は正確な尤度とはなっていないからである．従って，BIC によってモデル選択を行う場合には，他の方法で適用する必要がある．ここでは，BIC の適用方法として，Pelleg らによる X-means [26] で用いられた方法を採用する．まず， $J+1$ 次元係数値ベクトル \mathbf{c}^μ とそれが属するクラスタの重心ベクトル $\mathbf{a}^{(\mu)}$ との誤差 ϵ が $J+1$ 次元正規分布に従うと仮定し，データの従う密度関数 $\hat{P}(\mathbf{c}^\mu)$ を以下とする．

$$\hat{P}(\mathbf{c}^\mu) = \frac{N^{(\mu)}}{N} \frac{1}{(\sqrt{2\pi})^{J+1} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\mathbf{c}^\mu - \mathbf{a}^{(\mu)})^T \Sigma^{-1} (\mathbf{c}^\mu - \mathbf{a}^{(\mu)})\right) \quad (3.22)$$

ただし， $N^{(\mu)}$ はサンプル \mathbf{c}^μ が属するクラスタのサンプル数を表す． Σ は共分散行列を表し，その最尤推定量は以下で計算できる．

$$\hat{\Sigma} = \frac{1}{N-I} \sum_{i=1}^I \sum_{\mu \in G_i} (\mathbf{c}^\mu - \mathbf{a}^{(\mu)}) (\mathbf{c}^\mu - \mathbf{a}^{(\mu)})^T \quad (3.23)$$

このとき，最大対数尤度は以下となる．

$$\begin{aligned} L(\hat{\theta}) &= \log \prod_{\mu=1}^N \hat{P}(\mathbf{c}^\mu) \\ &= \sum_{\mu=1}^N \left(\log N^{(\mu)} - \log N - \frac{J+1}{2} \log(2\pi) - \frac{1}{2} \log |\hat{\Sigma}| \right. \\ &\quad \left. - \frac{1}{2} (\mathbf{c}^\mu - \mathbf{a}^{(\mu)})^T \hat{\Sigma}^{-1} (\mathbf{c}^\mu - \mathbf{a}^{(\mu)}) \right) \end{aligned} \quad (3.24)$$

モデル選択に無関係な項を除外すると， I^* を選択するための $BIC(I)$ は以下となる．

$$\begin{aligned} BIC(I) &= - \sum_{i=1}^I N_i \log N_i + \frac{N}{2} \log |\hat{\Sigma}| \frac{1}{2} \sum_{\mu=1}^N (\mathbf{c}^\mu - \mathbf{a}^{(\mu)})^T \hat{\Sigma}^{-1} (\mathbf{c}^\mu - \mathbf{a}^{(\mu)}) \\ &\quad + \frac{I(J+2)}{2} \log N \end{aligned} \quad (3.25)$$

Pelleg らによる X-means では共分散行列は対角行列とし，更に全ての次元において同じ分散値を仮定しているが，上述の提案法ではフルの共分散行列を仮定して

$\text{BIC}(I)$ を導出した。その理由は、量子化する係数値ベクトルに正規化を施さないので、係数の各次元によってそのスケールが大きく異なる場合があるからである。しかし、パーセプトロンの構造のために $\text{rank}(\widehat{\Sigma}) = R^*$ となっており、 $J+1 > R^*$ の場合には共分散行列 $\widehat{\Sigma}$ の逆行列が計算できないので、この場合、実際にフルの共分散行列を用いて $\text{BIC}(I)$ を計算することはできない。そこで、この場合には共分散行列は対角行列として $\text{BIC}(I)$ を計算するが、各次元で違う分散値を取れる点が Pelleg らの方法と異なる。

第 4 章

質的条件付き重回帰法

4.1 重回帰分析と数量化理論一類

多変量解析法の 1 つである重回帰分析は、ある数値変数(目的変数)を、他のいくつかの量的変数(説明変数)に基づいて線形回帰式として予測するものである。重回帰分析はモデル自体がわかりやすく応用に富み、計算も容易であるため、統計分野において非常によく用いられている手法である。また、説明変数に質的変数を含むデータに対しては、数量化理論一類がある。これは、質的変数を 0, 1 のダミー変数で表し、質的変数の各カテゴリを 1 つの独立した量的変数として重回帰分析を行う手法である。

データが $\{(x^\mu, y^\mu), \mu = 1, \dots, N\}$ で与えられたとすると、重回帰分析の出力式は以下で表される。ただし、 w はパラメータ w_k (定数項 w_0 を含む) から成るベクトルで、 x には w_0 に対応する変数 $x_0 = 1$ を含むとする。

$$y^\mu = w_0 + w_1 x_1^\mu + \dots + w_j x_j^\mu = \mathbf{w}^T \mathbf{x}^\mu \quad (4.1)$$

なお、数量化理論一類の場合は、入力 x にダミー化した質的変数を含めるとすれば、重回帰分析と同様に扱うことができる。

最適なパラメータ \hat{w} とは、以下の平均二乗誤差を最小にする w である。

$$E = \frac{1}{N} \sum_{\mu=1}^N (\mathbf{w}^T \mathbf{x}^\mu - y^\mu)^2 \quad (4.2)$$

極値条件 $\partial E / \partial \mathbf{w} = 0$ より、以下の正規方程式を解くことによって \hat{w} を求めることができる。

$$\left(\sum_{\mu=1}^N \mathbf{x}^\mu (\mathbf{x}^\mu)^T \right) \hat{\mathbf{w}} = \sum_{\mu=1}^N y^\mu \mathbf{x}^\mu \quad (4.3)$$

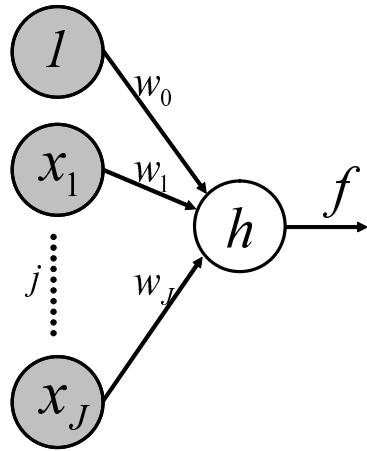


図 4.1: 重回帰分析を実現する单層パーセプトロン

逆行列が求まるとき，最小二乗解は以下で与えられる．

$$\hat{\mathbf{w}} = \left(\sum_{\mu=1}^N \mathbf{x}^\mu (\mathbf{x}^\mu)^T \right)^{-1} \left(\sum_{\mu=1}^N y^\mu \mathbf{x}^\mu \right) \quad (4.4)$$

$\hat{\mathbf{w}}$ は式 (4.4) によって一挙に求めることができるが，式 (4.2) を目的関数として図 4.1 のような单層パーセプトロンを学習することによっても求めることができる．

4.2 複数の関数を割り当てるモデル

しかし，重回帰分析や数量化理論一類は単純な線形回帰であるため，非線形のデータや区分線形のデータに対しては柔軟に対応することができず，十分な汎化性能は期待できない．

この章では，第3章と同様に，説明変数として質的変数と量的変数が混在した多変量データを扱う．そして，RF5法やRF6.4法のような多変量多項式モデルとは別のアプローチで，可読性と汎化性の高いモデルを構築することを考える．具体的には，RF6.4法のように質的変数によって条件付けられた回帰式を求めるモデルで，回帰式は線形回帰式とする．すなわち，全体のデータに複数の線形回帰式を割り当てて適合させる質的条件付き重回帰法について考える．

全体のデータに複数の関数を割り当てて適合させる回帰モデルとしては，RF6.4法のような質的条件付き多変量多項式回帰法以外にもマルチモジュールニューラルネット [27]，クラスタ分割関数回帰 [28] [29]，HME[30] [31] などが提案された．マルチモジュールニューラルネットやクラスタ分割関数回帰では，全体空間を量的

変数によって分割し，量的変数の値が近いサンプルが同じクラスタ（部分空間と同義）を構成する．しかし，データによってはその分割方法が最善とはならない場合がある．例えば，プロ野球選手の年俸をその選手の成績から推定するような問題では，似た成績の選手を集めて1つのクラスタとするよりも，チームや守備位置といった質的変数をもとにクラスタを構成する方が自然である．HMEではもともとは量的変数によってクラスタを構成するように考えられたが，ゲート変数として質的変数を用いることにより，質的変数でのクラスタ分割も可能である．

この複数の関数を割り当てる枠組みのもとに，量的変数と質的変数が混在するデータに対する質的条件付き重回帰を実現する，nc-MR法（nominally conditioned multiple regression）を提案する．nc-MR法では，質的変数空間を自動的かつ適切に部分空間群に分割し，各部分空間内ではそれぞれに適合する線形回帰式を量的変数を用いて推定する．適切な部分空間群を構成すれば，各部分空間内での回帰式は，線形でも良好な汎化性能が期待できる．さらに，線形回帰式は多変量多項式などと比べて可読性が良いという利点もある．なお，提案法とHMEとは，式や学習方法は大きく異なるが，質的空間を分割して各部分空間を線形回帰する考え方には類似しているといえる．

説明変数に量的変数と質的変数の両方を用いて線形回帰を行う方法としては，次の2つが最も単純な方法であるといえる．まず1つめは，数量化理論一類であり，この場合，各ルールの回帰式が質的変数の値によって定数部分の値のみ異なる．すなわち，各部分空間の回帰曲線（曲面）は，互いに平行となる．そして2つめは，質的変数の値の全ての組合せごとに個別な部分空間を考え，それの中でも量的変数を用いて通常の重回帰を行う方法である．ただし，この方法ではそれぞれの質的変数の値ごとに十分な数のサンプルが必要となるため，質的変数が多いデータに対しての適用は難しいといえる．更に，それぞれの部分空間での線形回帰式は互いに関係がなくなってしまう．例えば，プロ野球選手の年俸推定問題では，チームGとTの回帰式は異なるかもしれないが，それらは全く独立とは考えにくく，互いに何らかの関係と考えられるので，この方法は一般的にもあまり好ましくないといえる．nc-MR法は，これら2つの単純な方法の中間的な位置付けのモデルとなっている．すなわち，nc-MR法における各部分空間での回帰式は，数量化一類などの共通性はないが，個別に重回帰を行う方法のように，全く無関係というわけではない．多層パーセプトロンを用いることによって，各部分空間での回帰式に，なんらかの相関を持たせることが可能となる．

4.3 4層パーセプトロンを用いた質的条件付き重回帰法：nc-MR法

与えられるデータはRF6.4法の場合と同様な2次元の多変量データとし、各サンプルが変数群 $(q_1, \dots, q_K, x_1, \dots, x_J, y)$ で表現されるとする。ただし、 q_k は質的説明変数、 x_j は量的説明変数、 y は基準変数を表す。また、 L_k を質的変数 q_k がとり得るカテゴリーの数とする。各 q_k はRF6.4法の場合と同様に2進ダミー変数 q_{kl} で表現する。すなわち、 q_{kl} は、 q_k 値が第 l カテゴリーのとき1で、その他は0となる。

説明変数を用いて基準変数の挙動を説明する回帰モデルとして、以下のような回帰ルール集合 $(i = 1, \dots, I)$ から成る質的条件付き重回帰モデルを考える。

$$\text{if } \bigwedge_k \bigvee_{q_{kl} \in Q_k^i} q_{kl} \text{ then } y = \phi(\mathbf{x}; \mathbf{w}^i), \quad i = 1, \dots, I \quad (4.5)$$

回帰ルールはif-then型とし、条件部で規定する質的変数の部分空間に、実行部で規定する回帰式を適用することを意味する。なお、 Q_k^i は第*i*回帰ルールで使われる q_k に対応するダミー変数の集合、 \mathbf{x} は x_j から成るベクトル、 \mathbf{w}^i は第*i*回帰ルールの回帰式に現れるパラメータベクトル、*I*は回帰ルール数を表す。 $\phi(\mathbf{x}; \mathbf{w})$ は各回帰ルールでの線形回帰関数で、 $\phi(\mathbf{x}; \mathbf{w}) = \mathbf{w}^T \mathbf{x}$ である。ただし、 \mathbf{w} はバイアス w_0 を含み、 \mathbf{x} は $w_0 (= 1)$ に対応する x_0 を含むとする。

次に、式(4.5)が単一の4層パーセプトロンで表現できることを示す。まず、条件部を数値表現するために、以下の関数 c を導入する。

$$c(\mathbf{q}; \mathbf{v}^i) = \sigma \left(\sum_{k=1}^K \sum_{l=1}^{L_k} v_{kl}^i q_{kl} \right) \quad (4.6)$$

ただし、 v^i はパラメータ v_{kl}^i を要素とするベクトルで、 $\sigma(h)$ はシグモイド関数 $\sigma(h) = 1/(1 + e^{-h})$ とする。このとき、式(4.5)で定義された回帰ルール集合からの最終出力に対して、下式は十分な精度の近似となる。

$$F(\mathbf{q}, \mathbf{x}; \mathbf{v}^1, \dots, \mathbf{v}^I, \mathbf{w}^1, \dots, \mathbf{w}^I) = \sum_{i=1}^I c(\mathbf{q}; \mathbf{v}^i) \phi(\mathbf{x}; \mathbf{w}^i) \quad (4.7)$$

以上はRF6.4法の基本枠組みと同様であるが、nc-MR法の最終出力は、RF6.4法とは大きく異なる。

さて、nc-MR法では、 R を適切に用意すれば、下式は $\phi(\mathbf{x}; \mathbf{w}^i) = (\mathbf{w}^i)^T \mathbf{x}$ とする式(4.7)の最終出力を精度良く表現できる。

$$f(\mathbf{q}, \mathbf{x}; \boldsymbol{\theta}) = \sum_{j=0}^J c_j x_j, \quad c_j = \sum_{r=1}^R v_{jr} \sigma_r, \quad \sigma_r = \sigma \left(v_{r0} + \sum_{k=1}^K \sum_{l=1}^{L_k-1} v_{rkl} q_{kl} \right) \quad (4.8)$$

ただし， θ はパラメータ v_{jr}, v_{rkl} 及びバイアス v_{r0} から成るベクトルである。

式(4.8)は図4.2で表される4層パーセプトロンで学習することができる。 R は第2層の中間ユニット数に相当する。パラメータ v_{jr} の値によって各部分空間における w を十分に表現できるため、第3層と第4層の間のパラメータを1と固定しても問題ない。RF6.3法では、単一の3層パーセプトロンで複数の回帰ルールを一度に学習するが、提案法ではRF6.4法と同様にパーセプトロンにおいて質的変数を数値表現する部分を2層にし、全体では4層となっている。そうすることで、RF6.3法のような3層での表現より複雑な条件部を表現することが可能となる。

また、それに伴い、入力 q, x の次元が大きいときにはパラメータ数を減らすことができる。 q をダミー変数化した後の次元数を K_{all} 、 x のバイアスを含めた次元数を J_{all} とすると、3層パーセプトロンで学習する場合の総パラメータ数は $K_{all}J_{all}$ 個となり、それぞれの次元の積となるので、非常に多くのパラメータが必要となる。一方、nc-MR法の4層パーセプトロンの場合の総パラメータ数は $(K_{all} + J_{all})R$ 個となり、それぞれの次元の和と中間ユニット数との積になる。通常は R の値はあまり大きくななくても、十分に複雑なルールを表現できるので、多くの場合、3層に比べて4層の方がパラメータ数が減り、より安定した結果を得ることが可能となる。

更に、パラメータ数を減らすため、式(4.8)の σ_r の計算では、 l の範囲を1から $L_k - 1$ までとする。 q_{kl} はダミー変数であり、どれか1つのカテゴリが1で他のカテゴリは全て0になるので、 $\sum_{l=1}^{L_k} q_{kl} = 1$ という線形制約式が成り立つ。従って、以下に示すように、バイアス v_{r0} を入れれば、 l についての和は $L_k - 1$ までで十分である。それによって、パラメータ数を減らすだけでなく、モデルの冗長性を取り除けるので、より安定した結果が望める。

$$\begin{aligned} \sum_{k=1}^K \sum_{l=1}^{L_k} v_{rkl} q_{kl} &= \sum_{k=1}^K \sum_{l=1}^{L_k-1} v_{rkl} q_{kl} + \sum_{k=1}^K v_{rkL_k} \left(1 - \sum_{l=1}^{L_k-1} q_{kl} \right) \\ &= \sum_{k=1}^K v_{rkL_k} + \sum_{k=1}^K \sum_{l=1}^{L_k-1} (v_{rkl} - v_{rkL_k}) q_{kl} \end{aligned} \quad (4.9)$$

なお、上記のパラメータ数を減らす工夫は、RF6.4法にも適用することが可能である。

パーセプトロンの学習にはBPQ法を用いる。なお、学習の際に必要な微分値の計算については、付録に掲載した。

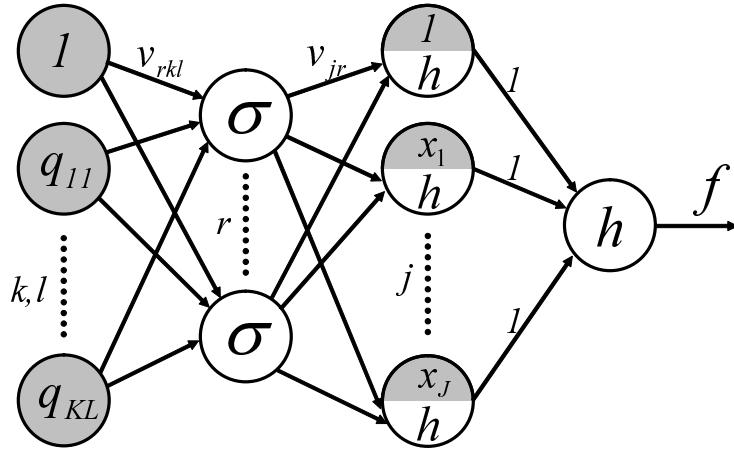


図 4.2: nc-MR 法を実現する 4 層パーセプトロン

4.4 回帰ルール復元における工夫

nc-MR 法では、RF6.4 法と同様に回帰ルールの条件部の復元を行う必要がある。その回帰ルール復元の方法は RF6.4 法と基本枠組みは同じであるが、RF6.4 法でのルール復元に改良を加えて、計算コストを削減している。そのための工夫について、以下に詳細に述べる。なお、以下の工夫は RF6.4 法に適用することも可能である。

RF6.4 法では、あるサンプルにおける係数値ベクトルの長さは多項式の項数と同じであり、サイズはそれほど大きくなかったので、全サンプルに対する係数値ベクトルに対して K 平均法を用いてベクトル量子化を行っても、計算コストはそれほどかからなかった。しかし、nc-MR 法では、あるサンプルにおける係数値ベクトルの長さは量的説明変数の種類数と同じになるので、量的説明変数の種類が多い場合には係数値ベクトルのサイズが大きくなり、ベクトル量子化にも比較的大きい計算コストがかかってしまう。

そこで、nc-MR 法では計算量削減のため、全サンプルをそのまま用いず、異なる質的変数の組合せを持つサンプルのみを用いる。各サンプルに対する係数値ベクトルは、質的変数の値のみで決定され、量的変数の値には影響されない。よって、全サンプルのうち、異なる質的変数の組合せとなるサンプルだけを用いてベクトル量子化を行えば、同じ質的変数の組合せのサンプルを多く含むデータに対しては、ベクトル量子化の計算量を大幅に削減することができる。しかし、その

まま異なる質的変数の組合せとなるサンプルだけを用いてベクトル量子化を行うと，全サンプルを用いてベクトル量子化を行った場合と異なる結果を得ることになってしまう。それはK平均法における重心計算の結果が異なることに起因する。

そこで，全サンプルを用いた場合と同じ結果が得られるように，K平均法における重心の計算式に変更を加える。通常，領域 S^i に含まれるベクトルの重心 r^i は以下の式で計算される。ただし，領域 S^i に含まれるベクトル数を $|S^i|$ とする。

$$\mathbf{r}^i = \frac{1}{|S^i|} \sum_{\mathbf{c}^\mu \in S^i} \mathbf{c}^\mu \quad (4.10)$$

これを以下のように変更することによって，重心を(4.10)で計算したときと同じ結果を得ることができる。

$$\mathbf{r}^i = \frac{1}{\sum_{\mathbf{c}^\nu \in S^i} n^\nu} \sum_{\mathbf{c}^\nu \in S^i} n^\nu \mathbf{c}^\nu \quad (4.11)$$

ただし，ベクトル \mathbf{c}^ν と質的変数の値が同じサンプルの個数を n^ν とする。

上記の方法でK平均法を行った後，c4.5決定木生成プログラムを用いて分類問題を解くことによって条件部を復元する。

4.5 モデル選択

nc-MR法でもRF6.4法と同様に，最適隠れユニット数 R^* 及び最適ルール数 I^* をなんらかのモデル選択法によって決定する必要がある。その方法として，交差検証法またはBICを用いる。それらの適用方法は，RF6.4法と同じである。ただし，RF6.4法では中間ユニット数は J, R の2つがあったが，nc-MR法では R のみなので，どちらの方法でモデル選択を行ったとしても，RF6.4法に比べて計算コストは少ない。

4.6 回帰ルールの位置関係

4.6.1 回帰ルールの近さ

4層パーセプトロンを学習後，K平均法によるベクトル量子化を行った後に分類問題を解けば，回帰ルールの集合として最終出力を得ることができる。nc-MR法によって，解釈がしやすく，後に有効に利用しやすい結果を得ることができるが，数式だけでは解釈し難いこともある。nc-MR法ではモデル選択によって最適な回帰ルール数が得られ，それぞれの回帰ルールごとの線形回帰式が得られるが，

各回帰ルールは全く独立な回帰式となっているわけではなく、回帰ルールは異なっても似たような式が得られる場合もある。例えば、プロ野球選手の年俸を推定する問題においては、モデル選択によってチームTとチームGでの推定式は異なるものが得られたとしても、数式を詳細に調べてみると、実は似たような式が得られているかもしれない。また、チームGとチームDでは、数式自体が全く異なるものとなっているかもしれない。つまり、異なる回帰ルールでも、回帰ルールごとの「近さ」があるといえる。

回帰ルールごとの「近さ」は、例えばそれぞれの回帰ルールごとの係数値ベクトルの距離を計算すれば、「近さ」を数値として表示することが可能である。その場合には、各係数値ベクトルの距離行列を計算することになる。

例えば、以下の回帰ルール集合について考える。

$$\left\{ \begin{array}{ll} \text{if } q_{11} \wedge q_{21} \text{ then } y = +1 + 2x_1 + 3x_2 + 4x_3 + 5x_4 + 6x_5 \\ \text{if } q_{11} \wedge q_{22} \text{ then } y = -1 + 4x_1 - 2x_2 + 5x_3 - 1x_4 + 2x_5 \\ \text{if } q_{11} \wedge q_{23} \text{ then } y = +3 - 3x_1 + 2x_2 - 2x_3 - 4x_4 - 3x_5 \\ \text{if } q_{12} \wedge q_{21} \text{ then } y = +1 + 2x_1 - 1x_2 + 3x_3 + 3x_4 + 4x_5 \\ \text{if } q_{12} \wedge q_{22} \text{ then } y = +2 + 1x_1 + 2x_2 - 1x_3 - 2x_4 - 4x_5 \\ \text{if } q_{12} \wedge q_{23} \text{ then } y = -2 + 3x_1 + 1x_2 + 1x_3 + 2x_4 + 2x_5 \end{array} \right. \quad (4.12)$$

上記の回帰ルール集合の係数値を見るだけでは、どの回帰ルール同士が似ていて、どの回帰ルール同士がより異なっているのかは分かりにくい。

このとき、各係数値ベクトルの距離行列 D (2乗ノルム) を計算すると、以下のようなになる。

$$D = \begin{pmatrix} 0 & 86 & 228 & 25 & 177 & 48 \\ 86 & 0 & 164 & 33 & 107 & 36 \\ 228 & 164 & 0 & 161 & 23 & 132 \\ 25 & 33 & 161 & 0 & 116 & 23 \\ 177 & 107 & 23 & 116 & 0 & 77 \\ 48 & 36 & 132 & 23 & 77 & 0 \end{pmatrix} \quad (4.13)$$

距離行列によって、例えば回帰ルール6と回帰ルール4が最も近いということや、回帰ルール1と回帰ルール3が最も異なった回帰式となっていることなどが分かる。また、その他にもどの回帰ルールとどの回帰ルールが近いのかが、回帰ルール集合の式の表現よりは、比較的分かりやすくなっている。

しかし、回帰ルール集合全体としては、それぞれの回帰ルールの近さは距離行列では直感的には分かりづらい。各回帰ルールの近さをより直感的に解釈するためには、距離行列による数字としての情報よりも、グラフ等を用いた視覚的な情

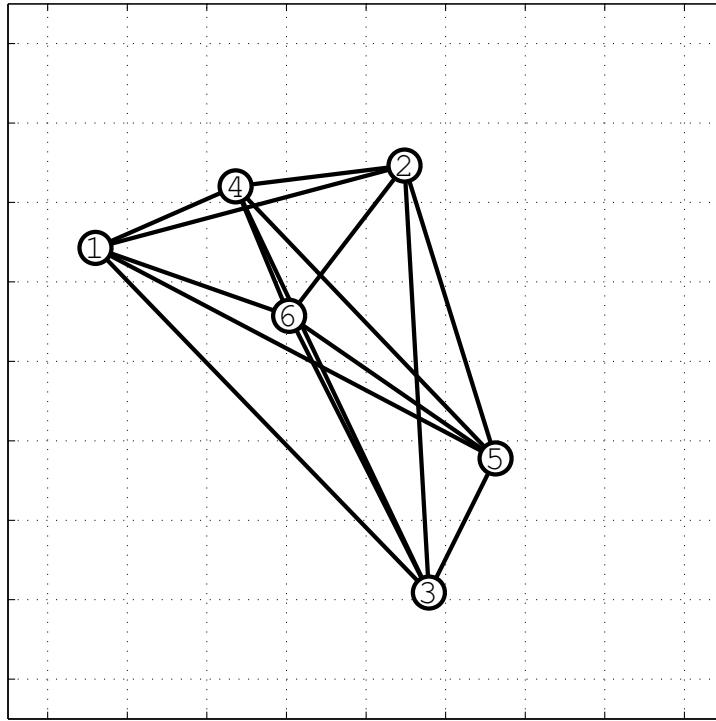


図 4.3: 式 (4.12) における各回帰ルールの位置関係

報の方が優れていると言える。そこで、各回帰ルールをノードとするネットワークに見立てて表示することを考える。式 (4.12) の回帰ルール集合の、各回帰ルールの位置関係を示すネットワークは、距離行列 D の値を用いて計算することで、図 4.3 のようになる。図 4.3 によって、例えば回帰ルール 3 と 5 は近いが他の回帰ルールとは少し離れていることや、また、回帰ルール 6 は回帰ルール集合全体の中心に位置していることなどが、視覚情報として直感的に理解できる。回帰ルールの位置関係をネットワークとして視覚的に示すことは、非常に意義があるといえる。

4.6.2 Spring Model

グラフの自動描画アルゴリズムについては古くから研究されており、有名なものとしては Spring Model [32] や Force-directed Placement[33] などがある。これらの方法を用いれば、距離行列からそれらの位置関係を適切に表示する 2 次元または 3 次元のグラフを自動的に描画できる。本研究では、アルゴリズムの簡潔性から Spring Model を用いて 2 次元グラフの自動描画を行う。Spring Model は、各頂点が

バネによって結ばれていると見立て、全てのバネの弾性エネルギーの総和が0になるように計算していく。nc-MR法における回帰ルールの位置関係の描画をするためのSpring Modelの適用について、その概要を以下に示す。

Spring Modelでは、バネを伸ばしたり縮めたりした時にバネが持つ内部エネルギーである、バネの弾性エネルギーの概念を用いている。弾性エネルギー E_{spr} は以下で計算される。

$$E_{spr} = \frac{1}{2} \sum_{i=1}^{I^*-1} \sum_{j=i+1}^{I^*} k_{ij} (s_{ij} - l_{ij})^2 \quad (4.14)$$

ただし、 k_{ij} はバネの張りの強さを表す定数で、距離行列 D の i, j 要素 d_{ij} と定数 K を用いて以下で定義する。

$$k_{ij} = K \times \frac{1}{d_{ij}} \quad (4.15)$$

また、 s_{ij} は2つの頂点 P_i, P_j 間のグラフ上での距離(距離行列の要素の値とは異なる)を、 l_{ij} は表示するグラフにおける2つの頂点 P_i, P_j 間の最も望ましい長さ(バネの自然長)をそれぞれ表す。 l_{ij} を以下で定義する。

$$l_{ij} = L_0 \times \frac{\sqrt{d_{ij}}}{\max(\sqrt{d_{ij}})} \quad (4.16)$$

ただし、 L_0 はグラフを表示するディスプレイの横幅(縦幅)を表す。

Spring modelでは、全てのバネの長さが自然長に近づくように計算していく。すなわち、 E_{spr} を目的関数として、 E_{spr} を最小化するように頂点を順に移動させる。平面上でグラフの表示を行うために、各頂点 P_i の位置を x 座標、y 座標を用いて $P_1(x_1, y_1), P_2(x_2, y_2), \dots, P_{I^*}(x_{I^*}, y_{I^*})$ とすると、式(4.14)は以下のようになる。

$$E_{spr} = \frac{1}{2} \sum_{i=1}^{I^*-1} \sum_{j=i+1}^{I^*} k_{ij} \left\{ (x_i - x_j)^2 + (y_i - y_j)^2 + l_{ij}^2 - 2l_{ij}\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \right\} \quad (4.17)$$

Spring Modelでは、すべての頂点を同時に動かすバッチ学習を行わず、バネの自然長から最もずれている頂点を1つずつ動かすオンライン学習によって、 E_{spr} を最小化していく。

自然長からのずれを表す基準として、全ての頂点について、以下で表される Δ_m を求める。

$$\Delta_m = \sqrt{\left(\frac{\partial E}{\partial x_m} \right)^2 + \left(\frac{\partial E}{\partial y_m} \right)^2} \quad (4.18)$$

Δ_m が最大となる頂点 P_m のみを動かして, E_{spr} を最小化する. 更新する頂点の初期位置の座標を $(x_m^{(0)}, y_m^{(0)})$ として, その座標を以下に従って更新していく.

$$x_m^{(t+1)} = x_m^{(t)} + \delta x, \quad y_m^{(t+1)} = y_m^{(t)} + \delta y, \quad t = 0, 1, 2, \dots \quad (4.19)$$

$\delta x, \delta y$ の計算には Newton-Raphson 法を用い, 以下の式を解くことで計算できる.

$$\begin{pmatrix} \frac{\partial^2 E_{spr}}{\partial x_m^2} & \frac{\partial^2 E_{spr}}{\partial x_m \partial y_m} \\ \frac{\partial^2 E_{spr}}{\partial x_m \partial y_m} & \frac{\partial^2 E_{spr}}{\partial y_m^2} \end{pmatrix} \begin{pmatrix} \delta x \\ \delta y \end{pmatrix} = - \begin{pmatrix} \frac{\partial E_{spr}}{\partial x_m} \\ \frac{\partial E_{spr}}{\partial y_m} \end{pmatrix} \quad (4.20)$$

なお, 具体的な微分値の計算については付録に掲載した. 式 (4.20) を解いて $\delta x, \delta y$ を求め, Δ_m の値が充分小さくなるまで (x_m, y_m) の更新を繰り返す.

Spring Model の処理手順は以下のようになる.

step1 全てのバネにおいて l_{ij} および $k_{[ij]}$ を計算し, 各頂点の初期位置を決定する.

step2 全ての頂点 P_i において Δ_i を計算する.

$\max(\Delta_i) < \epsilon$ となれば反復を終了する.

step3 Δ_i が最大となった頂点 P_m について, 以下を繰り返す.

step3.1 Newton-Raphson 法を用いて, 頂点 P_m における座標の更新量 δ_x, δ_y を計算して, (x_m, y_m) を更新する.

step3.2 Δ_m を計算する.

$\Delta_m < \epsilon$ となれば反復を終了し, step2 へ戻る.

4.7 nc-MR 法の処理手順

nc-MR 法の全体での処理手順を以下に示す.

step1 R^* の決定

$R = 1, 2, \dots$ として 4 層パーセプトロンを学習し, モデル選択基準によって, 最適隠れユニット数 R^* を決定する.

step2 係数ベクトルの量子化

パーセプトロンの学習によって得られた係数ベクトルを, K 平均法を用いてベクトル量子化する. $I = 1, 2, \dots$ としてベクトル量子化を行い, モデル選択基準によって最適ルール数 I^* を決定する.

step3 条件部の復元

決定木生成プログラムを用いて判別問題を解き，回帰ルールの条件部を復元する。

step4 回帰ルールの位置関係の表示

Spring Model を用いて，回帰ルールの位置関係を表示する。

第 5 章

多層パーセプトロンの重み共有

5.1 共有重み構造の多層パーセプトロン

多層パーセプトロンの学習能力は強力なので、十分な数の中間ユニットがあれば、非線形性を持ったデータに対しても柔軟に対応することが可能である。特に RF5 法や RF6.4 法は、パーセプトロンの学習によって得られる出力式を、多変量多項式という可読性のよい表現で獲得する。しかし、変量の数や多層パーセプトロンの中間ユニット数（多項式の項数）が多い場合には、多変量多項式は、重回帰や nc-MR 法に比べて可読性は悪くなる。

例えば、説明変数 $x = (x_1, x_2, \dots, x_{10})$ と目的変数 y の間に、以下のようないくつかの関係を持つデータについて考える。

$$y = 1 + 2x_1^2 x_2^3 x_3 + 3x_3^2 x_4^{0.5} (x_5 x_6)^{-1} \quad (5.1)$$

ただし、 x_7, \dots, x_{10} は回帰式中に現れない不要変数である。式 (5.1) は、非常に可読性のよい多変量多項式となっている。しかし、このデータに対して RF5 法を適用してこの多変量多項式を復元すると、以下のような非常に可読性の悪い多変量多項式が得られてしまう。

$$\begin{aligned} y = & 1.032 + 1.993 x_1^{2.012} x_2^{3.102} x_3^{0.982} x_4^{0.002} x_5^{0.014} x_6^{-0.102} x_7^{-0.083} x_8^{0.092} x_9^{0.012} x_{10}^{-0.122} \\ & + 3.093 x_1^{0.037} x_2^{-0.064} x_3^{2.102} x_4^{0.503} x_5^{-0.943} x_6^{-1.011} x_7^{0.122} x_8^{-0.053} x_9^{0.020} x_{10}^{-0.010} \end{aligned} \quad (5.2)$$

式 (5.2) の可読性が悪い最も大きな理由は、本来は不要であった変数が式中に現れていることである。通常、データにはノイズが含まれるために、不要変数に対するパラメータが完全に 0 になることはない。従って、不要なパラメータを削除することができれば、可読性の向上させることができる。不要パラメータの削除 (weight pruning) を行う最も簡単な方法は、パラメータの値の中で 0 に近いものを 0 として

しまって，そのパラメータは0のまま更新しないようにしてネットワークを再学習することである．しかし，その方法を用いる場合には，あるパラメータが不要であるかどうかを判別する基準が必要となる．データが多変量の場合には，その判別は全ての説明変数の組合せごとに必要となるので好ましくない．ネットワークの不要パラメータを削除する他の方法としては，LeCunらによるOBD(Optimal Brain Damage)法[34]などがある．OBD法では，ヘッセ行列の情報をを利用して，ネットワークの不要重みを削除する．

本章では，後に不要なパラメータを効率良く削除することも考慮して，ネットワークの重み共有法について考える．重み共有法とは，重みが取り得る値を制限し，数種類の値しか取れないようにする方法である．それにより，可読性の良い結果を望めるだけでなく，データに含まれるノイズへのオーバーフィットの軽減により，汎化性能の向上も期待できる．また，値が0に近く不要となる重みは，1つの集合(クラスタ)に集められるので，効率良く不要パラメータを削除することが可能となる．例えば，RF5法によって得られた式(5.2)に対して，共有重み構造で学習して得られる多変量多項式は以下となった．

$$\begin{aligned} y = & 1.031 + 2.008x_1^{2.083}x_2^{3.071}x_3^{0.971}x_4^{0.044}x_5^{0.044}x_6^{0.044}x_7^{0.044}x_8^{0.044}x_9^{0.044}x_{10}^{0.044} \\ & + 3.072x_1^{0.044}x_2^{0.044}x_3^{2.083}x_4^{0.513}x_5^{-1.002}x_6^{-1.002}x_7^{0.044}x_8^{0.044}x_9^{0.044}x_{10}^{0.044} \end{aligned} \quad (5.3)$$

これだけではまだ可読性のよい結果とはいえないが，なんらかの基準によって，式(5.3)の共有重みの値が0.044となるクラスタが不要であると分かれば，そのクラスタに属する重みを一挙に削除して，以下のような結果が得られる．

$$y = 1.034 + 2.007x_1^{2.080}x_2^{3.073}x_3^{0.971} + 3.051x_3^{2.080}x_4^{0.501}x_5^{-1.006}x_6^{-1.006} \quad (5.4)$$

これは，元の式(5.1)に極めて近く，可読性の良い式が得られているといえる．

ネットワークの重み共有は，ネットワークの複雑さを軽減し，簡素で汎化性能の高い結果を得るために手法として，広く用いられている．重み共有法は，元々は上記の例のように重みの集合をハードにクラスタリングするのではなく，ソフトにクラスタリングを行うsoft weight sharingとして，NowlanとHintonによって提案された[35][2]．また，soft weight sharingは効率良く大規模集積回路(VLSI)を実装するための手法として，多層パーセプトロンにも適用された[36]．ハードクラスタリングでは，同じクラスタに属する重みは全く同じ値しか取らないが，ソフトクラスタリングでは，同じクラスタに属する重みは，ある正規分布に従って近

い値を取るように学習される。更に、最近では、重み共有はオセロの評価関数を人工ニューラルネットで学習する際にも利用されている [37]。

可読性のよい結果を得るために、ソフトクラスタリングではなくハードクラスタリングが有効である。そこで、本研究ではハードクラスタリングによるパーセプトロンの重み共有に着目した。共有重みの値の決定、および各重みをどのクラスタに所属させるかを決める重みクラスタリング問題には多くの局所解が存在すると思われる。そこで、クラスタ数を増減しながらボトムアップとトップダウンの双方向にクラスタリングを行うことにより、大域最適な解（クラスタ数、クラスタ構成、共有重み）を探索する重み双方向クラスタリング学習法（BCW 法：bidirectional clustering of weights）[14] が提案された。各重みのクラスタリングにおいてクラスタの併合基準と分割基準が必要となるが、BCW 法では、ヘッセ行列を用いた最適基準を導入している。BCW 法はリカレントパーセプトロンを含む一般のニューラルネットに適用可能であるが、本研究では RF5 法、RF6 法における多層パーセプトロンへの適用を想定している。本章では、BCW 法の基本枠組み、及び本研究で加えた改良点、RF5 法と RF6.4 法に適用した際の処理手順について説明する。

5.2 重み双方向クラスタリング学習法：BCW 法

5.2.1 重み共有法の基本枠組み

パーセプトロンの全重みから成るベクトルを $w = (w_1, \dots, w_d, \dots, w_M)^T$ 、最小化すべき誤差関数を $E(w)$ とする。 S_g を g 番目のクラスタ（同じ共有重みを持つ重みの集合）とし、 G 個のクラスタから成るクラスタ集合を $\Omega(G) = \{S_1, \dots, S_g, \dots, S_G\}$ とする。ただし、ここでは集合 S_g の要素は重みの番号とする。各クラスタは空ではなく、互いに排他的で、すべての和集合が全重みをカバーする。すなわち、 $S_g \neq \emptyset$ 、 $S_g \cap S_{g'} = \emptyset$ ($g \neq g'$) および $S_1 \cup \dots \cup S_G = \{1, \dots, M\}$ である。また、クラスタ集合 $\Omega(G)$ に対応する共有重みベクトルを $u = (u_1, \dots, u_g, \dots, u_G)^T$ とする。ただし、 u_g は共有重みであるので、番号 m が $m \in S_g$ である重み w_m はすべて同じ値 $w_m = u_g$ となる。さらに、クラスタ集合 $\Omega(G)$ に対応する構造のパーセプトロンを学習して得られる共有重みの推定値を \hat{u} とする。

共有重み構造を持つパーセプトロンの簡単な例と、このときのパラメータの値を以下に示す。

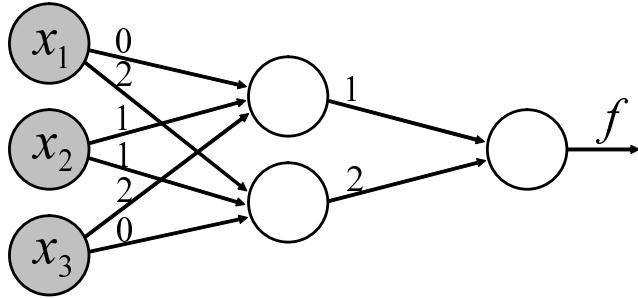


図 5.1: 共有重み構造を持つパーセプトロンの例

$$M = 8, \quad G = 3,$$

$$\mathbf{w} = (w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8)^T = (1, 2, 0, 2, 1, 1, 2, 0)^T,$$

$$S_1 = \{3, 8\}, \quad S_2 = \{1, 5, 6\}, \quad S_3 = \{2, 4, 7\}, \quad \Omega(3) = \{S_1, S_2, S_3\},$$

$$\mathbf{u} = (u_1, u_2, u_3)^T = (0, 1, 2)^T$$

次に，重みベクトル w と共有重みベクトル u の関係を考える。 e_m^M を第 m 成分のみ 1 で他は 0 の M 次元ベクトルとするとき， w は $M \times G$ の変換行列 A を用いて以下のように表せる．

$$\mathbf{w} = \mathbf{A}\mathbf{u}, \quad \mathbf{A} = \left(\sum_{m \in S_1} e_m^M, \dots, \sum_{m \in S_G} e_m^M \right) \quad (5.5)$$

図 5.1 の例では，変換行列は以下のようになる．

$$\mathbf{e}_1^8 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{e}_2^8 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \dots, \quad \mathbf{e}_8^8 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \quad (5.6)$$

$$\mathbf{A}\mathbf{u} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 0 \\ 2 \\ 1 \\ 1 \\ 2 \\ 0 \end{pmatrix} = \mathbf{w} \quad (5.7)$$

変換行列 A はクラスタ集合 $\Omega(G)$ と 1 対 1 に対応している。したがって解くべき基本問題は、誤差 $E(Au)$ を最小にするクラスタ集合 $\Omega(G^*)$ を求める問題として定式化できる。ただし、 G^* は指定されたクラスタ数である。

5.2.2 重みボトムアップクラスタリング

クラスタ集合 $\Omega(G)$ の中から適切な 2 クラスタを選択して併合し、 $\Omega(G-1)$ を得る一段ボトムアップクラスタリングの手順を説明する。

$\Omega(G)$ の中の 2 個のクラスタ S_g と $S_{g'}$ が併合されて $\tilde{S}_g = S_g \cup S_{g'}$ となるようにする。併合すべき 2 クラスタは併合に伴う誤差関数の増加が最小になるように選択したい。まず、全組合せで各パーセプトロンを学習して結果を調べる方法が考えられるが、組合せの総数が ${}_G C_2$ となるので、重みの数 G が多い場合には現実的でない。あるいは、共有重みベクトルに着目し、その要素間の距離 $(u_g - u_{g'})^2$ が最小の対を選択する方法も考えられるが、その対の併合が誤差関数の増加を最小にする保障はない。

そこで、併合に伴う誤差の増加量に基づく併合基準を考える。クラスタ S_g と $S_{g'}$ を併合した場合の誤差関数 $E(Au)$ の増加量は以下のように近似できる。ただし、 $w = Au$ とする。

$$E(A(u + \Delta u)) - E(Au) \approx g(w)^T A \Delta u + \frac{1}{2} \Delta u^T A^T H(w) A \Delta u \quad (5.8)$$

ここで、 $g(w)$ は勾配ベクトル、 $H(w)$ はヘッセ行列を表す。併合後の共有重みは同じになる制約から以下を得る。

$$(u + \Delta u)^T e_g^G = (u + \Delta u)^T e_{g'}^G \quad (5.9)$$

式(5.9)を制約条件として式(5.8)の右辺を最小にする変分 Δu をラグランジュ乗数法を用いて求める。式(5.8)と式(5.9)を用いたラグランジュ乗数法の目的関数 L は以下となる。

$$L = g(w)^T A \Delta u + \frac{1}{2} \Delta u^T A^T H(w) A \Delta u - \lambda (u + \Delta u)^T (e_g^G - e_{g'}^G) \quad (5.10)$$

ここで λ はラグランジュ乗数を表す。式(5.10)を Δu で微分し 0 とおいて解けば以下を得る。

$$\begin{aligned} & A^T g + A^T H(w) A \Delta u - \lambda (e_g^G - e_{g'}^G) = 0 \\ \iff & \Delta u = \lambda C_A d - C_A f_A, \\ & C_A = (A^T H(w) A)^{-1}, \quad f_A = A^T g(w), \quad d = e_g^G - e_{g'}^G \end{aligned} \quad (5.11)$$

式(5.11)を式(5.9)に代入し λ で解けば以下を得る.

$$\begin{aligned} & (\mathbf{u} + \lambda \mathbf{C}_A \mathbf{d} - \mathbf{C}_A \mathbf{f}_A)^T \mathbf{d} = 0 \\ \iff \quad & \lambda = \frac{\mathbf{f}_A^T \mathbf{C}_A \mathbf{d} - \mathbf{u}^T \mathbf{d}}{\mathbf{d}^T \mathbf{C}_A \mathbf{d}} \end{aligned} \quad (5.12)$$

また、式(5.11)を式(5.10)に代入して2倍すれば以下となる.

$$\begin{aligned} 2L &= 2\mathbf{g}(\mathbf{w})^T \mathbf{A}(\lambda \mathbf{C}_A \mathbf{d} - \mathbf{C}_A \mathbf{f}_A) + (\lambda \mathbf{C}_A \mathbf{d} - \mathbf{C}_A \mathbf{f}_A)^T \mathbf{A}^T \mathbf{H}(\mathbf{w}) \mathbf{A}(\lambda \mathbf{C}_A \mathbf{d} - \mathbf{C}_A \mathbf{f}_A) \\ &= \lambda^2 \mathbf{d}^T \mathbf{C}_A \mathbf{d} - \mathbf{f}_A^T \mathbf{C}_A \mathbf{f}_A \end{aligned} \quad (5.13)$$

ここで、 $\mathbf{f}_A^T \mathbf{C}_A \mathbf{f}_A$ はクラスタの併合に関わらず一定なので、除外してもよい. これに式(5.12)を代入することにより、クラスタ S_g と $S_{g'}$ を併合した場合の誤差関数の増加量に基づく併合基準 $Dissim(S_g, S_{g'})$ は以下のようになる.

$$\begin{aligned} Dissim(S_g, S_{g'}) &= \frac{(\mathbf{f}_A^T \mathbf{C}_A \mathbf{d} - \mathbf{u}^T \mathbf{d})^2}{\mathbf{d}^T \mathbf{C}_A \mathbf{d}}, \\ \mathbf{C}_A &= (\mathbf{A}^T \mathbf{H}(\mathbf{w}) \mathbf{A})^{-1}, \quad \mathbf{f}_A = \mathbf{A}^T \mathbf{g}(\mathbf{w}), \quad \mathbf{d} = \mathbf{e}_g^G - \mathbf{e}_{g'}^G \end{aligned} \quad (5.14)$$

$Dissim(S_g, S_{g'})$ は常に正で、値が大きい程その併合が望ましくないので、式(5.14)をクラスタ S_g と $S_{g'}$ の間の非類似性と呼ぶ. $Dissim(S_g, S_{g'})$ が最小のクラスタ対を選択して併合するのがクラスタ集合 $\Omega(G)$ から $\Omega(G-1)$ を得る一段ボトムアップクラスタリングである. なお、併合後はクラスタ集合 $\Omega(G-1)$ で構造化されたパーセプトロンを再度学習する.

オリジナルのBCW法では、十分に学習されたパラメータ $\hat{\mathbf{u}}$ に対してのみクラスタ併合を行うことを考え、極値条件 $\mathbf{A}^T \mathbf{g}(\mathbf{A}\hat{\mathbf{u}}) = 0$ を仮定して併合基準が導出されていた. しかし、本研究では、パラメータの学習が十分に進んでいない状況でもクラスタ併合基準が適用できるように、極値条件を仮定しないで導出した. その影響により、オリジナルのBCW法に比べて導出課程の式が複雑になっているが、最終的なクラスタ併合の基準としては、オリジナルのBCW法と比べてもあまり変わらないシンプルな式が得られた.

5.2.3 重みトップダウンクラスタリング

クラスタ集合 $\Omega(G)$ の中から適切なクラスタを1つ選択して分割し、 $\Omega(G+1)$ を得る一段トップダウンクラスタリングの手順を説明する.

$\Omega(G)$ の中のクラスタ S_g が2つのクラスタ S'_g と S_{G+1} に分割されるとする. すなわち、 $S_g = S'_g \cup S_{G+1}$ となる. このとき、分割後の変換行列 \mathbf{B} は $M \times (G+1)$ で以

下で与えられる .

$$\mathbf{B} = \left(\sum_{m \in S_1} \mathbf{e}_m^M, \dots, \sum_{m \in S'_g} \mathbf{e}_m^M, \dots, \sum_{m \in S_G} \mathbf{e}_m^M, \sum_{m \in S_{G+1}} \mathbf{e}_m^M \right) \quad (5.15)$$

分割すべきクラスタは誤差関数の減少が最大になるように選択したいので , 誤差関数の減少量に着目する . このとき , 分割直後の共有重み \mathbf{v} は $(G + 1)$ 次元で $\mathbf{v} = (\mathbf{u}^T, u_g)^T$ となる .

例えば図 5.1 の例で $S_2 = \{1, 5, 6\}$ を $S'_2 = \{1, 6\}$ と $S_4 = \{5\}$ に分割するとすれば , 変換行列 \mathbf{B} 及び共有重み \mathbf{v} は以下となる .

$$\mathbf{B} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & \mathbf{0} & 0 & \mathbf{1} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{v} = \begin{pmatrix} 0 \\ 1 \\ 2 \\ 1 \end{pmatrix} \quad (5.16)$$

分割に伴う誤差関数の減少量は以下のように近似できる .

$$E(\mathbf{B}\mathbf{v}) - E(\mathbf{B}(\mathbf{v} + \Delta\mathbf{v})) \approx - \left(\mathbf{g}(\mathbf{B}\mathbf{v})^T \mathbf{B} \Delta\mathbf{v} + \frac{1}{2} \Delta\mathbf{v}^T \mathbf{B}^T \mathbf{H}(\mathbf{B}\mathbf{v}) \mathbf{B} \Delta\mathbf{v} \right) \quad (5.17)$$

このとき , 式 (5.17) の右辺を最大にする変分 $\Delta\mathbf{v}$ を求める . 式 (5.17) を $\Delta\mathbf{v}$ で微分して 0 とおいて解けば以下を得る .

$$\begin{aligned} \mathbf{B}^T \mathbf{g}(\mathbf{B}\mathbf{v}) + \mathbf{B}^T \mathbf{H}(\mathbf{B}\mathbf{v}) \mathbf{B} \Delta\mathbf{v} &= 0 \\ \iff \Delta\mathbf{v} &= -\mathbf{C}_B^{-1} \mathbf{f}_B \\ \mathbf{C}_B &= \mathbf{B}^T \mathbf{H}(\mathbf{B}\mathbf{v}) \mathbf{B}, \quad \mathbf{f}_B = \mathbf{B}^T \mathbf{g}(\mathbf{B}\mathbf{v}) \end{aligned} \quad (5.18)$$

これを式 (5.17) に代入することにより , クラスタ S_g を S'_g と S_{G+1} に分割した場合の誤差関数の減少量に基づく分割基準 $GenUtil(S_g, S_{G+1})$ は以下のようになる .

$$\begin{aligned} GenUtil(S_g, S_{G+1}) &= \mathbf{f}_B^T \mathbf{C}_B \mathbf{f}_B \\ \mathbf{C}_B &= (\mathbf{B}^T \mathbf{H}(\mathbf{B}\mathbf{v}) \mathbf{B})^{-1}, \quad \mathbf{f}_B = \mathbf{B}^T \mathbf{g}(\mathbf{B}\mathbf{v}) \end{aligned} \quad (5.19)$$

式 (5.19) を基に , クラスタ S_g を S'_g と S_{G+1} に分割する是非を判断する . $GenUtil(S_g, S_{G+1})$ は常に正で , 値が大きい程好ましい分割となる .

オリジナルの BCW 法では，クラスタ併合基準の導出と同様に，十分に学習されたパラメータ \hat{u} に対してクラスタ分割を行うことを考え， \hat{u} が極値条件を満たすことを仮定して分割基準が導出されていた。しかし，本研究では分割前のパラメータの学習が十分に進んでいない状況でも分割基準が適用できるように，極値条件を仮定しないで導出した。

クラスタの併合においては，併合する 2 つのクラスタを選べば，その併合の仕方は 1 通りしかなかったのに対し，クラスタの分割においては，分割の仕方の総数は，あるクラスタ S_g の要素数が n_g のとき $(2^{n_g} - 2)/2 = 2^{n_g-1} - 1$ となる。従ってそれらをすべて調べると組合せ爆発に陥り，現実的ではない。そこで，以下では組合せ爆発に陥らないための分割方法について考える。

分割方法 1 クラスタ S_g の中から 1 つの重み $d \in S_g$ だけを分割する。このとき，クラスタ S_g 内の分割方法は n_g 通りとなる。

分割方法 2 クラスタ S_g に属する重みを，その勾配 $g(Bv)$ の順に並び替え，勾配が大きいクラスタと小さいクラスタに分割する。式(5.19)の値は勾配の大きさに大きく左右されるため，勾配の順に並べて分割することで， $GenUtil(S_g, S_{G+1})$ がより大きな値を取ることを期待できる。このとき，クラスタ S_g 内の分割方法は $n_g - 1$ 通りとなる。

分割方法 3 クラスタ S_g に属する重みに対して，その重みの重み共有無しでの学習結果に対して K 平均法を適用して，2 つのクラスタに分割する。重み共有無しでの学習のパラメータ値が異なるほど，重み共有時に違うクラスタになりやすいので，この方法によってうまく分割されることが期待できる。このとき，クラスタ S_g 内の分割方法は 1 通りに定まる。

上記の 3 つの分割方法のうち，オリジナルの BCW 法では分割方法 1 のみを採用していた。しかし，分割方法 1 では分割したクラスタのうち片方は要素が 1 つのクラスタとなってしまうので，好ましくない。そこで，本研究では新たに分割方法 2, 3 を考えた。パーセプトロンの再学習などに必要な計算コストに比べて，上記の 3 つの分割方法を適用して $GenUtil(S_g, S_{G+1})$ を計算するコストはかなり小さいため，3 つのうちのどれか 1 つの方法だけを用いる必要はない。すなわち，上記の 3 つの分割方法を全て試した上で，その中で最も $GenUtil(S_g, S_{G+1})$ が大きくなる分割を選べばよい。

以上が，クラスタ集合 $\Omega(G)$ から $\Omega(G+1)$ を得る一段トップダウンクラスタリングである。なお，分割後はクラスタ集合 $\Omega(G+1)$ で構造化されたパーセプトロンで再度学習する。

5.2.4 重み双方向クラスタリング

一般に，重みクラスタリング問題には多くの局所解が存在するといえる。従つて，ボトムアップまたはトップダウンの一方向クラスタリングでは大域最適でない局所解に陥ってしまうことが考えられる。そこで，ボトムアップとトップダウンの双向向の重みクラスタリングを繰り返して，大域最適解を探索する。

BCW 法の動作パラメータには，クラスタ数 G と双向向クラスタリングの探索幅 h があり，ともにユーザが指定する。最適なクラスタ数 G^* が不明のときは，最小の汎化誤差を与える G を選択する方法が考えられる。探索幅 h が大きい程，探索範囲は広がるが，計算量が増大する。しかし，問題に応じて臨界の h が存在し，それ以上大きくしても解品質改善の面では効果がないと思われる。

オリジナルの BCW 法では，それぞれの G の値に対して別々に BCW を走らせていました。すなわち， $G = 2, 3, \dots$ を初期クラスタ数として，探索幅 h で双向向に探索していく。そして，最も良い汎化性能が得られたときの G を最適クラスタ数 G^* として選択する。しかし，この方法だと計算量も飛躍的に増大してしまう。そこで，本研究では双向向クラスタリングの方法に改良を加えた。改良した BCW 法では，初期クラスタ数を $G = 2$ に固定し， $G = 2$ から $G = 2 + h$ の範囲で BCW 法を走らせて，その中で最も汎化性能のよいときの G を G^* として選択する。

$G = 2$ での初期クラスタの構造を決めるのに，オリジナルの BCW 法では再学習無しのマルチステップボトムアップクラスタリングを用いていた。これもかなりの計算量となるので，本研究で改良した BCW 法では，この代わりにスカラー量子化を用いて初期クラスタの構造を決定する。それにより，マルチステップボトムアップクラスタリングの処理を省き，計算量を少なくすることができる。スカラー量子化を行うための方法としては，再び K 平均法を採用する。

本研究で改良した BCW 法の処理手順を以下に示す。クラスタ集合 $\Omega(G)$ の変換行列の種類は有限のため，BCW 法は常に収束するといえる。

step1 パーセプトロンを学習し，初期クラスタ集合 $\Omega(M)$ を得る。

step2 $\Omega(M)$ に対してスカラー量子化することにより， $\Omega_1(2)$ を得る。 $\Omega_1(2)$ の変換行列 $A^{(0)}$ を記憶する。 $t \leftarrow 1$ とする。

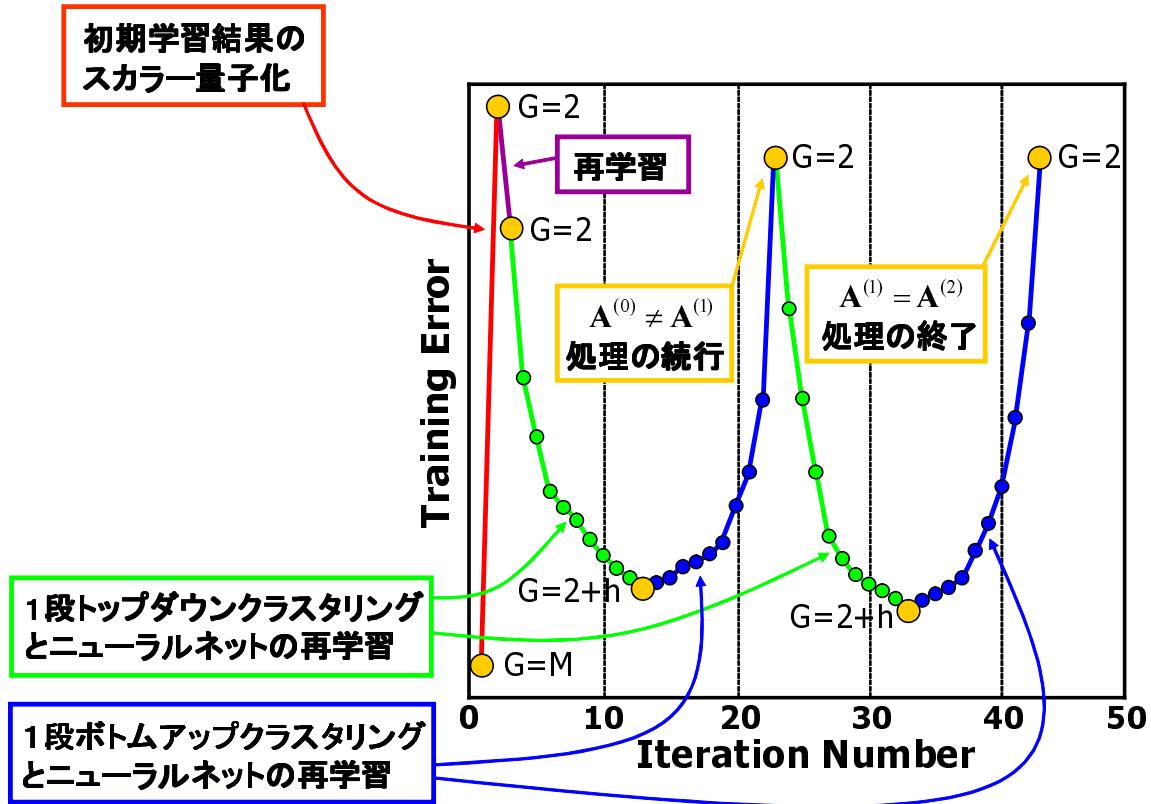


図 5.2: BCW 法の動作の様子

step3 1段トップダウンクラスタリングをパーセプトロンの再学習と共に $\Omega_1(2)$ から $\Omega(2+h)$ まで繰り返す。それぞれの G に対して性能が向上した場合は、そのクラスタ構造及び重みの値を記憶する。

step4 1段ボトムアップクラスタリングをパーセプトロンの再学習と共に $\Omega(2+h)$ から $\Omega_2(2)$ まで繰り返す。それぞれの G に対して性能が向上した場合は、そのクラスタ構造及び重みの値を記憶する。また、 $\Omega_2(2)$ の変換行列 $A^{(t)}$ を記憶する。

step5 もし $A^{(t)}$ がそれ以前の変換行列 $A^{(t-1)}, \dots, A^{(0)}$ のどれかと一致すれば、それぞれの G に対して最も性能が良かったときのクラスタ集合 $\Omega(G)$ を最終的な解として終了する。一致しなければ、 $t \leftarrow t + 1$, $\Omega_1(2) \leftarrow \Omega_2(2)$ として Step3 へ戻る。

5.3 多変量多項式回帰法への適用

5.3.1 共有重み構造での勾配の計算

多変量多項式回帰法 (RF5 法 , RF6.4 法) において , 多層パーセプトロンの学習には BPQ 法を用いるが , その学習の際に , 勾配ベクトルを計算する必要がある . 共有重み構造における共有重み \tilde{w} の勾配 $g(\tilde{w})$ は , 重みを共有しない場合の勾配 $g(w)$ と変換行列 A を用いることにより , 以下のように計算することが可能である .

$$g(\tilde{w}) = Ag(u) = AA^T g(w) \quad (5.20)$$

式 (5.20) を用いれば , 共有重み構造での勾配を別途計算する必要がないため , 非常に簡単に共有重み構造での学習を実装することが可能である .

5.3.2 RF5 法への適用

BCW 法を RF5 法へ適用した場合の処理手順を以下に示す . なお , 最適クラスタ数 G^* の選択については , 最適隠れユニット数 J^* の選択と同様に , 交差検証法または BIC を最適モデル選択基準として用いることができる .

step1 最適な項数の決定

重み共有無しで $J = 1, 2, \dots$ として 3 層パーセプトロンを学習し , モデル選択基準によって , 最適隠れユニット数 J^* を決定する .

step2 重み共有と最適クラスタ数の決定

$J = J^*$ として , BCW 法により共有重み構造で 3 層パーセプトロンを学習する . モデル選択基準によって , 最適クラスタ数 G^* を決定する .

step3 重み枝刈り

$J = J^*, G = G^*$ のもとで学習して得られた結果に対して , 共有重みの値が最も小さいクラスタに属する重みを 0 に固定して再度共有重み構造のパーセプトロンを学習する . モデル選択基準による汎化性能が向上した場合は , 次に共有重みの値が小さいクラスタに属する重みを 0 に固定して再度学習するという過程を , 汎化性能が劣化するまで繰り返す . 最も良い汎化性能が得られたときの結果を最終結果として , 処理を終了する .

5.3.3 RF6.4 法への適用

BCW 法を RF6.4 法へ適用した場合の処理手順を以下に示す .

step1 J^*, R^* の決定

重み共有無しで $J = 1, 2, \dots, R = 1, 2, \dots$ として 4 層パーセプトロンを学習し、モデル選択基準によって、最適隠れユニット数 J^*, R^* を決定する。

step2 重み共有と最適クラスタ数の決定

$J = J^*, R = R^*$ として、BCW 法により共有重み構造で 4 層パーセプトロンを学習する。モデル選択基準によって、最適クラスタ数 G^* を決定する。

step3 重み枝刈り

$J = J^*, R = R^*, G = G^*$ のもとで学習して得られた結果に対して、共有重みの値が最も小さいクラスタに属する重みを 0 に固定して再度共有重み構造のパーセプトロンを学習する。モデル選択基準による汎化性能が向上した場合は、次に共有重みの値が小さいクラスタに属する重みを 0 に固定して再度学習するという過程を、汎化性能が劣化するまで繰り返す。最も良い汎化性能が得られたときの結果を最終結果として、処理を終了する。

step4 係数ベクトルの量子化

重み枝刈り後に得られた係数ベクトルを、K 平均法を用いてベクトル量子化する。 $I = 1, 2, \dots$ としてベクトル量子化を行い、モデル選択基準によって最適ルール数 I^* を決定する。

step5 条件部の復元

決定木生成プログラムを用いて判別問題を解き、回帰ルールの条件部を復元する。

第 6 章

人工データを用いた実験による検証

人工データを用いた実験により，RF5 法，RF6.4 法，nc-MR 法，及び RF5 法と RF6.4 法に BCW 法を適用した場合に，元の数法則をデータから復元できるかどうかを検証する．また，モデル選択の妥当性についても考察する．

6.1 多変量多項式回帰法 (RF5 法，RF5+BCW 法)

6.1.1 実験における諸設定

実験データ

多変量多項式回帰の回帰式として，式 (6.1) を考える．

$$y = 2 - 3x_1^{1.5}x_2^2x_3^{1.5} + x_2^{0.5}x_3^{0.5}x_4^{1.5}x_5^{1.5} \quad (6.1)$$

量的説明変数として，回帰式に現れる x_1, \dots, x_5 の他に，回帰に全く寄与しない 10 個の不要な説明変数 x_6, \dots, x_{15} を加えた．データ生成においては，各説明変数 x_k の値は $(0, 1)$ の範囲の一様乱数とし，対応する目的変数 y の値を，式 (6.1) を用いて算出した．サンプル数は $N = 300$ で，教師信号 y には正規ノイズを加えた．ノイズの平均を 0，標準偏差を 0.1, 0.3, 0.5 と変化させて出力に加え，それぞれのデータごとに実験を行った．以降では，ノイズの標準偏差を $nstd$ と表記する．

パラメータの設定と終了条件

共有重み構造無しでの学習では，重み w_{jk} の初期値は $(-1, 1)$ の一様乱数とし，重み w_j の初期値は 0，バイアス項 w_0 の初期値は教師信号の平均値とした．

BCW 法による共有重み構造での学習では，共有重み構造の適用を入力・中間層間の重み（多項式の指數部分） w_{jk} に限定した．双方向クラスタリングの初期クラスタ数は $G = 2$ とし，探索幅は $h = 10$ とした．

パーセプトロンの学習は，勾配ベクトルのどの成分も 10^{-5} 未満となったとき，または学習の繰り返し回数が 20,000 回を超えたとき停止させた。

モデル選択の方法としては BIC と交差検証法を用い， $J = 1, \dots, 5$ としてそれぞれの場合の交差検証誤差と BIC の値を比較した。同様に G の値ごとにも比較した。

6.1.2 実験結果

まず，全サンプルを用いて， J の値ごとにパラメータの初期値を変えて 100 回ずつ学習を行った。このとき $BIC(J)$ の値が最小となった回数を表 6.1 に示す。また， J の値ごとの $BIC(J)$ の最小値を表 6.2 に示す。

表 6.1: $BIC(J)$ が最小となった回数 (RF5 法，人工データ)

J	1	2	3	4	5	合計
nstd = 0.1	0	94	6	0	0	100
nstd = 0.3	0	93	7	0	0	100
nstd = 0.5	0	81	19	0	0	100

表 6.2: $BIC(J)$ の最小値 (RF5 法，人工データ)

J	1	2	3	4	5
nstd = 0.1	860.7	242.7	276.1	306.0	341.9
nstd = 0.3	911.6	579.5	605.4	633.3	666.2
nstd = 0.5	1384.0	709.5	749.0	779.9	809.5

同様に，10-fold 交差検証によるモデル選択を行った。 J の値ごとにパラメータの初期値を変えて 10 回ずつ交差検証を行ったときの，交差検証誤差の最小値を表 6.3 に示す。なお，10 回の交差検証でのセグメントの分割方法は全て同じとした。

表 6.3: $MSE_{CV}(J)$ の最小値 (RF5 法，人工データ)

J	1	2	3	4	5
nstd = 0.1	0.9636	0.0181	0.0284	0.0476	0.0671
nstd = 0.3	4.4716	0.1264	0.1758	0.1873	0.1918
nstd = 0.5	24.9941	0.2791	0.2973	0.3788	0.4607

次に , $J = 2$ として , BCW 法による共有重み構造でのパーセプトロンの学習を行った . パラメータの初期値を変えて 100 回学習を行った後に , BCW 法を適用した . このとき $\text{BIC}(G)$ の値が最小となった回数を表 6.4 に示し , G の値ごとの $\text{BIC}(G)$ の最小値を表 6.5 に示す .

表 6.4: $\text{BIC}(G)$ が最小となった回数 (RF5+BCW 法 , 人工データ)

G	2	3	4	5	6	7	8	9	10	11	12	合計
nstd = 0.1	0	0	96	4	0	0	0	0	0	0	0	100
nstd = 0.3	0	0	89	11	0	0	0	0	0	0	0	100
nstd = 0.5	0	0	69	16	9	6	0	0	0	0	0	100

表 6.5: $\text{BIC}(G)$ の最小値 (RF5+BCW 法 , 人工データ)

G	2	3	4	5	6	
	7	8	9	10	11	12
nstd = 0.1	1029.5	633.4	171.9	173.3	173.6	
nstd = 0.3	1021.2	670.5	506.9	507.6	508.6	
nstd = 0.5	1029.5	742.6	653.2	652.1	651.7	
	174.4	177.1	179.6	182.2	184.8	187.6
	510.6	513.0	515.6	518.3	521.1	523.9
	651.5	653.1	654.6	656.1	658.2	661.1

同様に , パラメータの初期値を変えて 10 回ずつ 10-fold 交差検証を行ったときの , 交差検証誤差の最小値を表 6.6 に示す .

表 6.6: $\text{MSE}_{\text{CV}}(G)$ の最小値 (RF5+BCW 法 , 人工データ)

G	2	3	4	5	6	
	7	8	9	10	11	12
nstd = 0.1	3.0702	0.2343	0.0125	0.0157	0.0163	
nstd = 0.3	2.9047	0.2996	0.1003	0.1042	0.1081	
nstd = 0.5	3.1108	0.4525	0.2342	0.2398	0.2416	
	0.0165	0.0172	0.0178	0.0169	0.0179	0.0180
	0.1100	0.1194	0.1172	0.1198	0.1217	0.1232
	0.2477	0.2538	0.2635	0.2700	0.2710	0.2746

また, $nstd = 0.3$ のときの, ある試行における BCW 法での誤差の変化の様子を図 6.1 に示す .

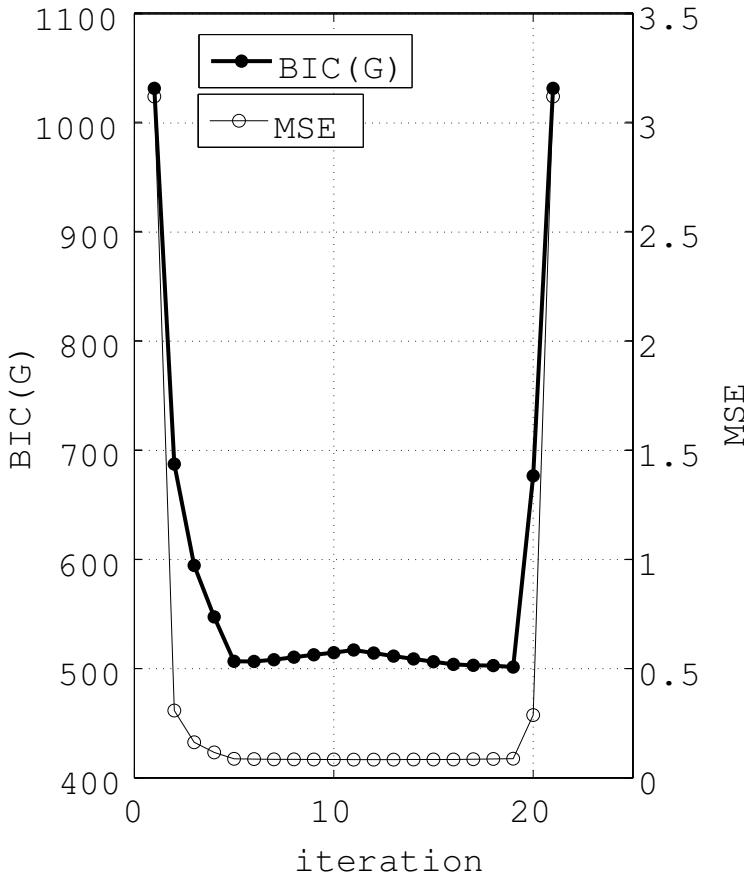


図 6.1: 誤差の変化の様子 (RF5+BCW 法, 人工データ, $nstd = 0.3$)

最後に, $nstd = 0.3$ のデータでのある試行における RF5 法の最終出力を示す . 共有重み構造無しの, $J = 2$ における通常の RF5 法での出力の式を以下に示す .

$$\begin{aligned}
 y = & 1.6819 - 2.9793x_1^{1.4956}x_2^{2.0087}x_3^{1.5071}x_4^{0.0011}x_5^{-0.0116}x_6^{-0.0055}x_7^{-0.0004} \\
 & \times x_8^{-0.0082}x_9^{-0.0025}x_{10}^{-0.0039}x_{11}^{0.0006}x_{12}^{-0.0023}x_{13}^{-0.0004}x_{14}^{0.0047}x_{15}^{0.0038} \\
 & + 1.1242x_1^{-0.0726}x_2^{0.4553}x_3^{0.4566}x_4^{1.4767}x_5^{1.4478}x_6^{-0.0280}x_7^{0.0083} \\
 & \times x_8^{-0.0264}x_9^{0.0238}x_{10}^{-0.0193}x_{11}^{0.0062}x_{12}^{-0.0203}x_{13}^{-0.0171}x_{14}^{0.0340}x_{15}^{-0.0017} \quad (6.2)
 \end{aligned}$$

BCW 法による共有重み構造での, $J = 2, G = 4$ における RF5 法の出力の式を以下に示す .

$$y = 1.9632 - 3.0116x_1^{1.4990}x_2^{1.9951}x_3^{1.4990} + 1.0183x_2^{0.4879}x_3^{0.4879}x_4^{1.4990}x_5^{1.4990} \quad (6.3)$$

6.1.3 考察

まず , BIC(J) による最適隠れユニット数の選択では , 表 6.1 より $nstd = 0.1, 0.3$ のときで 9 割以上 , $nstd = 0.5$ のときで 8 割以上の試行で正解である $J^* = 2$ を選択することができている . 数回の試行で $J = 3$ が選択されてしまっているが , このときのほとんどの試行で , $J = 2$ での学習が局所解に陥っていた . また , 表 6.2, 6.3 に示されるように , BIC(J), MSE_{CV}(J) の最小値でも $J^* = 2$ を選択することができている .

次に , BIC(G) による最適クラスタ数の選択でも , 表 6.4 より , 多くの試行で正解である $G^* = 4$ を選択できていることがわかる . しかし , データに加えたノイズが大きくなるにつれて , 正解よりも大きいクラスタ数を選択することが増えている . 特に , 表 6.5 より BIC(G) の最小値では , $nstd = 0.5$ のとき $G = 7$ を選択してしまっている . クラスタ数が 1 増えても , BIC(G) のペナルティ部分の項の値はそれほど増えないため , 多くの試行を繰り返すと , 程に BIC(G) が正解のときよりも小さくなることが起きると考えられる .

図 6.1 は双方向クラスタリングが 1 回だけ行われた場合の例であるが , 今回の人工データでは , 同様に双方向クラスタリングが 1 回行われるだけで収束することが多く見られた . 図より , BIC(4) の値はトップダウンクラスタリングの部分では値が大きく , ボトムアップクラスタリングのところで小さい値が得られていることから , トップダウンクラスタリングのときには , 正解のパラメータ値が得られていないと考えられるので , 双方向のクラスタリングがうまく働いているといえる .

式 (6.3) は最も 0 に近い共有重みの値を 0 として枝刈りした後に , パーセプトロンを再学習して得られた結果を示している . 式 (6.2) と (6.3) を比べると , BCW 法による重み共有によって , 明らかに可読性の良い式を得ることができている . 重み共有無しでも正解に近いパラメータ値が得られているが , 重み共有を適用することで , より正解に近いパラメータ値が得られていることがわかる . また , 全てのデータにおいて , 交差検証誤差が共有重みを適用したときの方が小さい値となっていることから , 可読性が良くなっただけでなく , 汎化性能も向上したといえる .

6.2 質的条件付き多変量多項式回帰法 (RF6.4法, RF6.4+BCW法)

6.2.1 実験における諸設定

実験データ

質的条件付き多変量多項式回帰の回帰ルール集合として、式(6.4)を考える。

$$\left\{ \begin{array}{ll} \text{if } q_{21} \wedge q_{31} \text{ then} \\ \quad y = 1 + 2x_1^2 x_2 x_3^{0.5} + 3x_2 x_3^{0.5} x_4^2 x_5, \\ \text{if } q_{22} \wedge q_{31} \text{ then} \\ \quad y = 2 - x_1^2 x_2 x_3^{0.5} + x_2 x_3^{0.5} x_4^2 x_5, \\ \text{if } (q_{12} \vee q_{13}) \wedge q_{32} \text{ then} \\ \quad y = -3 + x_1^2 x_2 x_3^{0.5} - x_2 x_3^{0.5} x_4^2 x_5, \\ \text{if } q_{11} \wedge q_{32} \text{ then} \\ \quad y = 2 + 3x_1^2 x_2 x_3^{0.5} + 4x_2 x_3^{0.5} x_4^2 x_5, \end{array} \right. \quad (6.4)$$

量的説明変数として x_1, \dots, x_{15} , 質的説明変数として q_1, \dots, q_4 を用い, 質的説明変数のカテゴリ数は $L_1 = 3, L_2 = L_3 = 2, L_4 = 5$ とした。ただし, 式から明らかなように, x_6, \dots, x_{15}, q_4 は回帰ルール集合に現れない不要な変数である。訓練データのサンプル数は $N = 500$ とした。各サンプルの量的変数 x_j の値は $(-1, 1)$ の範囲の一様乱数とし, 質的変数 q_k の値はランダムにどれか1つのカテゴリをとるようにした。そして, 式(6.4)のルール集合を元に最終出力を計算し, 正規ノイズを加えた。ノイズの平均を0, 標準偏差を0.1, 0.3, 0.5と変化させて出力に加え, それぞれのデータごとに実験を行った。

パラメータの設定と終了条件

共有重み構造無しでの学習では, 重み v_{jr}, v_{rkl}, w_{jk} の初期値を $(-1, 1)$ の一様乱数とした。

BCW法による共有重み構造での学習では, RF5法の場合と同様に, 共有重み構造の適用は w_{jk} に限定した。双方向クラスタリングの初期クラスタ数は $G = 2$ とし, 探索幅は $h = 10$ とした。

パーセプトロンの学習は, 勾配ベクトルのどの成分も 10^{-5} 未満となったとき, または学習の繰り返し回数が20,000回を超えたとき停止させた。

モデル選択の方法としては BIC と交差検証法を用い, $J = 1, \dots, 5, R = 1, \dots, 5, I = 1, \dots, 10$ としてそれぞれの場合の交差検証誤差と BIC の値を比較した。同様に G の値ごとにも比較した。

6.2.2 実験結果

まず, 全サンプルを用いて, J, R の値ごとにパラメータの初期値を変えて 100 回ずつ学習を行った. このとき $BIC(J, R)$ の値が最小となった回数を表 6.7~6.9 に示す. また, J, R の値ごとの $BIC(J, R)$ の最小値を表 6.10~6.12 に示す.

表 6.7: $BIC(J, R)$ が最小となった回数 (RF6.4 法, 人工データ, $nstd = 0.1$)

models	$J = 1$	$J = 2$	$J = 3$	$J = 4$	$J = 5$
R = 1	0	0	0	0	0
R = 2	0	0	0	0	0
R = 3	0	91	0	0	0
R = 4	0	9	0	0	0
R = 5	0	0	0	0	0

表 6.8: $BIC(J, R)$ が最小となった回数 (RF6.4 法, 人工データ, $nstd = 0.3$)

models	$J = 1$	$J = 2$	$J = 3$	$J = 4$	$J = 5$
R = 1	0	0	0	0	0
R = 2	0	0	0	0	0
R = 3	0	86	0	0	0
R = 4	0	14	0	0	0
R = 5	0	0	0	0	0

表 6.9: $BIC(J, R)$ が最小となった回数 (RF6.4 法, 人工データ, $nstd = 0.5$)

models	$J = 1$	$J = 2$	$J = 3$	$J = 4$	$J = 5$
R = 1	0	0	0	0	0
R = 2	0	0	0	0	0
R = 3	0	64	1	0	0
R = 4	0	29	0	0	0
R = 5	0	6	0	0	0

表 6.10: $BIC(J, R)$ の最小値 (RF6.4 法, 人工データ, $nstd = 0.1$)

models	$J = 1$	$J = 2$	$J = 3$	$J = 4$	$J = 5$
R = 1	2784.4	2803.5	2840.8	2855.7	876.6
R = 2	2039.6	2006.9	2014.4	1995.7	109.8
R = 3	2064.3	611.3	699.7	724.4	749.2
R = 4	2089.2	648.6	658.8	692.7	726.1
R = 5	2123.5	638.0	674.4	717.6	761.4

表 6.11: $BIC(J, R)$ の最小値 (RF6.4法, 人工データ, $nstd = 0.3$)

models	J = 1	J = 2	J = 3	J = 4	J = 5
R = 1	2690.8	2714.5	2738.6	2721.1	2699.1
R = 2	2174.2	2022.0	2011.2	2024.5	2070.7
R = 3	2084.7	1134.9	1184.4	1205.1	1236.9
R = 4	2113.8	1152.2	1190.5	1231.5	1265.3
R = 5	2144.9	1178.4	1213.5	1255.9	1278.2

表 6.12: $BIC(J, R)$ の最小値 (RF6.4法, 人工データ, $nstd = 0.5$)

models	J = 1	J = 2	J = 3	J = 4	J = 5
R = 1	2737.3	2764.7	2780.8	2814.0	2822.3
R = 2	2108.0	1976.9	2013.9	2032.7	2074.2
R = 3	2046.5	1370.0	1410.0	1437.7	1485.8
R = 4	2075.6	1388.4	1422.2	1455.6	1490.8
R = 5	2099.1	1417.1	1458.4	1487.0	1509.6

J, R の値ごとにパラメータの初期値を変えて 10 回ずつ交差検証を行ったときの, 交差検証誤差の最小値を表 6.13 ~ 6.15 に示す。なお, 10 回の交差検証でのセグメントの分割方法は全て同じとした。

表 6.13: $MSE_{CV}(J, R)$ の最小値 (RF6.4法, 人工データ, $nstd = 0.1$)

models	J = 1	J = 2	J = 3	J = 4	J = 5
R = 1	100.4736	90.0199	85.7472	74.8545	66.9089
R = 2	5.4107	4.1322	3.6127	2.9631	3.5083
R = 3	5.2452	0.0208	0.0236	0.0262	0.0290
R = 4	5.0917	0.0284	0.0276	0.0270	0.0265
R = 5	5.0928	0.0287	0.0278	0.0276	0.0268

表 6.14: $MSE_{CV}(J, R)$ の最小値 (RF6.4法, 人工データ, $nstd = 0.3$)

models	J = 1	J = 2	J = 3	J = 4	J = 5
R = 1	88.4172	71.6387	65.6333	52.5273	41.9117
R = 2	7.5554	4.3268	4.5803	3.2026	4.1452
R = 3	6.6075	0.1525	0.1604	0.1698	0.1734
R = 4	6.5141	0.1661	0.1701	0.1752	0.1789
R = 5	6.4588	0.1628	0.1714	0.1764	0.1815

表 6.15: $\text{MSE}_{\text{CV}}(J, R)$ の最小値 (RF6.4 法 , 人工データ, $nstd = 0.5$)

models	J = 1	J = 2	J = 3	J = 4	J = 5
R = 1	82.3794	75.3377	65.8800	61.6669	52.2455
R = 2	5.7976	2.7782	2.6083	2.2764	2.1751
R = 3	4.9548	0.3112	0.3482	0.3770	0.3816
R = 4	4.8746	0.3059	0.3571	0.3898	0.3953
R = 5	4.7123	0.2992	0.3741	0.3923	0.3920

次に , $J = 2, R = 3$ として , BCW 法による共有重み構造でのパーセプトロンの学習を行った . パラメータの初期値を変えて 100 回学習を行った後に , BCW 法を適用した . このとき $\text{BIC}(G)$ の値が最小となった回数を表 6.16 に示し , G の値ごとの $\text{BIC}(G)$ の最小値を表 6.17 に示す .

表 6.16: $\text{BIC}(G)$ が最小となった回数 (RF6.4+BCW 法 , 人工データ)

G	2	3	4	5	6	7	8	9	10	11	12	合計
nstd = 0.1	0	0	87	13	0	0	0	0	0	0	0	100
nstd = 0.3	0	0	83	12	5	0	0	0	0	0	0	100
nstd = 0.5	0	0	65	20	9	6	0	0	0	0	0	100

表 6.17: $\text{BIC}(G)$ の最小値 (RF6.4+BCW 法 , 人工データ)

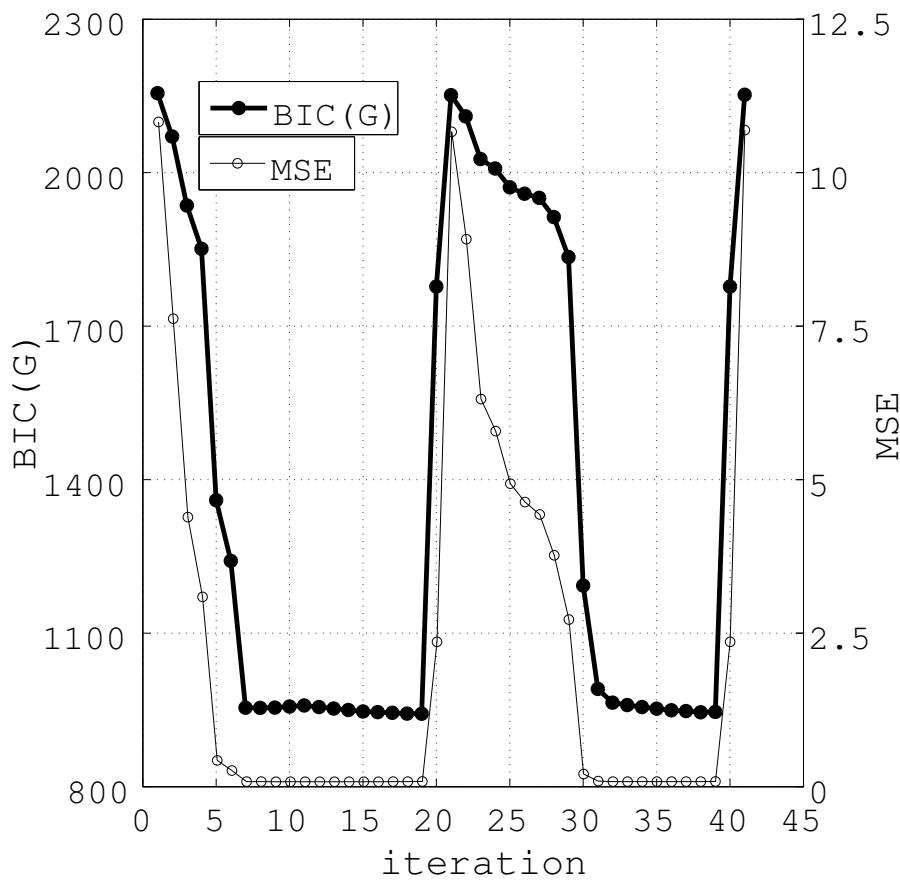
G	2	3	4	5	6
nstd = 0.1	2094.2	1743.7	486.2	486.8	487.1
nstd = 0.3	2164.2	1815.0	912.8	914.1	915.7
nstd = 0.5	2268.4	1992.8	1284.3	1286.1	1291.6
	7	8	9	10	11
	490.0	490.0	492.6	495.7	498.7
	926.6	920.2	920.7	924.3	927.9
	1300.8	1302.2	1305.3	1311.4	1316.5
	12				
					503.9
					929.0
					1320.0

同様に , パラメータの初期値を変えて 10 回ずつ 10-fold 交差検証を行ったときの , 交差検証誤差の最小値を表 6.18 に示す .

表 6.18: $MSE_{CV}(G)$ の最小値 (RF6.4+BCW 法 , 人工データ)

G	2	3	4	5	6	
nstd = 0.1	8.4761	2.0606	0.0142	0.0148	0.0151	
nstd = 0.3	9.6284	2.2784	0.1281	0.1299	0.1372	
nstd = 0.5	11.4049	2.5080	0.2611	0.2710	0.2691	
	7	8	9	10	11	12
	0.0163	0.0162	0.0172	0.0180	0.0184	0.0184
	0.1374	0.1389	0.1391	0.1390	0.1392	0.1399
	0.2710	0.2748	0.2800	0.2808	0.2833	0.2868

また , $nstd = 0.3$ のときの , ある試行における BCW 法での誤差の変化の様子を図 6.2 に示す .

図 6.2: 誤差の変化の様子 (RF6.4+BCW 法 , 人工データ , $nstd = 0.3$)

次に , $J = 2, R = 3$ でのパーセプトロンの学習によって得られたパラメータに對して , K 平均法を用いてベクトル量子化を行った . パラメータの初期値を変え て 100 回学習を行った後に , ベクトル量子化を行った . このとき $BIC(I)$ の値が最小 となった回数を表 6.19 , 6.20 に示す . また , I の値ごとの $BIC(I)$ の最小値を表 6.21 , 6.22 に示す . なお , 表 6.19 , 6.21 は , BCW 法を適用しない通常の RF6.4 法での 結果を示し , 表 6.20 , 6.22 は , $G = 4$ として BCW 法による共有重み構造を適用した RF6.4 法での結果を示す .

表 6.19: $BIC(I)$ が最小となった回数 (RF6.4 法 , 人工データ)

I	1	2	3	4	5	6	7	8	9	10	合計
nstd = 0.1	0	0	0	83	10	2	1	0	0	4	100
nstd = 0.3	0	0	0	82	9	3	1	1	1	3	100
nstd = 0.5	0	0	0	79	9	2	0	1	2	7	100

表 6.20: $BIC(I)$ が最小となった回数 (RF6.4+BCW 法 , 人工データ)

I	1	2	3	4	5	6	7	8	9	10	合計
nstd = 0.1	0	0	0	85	10	2	0	0	0	3	100
nstd = 0.3	0	0	0	82	10	5	0	0	0	3	100
nstd = 0.5	0	0	0	81	12	4	0	0	1	2	100

表 6.21: $BIC(I)$ の最小値 (RF6.4 法 , 人工データ)

I	1	2	3	4	5
nstd = 0.1	-41.9	-183.6	-304.5	-864.8	-865.5
nstd = 0.3	-42.3	-180.5	-305.4	-872.2	-881.9
nstd = 0.5	-39.2	-233.3	-341.3	-745.9	-757.2
	6	7	8	9	10
	-874.1	-914.4	-973.8	-993.9	-998.5
	-892.8	-902.2	-971.0	-988.3	-999.8
	-875.0	-870.8	-927.3	-931.9	-952.7

表 6.22: $BIC(I)$ の最小値 (RF6.4+BCW 法 , 人工データ)

I	1	2	3	4	5
nstd = 0.1	-41.3	-181.6	-305.8	-864.4	-875.5
nstd = 0.3	-41.9	-225.8	-326.5	-871.4	-883.6
nstd = 0.5	-41.2	-232.8	-332.7	-804.7	-811.8
	6	7	8	9	10
	-894.1	-916.3	-994.0	-1000.2	-1012.9
	-893.8	-908.9	-1002.3	-1028.8	-1041.3
	-874.6	-884.9	-947.8	-985.8	-999.5

同様に , パラメータの初期値を変えて 10 回ずつ学習した後にベクトル量子化を行い , 10-fold 交差検証を行ったときの , 交差検証誤差の最小値を表 6.23 , 6.24 に示す .

表 6.23: $MSE_{CV}(I)$ の最小値 (RF6.4 法 , 人工データ)

I	1	2	3	4	5
nstd = 0.1	362.5197	137.9922	14.9615	0.0179	0.0184
nstd = 0.3	383.7399	147.1613	14.6635	0.1213	0.1270
nstd = 0.5	342.7193	182.6128	15.9153	0.2649	0.2721
	6	7	8	9	10
	0.0191	0.0183	0.0188	0.0190	0.0194
	0.1296	0.1322	0.1324	0.1321	0.1347
	0.2730	0.2741	0.2818	0.2844	0.2851

表 6.24: $MSE_{CV}(I)$ の最小値 (RF6.4+BCW 法 , 人工データ)

I	1	2	3	4	5
nstd = 0.1	344.2711	147.9752	14.5597	0.0129	0.0138
nstd = 0.3	379.3678	148.8406	15.8526	0.1006	0.1071
nstd = 0.5	332.1736	152.8843	14.5439	0.2579	0.2625
	6	7	8	9	10
	0.0131	0.0142	0.0138	0.0140	0.0141
	0.1066	0.1065	0.1124	0.1139	0.1178
	0.2632	0.2700	0.2708	0.2713	0.2708

最後に , $nstd = 0.3$ のデータである試行において , c4.5 プログラムを適用して得られた決定木 , およびその各リーフに対応する代表点を以下に示す . ただし括弧中の数字は , 決定木の各リーフまで到達したベクトル量子化のサンプル数である .

$$\begin{aligned} q_{31} &= 1: \\ | \quad q_{21} &= 1: \text{rule1 (15)} \\ | \quad q_{22} &= 1: \text{rule2 (15)} \\ q_{32} &= 1: \\ | \quad q_{11} &= 1: \text{rule4 (10)} \\ | \quad q_{12} &= 1: \text{rule3 (10)} \\ | \quad q_{13} &= 1: \text{rule3 (10)} \end{aligned}$$

BCW 法による共有重み構造の適用にかかわらず , 多くの試行で上のような決定木が得られた .

共有重み構造無しの , $J = 2, R = 3, I = 4$ における通常の RF6.4 法での出力の式を以下に示す .

$$\left\{ \begin{array}{l} \text{if } q_{21} \wedge q_{31} \text{ then} \\ \quad y = 1.1152 + 1.9854x_1^{1.9191}x_2^{0.9792}x_3^{0.4997}x_4^{0.1323}x_5^{0.0638}x_6^{0.0095}x_7^{0.0205}x_8^{-0.0025} \\ \quad \quad \quad \times x_9^{-0.0030}x_{10}^{0.0024}x_{11}^{-0.0371}x_{12}^{-0.0088}x_{13}^{-0.0297}x_{14}^{-0.0314}x_{15}^{0.0152} \\ \quad \quad \quad + 2.9571x_1^{-0.0280}x_2^{1.0430}x_3^{0.5034}x_4^{2.0430}x_5^{1.0176}x_6^{-0.0094}x_7^{-0.0167}x_8^{0.0138} \\ \quad \quad \quad \times x_9^{0.0011}x_{10}^{0.0028}x_{11}^{0.0055}x_{12}^{-0.0010}x_{13}^{0.0040}x_{14}^{0.0138}x_{15}^{0.0025} \\ \text{if } q_{22} \wedge q_{31} \text{ then} \\ \quad y = 1.9677 - 0.9894x_1^{1.9191}x_2^{0.9792}x_3^{0.4997}x_4^{0.1323}x_5^{0.0638}x_6^{0.0095}x_7^{0.0205}x_8^{-0.0025} \\ \quad \quad \quad \times x_9^{-0.0030}x_{10}^{0.0024}x_{11}^{-0.0371}x_{12}^{-0.0088}x_{13}^{-0.0297}x_{14}^{-0.0314}x_{15}^{0.0152} \\ \quad \quad \quad + 0.9974x_1^{-0.0280}x_2^{1.0430}x_3^{0.5034}x_4^{2.0430}x_5^{1.0176}x_6^{-0.0094}x_7^{-0.0167}x_8^{0.0138} \\ \quad \quad \quad \times x_9^{0.0011}x_{10}^{0.0028}x_{11}^{0.0055}x_{12}^{-0.0010}x_{13}^{0.0040}x_{14}^{0.0138}x_{15}^{0.0025} \\ \text{if } (q_{12} \vee q_{13}) \wedge q_{32} \text{ then} \\ \quad y = -3.0033 + 0.9965x_1^{1.9191}x_2^{0.9792}x_3^{0.4997}x_4^{0.1323}x_5^{0.0638}x_6^{0.0095}x_7^{0.0205}x_8^{-0.0025} \\ \quad \quad \quad \times x_9^{-0.0030}x_{10}^{0.0024}x_{11}^{-0.0371}x_{12}^{-0.0088}x_{13}^{-0.0297}x_{14}^{-0.0314}x_{15}^{0.0152} \\ \quad \quad \quad - 0.9970x_1^{-0.0280}x_2^{1.0430}x_3^{0.5034}x_4^{2.0430}x_5^{1.0176}x_6^{-0.0094}x_7^{-0.0167}x_8^{0.0138} \\ \quad \quad \quad \times x_9^{0.0011}x_{10}^{0.0028}x_{11}^{0.0055}x_{12}^{-0.0010}x_{13}^{0.0040}x_{14}^{0.0138}x_{15}^{0.0025} \\ \text{if } q_{11} \wedge q_{32} \text{ then} \\ \quad y = 2.2090 + 2.9646x_1^{1.9191}x_2^{0.9792}x_3^{0.4997}x_4^{0.1323}x_5^{0.0638}x_6^{0.0095}x_7^{0.0205}x_8^{-0.0025} \\ \quad \quad \quad \times x_9^{-0.0030}x_{10}^{0.0024}x_{11}^{-0.0371}x_{12}^{-0.0088}x_{13}^{-0.0297}x_{14}^{-0.0314}x_{15}^{0.0152} \\ \quad \quad \quad + 3.9445x_1^{-0.0280}x_2^{1.0430}x_3^{0.5034}x_4^{2.0430}x_5^{1.0176}x_6^{-0.0094}x_7^{-0.0167}x_8^{0.0138} \\ \quad \quad \quad \times x_9^{0.0011}x_{10}^{0.0028}x_{11}^{0.0055}x_{12}^{-0.0010}x_{13}^{0.0040}x_{14}^{0.0138}x_{15}^{0.0025} \end{array} \right. \quad (6.5)$$

BCW 法による共有重み構造での , $J = 2, R = 3, G = 4, I = 4$ における RF6.4 法の出

力の式を以下に示す。

$$\left\{ \begin{array}{l} \text{if } q_{21} \wedge q_{31} \text{ then} \\ \quad y = 1.0076 + 1.9823x_1^{2.0109}x_2^{1.0520}x_3^{0.4981} + 2.9832x_2^{1.0520}x_3^{0.4981}x_4^{2.0109}x_5^{1.0520} \\ \text{if } q_{22} \wedge q_{31} \text{ then} \\ \quad y = 2.0123 - 1.0001x_1^{2.0109}x_2^{1.0520}x_3^{0.4981} + 0.9699x_2^{1.0520}x_3^{0.4981}x_4^{2.0109}x_5^{1.0520} \\ \text{if } (q_{12} \vee q_{13}) \wedge q_{32} \text{ then} \\ \quad y = -2.9800 + 1.1275x_1^{2.0109}x_2^{1.0520}x_3^{0.4981} - 0.9910x_2^{1.0520}x_3^{0.4981}x_4^{2.0109}x_5^{1.0520} \\ \text{if } q_{11} \wedge q_{32} \text{ then} \\ \quad y = 1.9752 + 3.0646x_1^{2.0109}x_2^{1.0520}x_3^{0.4981} + 4.0423x_2^{1.0520}x_3^{0.4981}x_4^{2.0109}x_5^{1.0520} \end{array} \right. \quad (6.6)$$

6.2.3 考察

まず, $\text{BIC}(J, R)$ による最適隠れユニット数の選択では, 表 6.7 ~ 6.9 より, $nstd = 0.5$ のデータにおける 1 回の試行を除いた全ての試行で, 正解である $J^* = 2$ を選択できている。また, $nstd = 0.1$ のデータでは 9 割以上, $nstd = 0.3$ のデータでは 8 割以上, $nstd = 0.5$ のデータでは 6 割以上の試行で, $R^* = 3$ を選択している。復元する回帰ルール集合の, 定数項を含む係数値を並べた行列を C とすると, RF6.4 法ではパーセプトロンの構造から, $R \geq \text{rank}(C)$ のときその式を復元できる。式 (6.4)においては, 定数項を含む項数が 3 なので, $\text{rank}(C) = 3$ となり, R は 3 以上であれば元の式を復元することができる。 R は最終出力の式中には陽に現れないで, 3 より大きい値を選択してしまったとしても問題はないといえる。表 6.10 ~ 6.15 より, $\text{BIC}(J, R)$ 及び $\text{MSE}_{\text{CV}}(J, R)$ の最小値でも, 正解である $J^* = 2$ を選択できており, R^* も正解の回帰ルール集合を表現するために必要な 3 以上の値を選択できている。

$\text{BIC}(G)$ 及び $\text{MSE}_{\text{CV}}(G)$ による最適クラスタ数の選択でも, 表 6.16 ~ 6.18 より, 多くの試行で, 正解である $G^* = 4$ を選択できている。しかし, $\text{BIC}(G)$ の値は, 正解である $G = 4$ とそれ以上の G の値のときではほとんど値が変わっていないことがわかる。

図 6.2 は双方向クラスタリングが 2 回行われた場合の例である。RF5 法の実験では 1 回で収束しちゃうことが多かったが, RF6.4 法の実験においては, 平均して 2 ~ 3 回の双方向クラスタリングが行われていた。RF5 法に比べて RF6.4 法はパラメータ数が多く, より複雑な式となっているので, 双方向クラスタリングの回数も多めに必要であると思われる。

最適ルール数の選択においては, 表 6.19, 6.20 より, $\text{BIC}(I)$ が最小となった回数では, 共有重み構造の適用にかかわらず, 正解である $I^* = 4$ を選択できている。

また、表 6.23, 6.24 より、 $\text{MSE}_{\text{CV}}(I)$ の最小値でも $I^* = 4$ を選択できている。しかし、表 6.21, 6.22 に示されるように、 $\text{BIC}(I)$ の最小値では $I^* = 10$ が選択されてしまっている。表 6.19~6.20において、数回の試行で $I = 10$ のときに $\text{BIC}(I)$ が最小となっているが、このときの値が、多くの試行における $I = 4$ のときの $\text{BIC}(I)$ の値より小さくなってしまったために、 $I = 10$ が選択された。従って、RF6.4 法の最適ルール数の選択においては、 $\text{BIC}(I)$ の最小値を用いるのは好ましくないといえる。なお、BCW 法による共有重み構造の適用は多項式の指數部分だけであり、係数部分の値は BCW 法の適用にかかわらずあまり変わらないため、表 6.19 と 6.20, 6.21 と 6.22 では値がかなり近くなっている。しかし、表 6.23, 6.24 より、BCW 法による共有重み構造の適用によって、汎化性能が向上していることがわかる。

式 (6.6) は、RF5 法の実験と同様に、最も 0 に近い共有重みの値を 0 として枝刈りした後に、パーセプトロンを再学習して得られた結果を示している。RF6.4 法では、最終的な出力は回帰ルール集合となるため、BCW 法による共有重み構造の適用によって、RF5 法よりも大きく可読性を向上させることができている。

6.3 質的条件付き重回帰法 (nc-MR 法)

6.3.1 実験における諸設定

実験データ

質的条件付き重回帰の回帰ルール集合として、式 (6.7) を考える。

$$\left\{ \begin{array}{l} \text{if } (q_{11} \vee q_{12}) \wedge (q_{21} \vee q_{23}) \wedge q_{31} \text{ then} \\ \quad y = -9 - 8x_1 + 4x_2 - 8x_3 - 5x_4 - 5x_5 - 6x_6 - 1x_7 + 2x_8 + 9x_9 + 3x_{10} \\ \text{if } (q_{11} \vee q_{12}) \wedge q_{21} \wedge q_{32} \text{ then} \\ \quad y = -3 - 1x_1 + 4x_2 - 1x_3 + 7x_4 - 5x_5 + 7x_6 - 3x_7 - 8x_8 - 5x_9 + 9x_{10} \\ \text{if } q_{13} \wedge (q_{21} \vee q_{23}) \text{ then} \\ \quad y = 3 - 1x_1 + 5x_2 + 4x_3 - 5x_4 - 9x_5 - 7x_6 - 4x_7 - 9x_8 - 7x_9 + 3x_{10} \\ \text{if } (q_{11} \vee q_{12}) \wedge q_{23} \wedge q_{32} \text{ then} \\ \quad y = 4 - 3x_1 + 8x_3 + 6x_4 - 8x_5 - 7x_6 - 3x_7 - 1x_8 + 7x_9 + 7x_{10} \\ \text{if } q_{22} \text{ then} \\ \quad y = 4 - 7x_1 + x_2 - 5x_3 + 8x_4 + 3x_5 + 10x_6 - 2x_7 + 7x_8 - 5x_9 - 10x_{10} \end{array} \right. \quad (6.7)$$

量的説明変数として x_1, \dots, x_{15} , 質的説明変数として q_1, \dots, q_4 を用い, 質的説明変数のカテゴリ数は $L_1 = L_2 = 3, L_3 = 2, L_4 = 5$ とした。ただし、式から明らかなように, $x_{11}, \dots, x_{15}, q_4$ は回帰ルール集合に現れない不要な変数である。訓練データのサンプル数は $N = 500$ とした。各サンプルの量的変数 x_j の値は $(-1, 1)$ の範囲の一様乱数とし, 質的変数 q_k の値はランダムにどれか 1 つのカテゴリをとるよう

にした。そして、式(6.7)のルール集合を元に最終出力を計算し、正規ノイズを加えた。ノイズの平均を0, 標準偏差を0.1, 0.3, 0.5と変化させて出力に加え、それぞれのデータごとに実験を行った。

パラメータの設定と終了条件

各重みの初期値は $(-1, 1)$ の一様乱数とした。パーセプトロンの学習は、勾配ベクトルのどの成分も 10^{-6} 未満となったとき、または学習の繰り返し回数が20,000回を超えたとき停止させた。

モデル選択の方法としてはBICと交差検証法を行い、 $R = 1, \dots, 10, I = 1, \dots, 10$ としてそれぞれの場合の交差検証誤差とBICの値を比較した。

6.3.2 実験結果

step1で、 R の値ごとにパラメータの初期値を変えて100回ずつ学習を行った。このとき $BIC(R)$ の値が最小となった回数を表6.25に示す。また、 R の値ごとの $BIC(R)$ の最小値を表6.26に示す。

表 6.25: $BIC(R)$ が最小となった回数 (nc-MR 法, 人工データ)

R	1	2	3	4	5	6	7	8	9	10	合計
nstd = 0.1	0	0	0	0	91	9	0	0	0	0	100
nstd = 0.3	0	0	0	0	84	16	0	0	0	0	100
nstd = 0.5	0	0	0	0	79	21	0	0	0	0	100

表 6.26: $BIC(R)$ の最小値 (nc-MR 法, 人工データ)

R	1	2	3	4	5
nstd = 0.1	2790.8	2737.5	2485.4	2305.1	758.8
nstd = 0.3	2785.0	2750.3	2528.5	2290.2	1328.1
nstd = 0.5	2737.8	2632.9	2480.2	2272.6	1507.4
	6	7	8	9	10
	805.3	873.2	918.0	972.0	1035.0
	1369.7	1400.8	1435.4	1479.7	1492.5
	1557.0	1600.4	1634.7	1678.9	1714.8

同様に、10-fold交差検証によるモデル選択を行った。 J の値ごとにパラメータの初期値を変えて10回ずつ交差検証を行ったときの、交差検証誤差の最小値を表

6.27 に示す。なお、10 回の交差検証でのセグメントの分割方法は全て同じとした。

表 6.27: $MSE_{CV}(R)$ の最小値 (nc-MR 法, 人工データ)

R	1	2	3	4	5
nstd = 0.1	111.8680	75.2753	28.1162	17.8738	0.0151
nstd = 0.3	105.7365	83.8325	32.2742	16.3635	0.1690
nstd = 0.5	88.5567	49.8401	27.6252	20.9539	0.3829
	6	7	8	9	10
	0.0176	0.0231	0.0226	0.0265	0.0261
	0.2131	0.2380	0.2473	0.2677	0.4530
	0.4104	0.4824	0.4962	0.4972	0.4979

次に step2 で, $R = 5$ でのパーセプトロンの学習によって得られたパラメータに対して, K 平均法を用いてベクトル量子化を行った。パラメータの初期値を変えて 100 回学習を行った後に, ベクトル量子化を行った。このとき $BIC(I)$ の値が最小となった回数を表 6.28 に示す。また, I の値ごとの $BIC(I)$ の最小値を表 6.29 に示す。

表 6.28: $BIC(I)$ が最小となった回数 (nc-MR 法, 人工データ)

I	1	2	3	4	5	6	7	8	9	10	合計
nstd = 0.1	0	0	0	0	89	10	1	0	0	0	100
nstd = 0.3	0	0	0	0	48	17	10	9	6	10	100
nstd = 0.5	0	0	0	0	31	27	12	10	8	12	100

表 6.29: $BIC(I)$ の最小値 (nc-MR 法, 人工データ)

I	1	2	3	4	5
nstd = 0.1	132.1	-244.0	-803.4	-1599.0	-7408.4
nstd = 0.3	489.2	-3.4	-493.6	-1580.0	-5414.8
nstd = 0.5	837.8	457.1	18.5	-529.4	-3614.7
	6	7	8	9	10
	-7317.4	-7292.2	-7242.6	-7199.9	-7163.4
	-5348.0	-5422.5	-5451.6	-5516.4	-5569.5
	-3806.2	-4087.9	-4120.4	-4137.4	-4218.6

同様に, パラメータの初期値を変えて 10 回ずつ学習した後にベクトル量子化を行い, 10-fold 交差検証を行ったときの, 交差検証誤差の最小値を表 6.30 に示す。

表 6.30: $MSE_{CV}(I)$ の最小値 (nc-MR 法 , 人工データ)

I	1	2	3	4	5
nstd = 0.1	111.6966	75.4667	27.6605	14.3625	0.0138
nstd = 0.3	105.4579	82.4465	29.9780	12.1713	0.1471
nstd = 0.5	88.6343	50.2931	27.9469	27.1803	0.3212
	6	7	8	9	10
	0.0138	0.0152	0.0158	0.0150	0.0143
	0.1565	0.1605	0.1675	0.1658	0.1693
	0.3624	0.3734	0.3883	0.3815	0.3844

Step3 で , $nstd = 0.3$ のデータである試行において , c4.5 プログラムを適用して得られた決定木 , およびその各リーフに対応する代表点を以下に示す . ただし括弧中の数字は , 決定木の各リーフまで到達したベクトル量子化のサンプル数である .

$q_{22} = 1$: rule5 (30)

$q_{21} = 1$:

| $q_{31} = 1$:

| | $q_{11} = 1$: rule1 (5)

| | $q_{12} = 1$: rule1 (5)

| | $q_{13} = 1$: rule3 (5)

| $q_{32} = 1$:

| | $q_{11} = 1$: rule2 (5)

| | $q_{12} = 1$: rule2 (5)

| | $q_{13} = 1$: rule3 (5)

$q_{23} = 1$:

| $q_{31} = 1$:

| | $q_{11} = 1$: rule1 (5)

| | $q_{12} = 1$: rule1 (5)

| | $q_{13} = 1$: rule3 (5)

| $q_{32} = 1$:

| | $q_{11} = 1$: rule4 (5)

| | $q_{12} = 1$: rule4 (5)

| | $q_{13} = 1$: rule3 (5)

最後に, $R = 5, I = 5$ における nc-MR 法での出力の式を以下に示す .

$$\left\{
 \begin{array}{l}
 \text{if } (q_{11} \vee q_{12}) \wedge (q_{21} \vee q_{23}) \wedge q_{31} \text{ then} \\
 \quad y = -8.9405 - 7.9333x_1 + 4.0718x_2 - 7.9151x_3 - 4.9578x_4 - 5.0626x_5 \\
 \quad \quad - 5.9717x_6 - 0.9834x_7 + 1.9399x_8 + 8.9693x_9 + 3.0411x_{10} \\
 \quad \quad + 0.0628x_{11} + 0.0217x_{12} + 0.0627x_{13} - 0.0851x_{14} + 0.0102x_{15} \\
 \text{if } (q_{11} \vee q_{12}) \wedge q_{21} \wedge q_{32} \text{ then} \\
 \quad y = -2.9528 - 1.0055x_1 + 4.0753x_2 - 0.9309x_3 + 6.9248x_4 - 5.0986x_5 \\
 \quad \quad + 7.0128x_6 - 3.1530x_7 - 7.9232x_8 - 4.8106x_9 + 9.1747x_{10} \\
 \quad \quad + 0.1393x_{11} + 0.0133x_{12} - 0.0521x_{13} - 0.0079x_{14} - 0.1268x_{15} \\
 \text{if } q_{13} \wedge (q_{21} \vee q_{23}) \text{ then} \\
 \quad y = 2.9748 - 0.9363x_1 + 5.0117x_2 + 4.1227x_3 - 5.0821x_4 - 9.0607x_5 \\
 \quad \quad - 7.0295x_6 - 3.8950x_7 - 9.0370x_8 - 6.9625x_9 + 2.9386x_{10} \\
 \quad \quad + 0.1973x_{11} - 0.0170x_{12} - 0.1034x_{13} - 0.0287x_{14} - 0.0196x_{15} \\
 \text{if } (q_{11} \vee q_{12}) \wedge q_{23} \wedge q_{32} \text{ then} \\
 \quad y = 4.0228 - 2.8589x_1 - 0.0465x_2 + 8.1287x_3 + 5.9308x_4 - 7.9713x_5 \\
 \quad \quad - 6.7350x_6 - 3.0863x_7 - 0.7004x_8 + 7.1023x_9 + 7.1625x_{10} \\
 \quad \quad + 0.1647x_{11} + 0.0291x_{12} - 0.0413x_{13} - 0.2152x_{14} + 0.0362x_{15} \\
 \text{if } q_{22} \text{ then} \\
 \quad y = 3.9601 - 7.0385x_1 + 0.9954x_2 - 4.9896x_3 + 7.9972x_4 + 2.9766x_5 \\
 \quad \quad + 9.9886x_6 - 2.0065x_7 + 7.0264x_8 - 4.9860x_9 - 9.9356x_{10} \\
 \quad \quad - 0.0359x_{11} + 0.0375x_{12} - 0.0436x_{13} + 0.0275x_{14} + 0.0444x_{15}
 \end{array}
 \right. \tag{6.8}$$

6.3.3 考察

表 6.25 より, 多くの試行で $R^* = 5$ を選択できていることがわかる . もとの式は回帰ルール数が 5 なので, $\text{rank}(\mathbf{C}) = 5$ となり, $R \geq 5$ ならば元の式をうまく復元できる . したがって, 数回の試行で $R = 6$ を選択しているが, RF6.4 法の場合と同様に大きな問題とはならない . なお, このときのパラメータ値を確認したところ, その多くの場合で $R = 5$ の学習が局所最適解に陥っていた . それが $R = 6$ を選択してしまった主な原因と考えられる . 表 6.26, 6.27 より, $\text{BIC}(R)$ 及び $\text{MSE}_{\text{CV}}(R)$ の最小値では, $R^* = 5$ を選択できている .

表 6.28 より, $nstd = 0.1$ で約 9 割, $nstd = 0.3$ で約 5 割, $nstd = 0.5$ で約 3 割の試行で正解である $I = 5$ を選択できている . 表 6.29 より, $\text{BIC}(I)$ の最小値では, $nstd = 0.1$ では正解を選択できているが, 他の場合には $I^* = 10$ を選択てしまっている . 従って, ノイズが大きくなるにつれて, 正解よりも大きい I の値を選択してしまう傾向があることがわかる . 表 6.30 より, $\text{MSE}_{\text{CV}}(I)$ の最小値では, $I^* = 5$ を選択できている . また, 係数ベクトルの量子化を行った方が, 僅かに汎化性能が向上しているのがわかる .

なお，今回の実験では人工データ作成時に正規ノイズを加えているが，ノイズを加えない場合の提案法の動作についても確認した．その場合， R や I を増加させて，真の回帰ルール集合を説明できる値になった瞬間に，訓練誤差が0となった．このとき，ノイズの統計モデリングで導出された $BIC(R)$ 及び $BIC(I)$ の値は $-\infty$ となってしまい，それ以降 R, I を増加させても $-\infty$ のままであった．この場合は， $BIC(R)$ 及び $BIC(I)$ が $-\infty$ となる最小の R, I を R^*, I^* とすればよい．しかし，提案法を実データに適用する場合においては，このようなことが起こることはほとんど無いと考えられる．

式(6.8)より，nc-MR法によってもとの回帰ルール集合とほぼ等価なものを復元することができている．

6.4 まとめ

人工データを用いた実験によって，RF5法，RF5+BCW法，RF6.4法，RF6.4+BCW法，nc-MR法が正しく式を復元できるかどうかを検証した．重み共有無しのRF5法及びRF6.4法でも，もとの式に近いものを復元することができたが，不要変数が多く含まれている場合などには，可読性の悪い結果を得ることになってしまう．しかし，BCW法による重み共有を適用することで，不要変数に対するパラメータを簡単に枝刈りすることができ，非常に可読性の良い式を得ることができることが確認できた．更に，可読性を向上させるだけでなく，汎化性能を向上させることができることも確認できた．また，nc-MR法でも多項式回帰の場合と同様に，もとの回帰ルール集合とほぼ等価なものを復元することができた．

BIC と交差検証法によるモデル選択を行ったが，ほとんどの試行で共に正解のモデルを選択することができていた．しかし，最適ルール数の選択において $BIC(I)$ の最小値を用いると，正解のルール数より大きい値を選択してしまうことがあった．このときの $BIC(I)$ の値は， I が正解のルールより大きい場合は非常に近い値となっていた．従って， $BIC(I)$ の最小値をモデル選択の方法として用いるのは好ましくないかもしれない．人工データは真のルールというものが存在する特殊なデータであり，このようなデータ特有の現象であることも考えられるので，現実データによる実験でも検証する必要がある．

第 7 章

現実データを用いた実験による検証

7.1 現実データへの適用

第 6 章では、人工データを用いた実験によって、提案モデルが元の回帰式、または回帰ルール集合をうまく復元できることを示した。しかし、人工データと現実データでは、データの性質が大きく異なる部分があるので、様々な現実データを用いた検証が必要である。現実データに適用する場合の、人工データに適用する場合との違いとしては、主に以下の 3 つが挙げられる。

1. 真の分布、真のモデルの復元をするわけではない

人工データを用いた実験では、真のパラメータ値やモデルというものがはっきりわかり、それを復元できるかどうかの検証が最も重要であった。しかし、現実のデータには通常、真の分布やモデルというものは存在しないことが多い。仮に真の分布やモデルが存在すると仮定しても、その真のモデルを推定するという問題設定は適当ではないといえる。つまり、モデリングの最も重要な目的は、「真の」モデルを求めることではなく、「よい」モデルを求めることであるといえる。ここでの「よい」モデルというのは、未知のデータに対する汎化性能や予測結果の可読性のことをいう。

2. データに事前処理を施す必要がある

人工データを用いた実験では、データのスケールなどが自由に設定でき、データの加工などを行わなくてもそのままモデルを適用することができる。しかし、現実データではモデルに適用する前に、データの加工が必要な場合が多い。例えば、説明変数や目的変数のスケールを揃えるためには、データに正規化を施す。また、あるサンプルの変数に欠損値がある場合には、なんらかの方法でその欠損値

を補うか，そのサンプルを取り除くなどの処理が必要である。このような事前処理は，予測結果やその汎化性能などに大きな影響を及ぼす。

3. 変数に「意味」がある

人工データを用いた実験では，各変数には特に意味があるわけではないので，最終的な回帰式が得られたらそれで終わりである。しかし，現実データの場合には，各変数にはなんらかの意味があり，最終的な回帰式が得られた後にそれを解析することが重要である。そもそも多くの場合には，回帰式自体を必要としてモデリングを行うわけではなく，その得られた結果を解析することを目的としている。また逆に，変数に意味があるということは，それを事前知識としてモデリングに有效地に活用できる可能性がある。

上記のような違いも考慮して，この章では様々なデータを用いて検証を行う。まず，特定のデータ2つについて予測結果を深く考察した後，他のデータを用いて性能の比較を行う。ここでは，データの内容が非常にわかりやすい例として，ダイヤモンドの価格データとプロ野球選手の年俸データの予測結果について詳しく掲載した。

7.2 ダイヤモンドの価格データ

本研究で扱う回帰モデルは全て，多変量データに対して適応可能であり，多層パーセプトロンを用いるので，単変量データより多変量データへの適用の方が適している。しかしこの節では，回帰式のグラフを表示でき，視覚的に検証ができるという理由で，説明変数が1変数であるダイヤモンドの価格データ[38]を用いた検証を行う。

7.2.1 実験における諸設定

実験データ

ダイヤモンドの価格データ¹は，ダイヤモンドの重さと価格に関するデータで，1つの量的説明変数 x ，3つの質的説明変数 q_1, q_2, q_3 及び目的変数 y から成る。 x はダイヤモンドの重さ(カラット)を， y はその価格(シンガポールドル)を表す。また質的変数は， q_1 は色のグレードD, E, F, G, H, I(無色に近いものからD, E, …)

¹データは Jurnal of Statistics Education のホームページより入手した。

を， q_2 は透明度 IF, VVS1, VVS2, VS1, VS2 (透明に近いものから IF, VVS1, …) を， q_3 は鑑定書の発行機関 GIA(Gemological Institute of America), IGI(International Gemmological Institute), HRD(Hoge Raad voor Diamant) をそれぞれ表す．データのサンプル数は $N = 308$ である．データの散布図を図 7.1 に示す．なお，質的変数の異なるサンプルを異なる色で表示している．

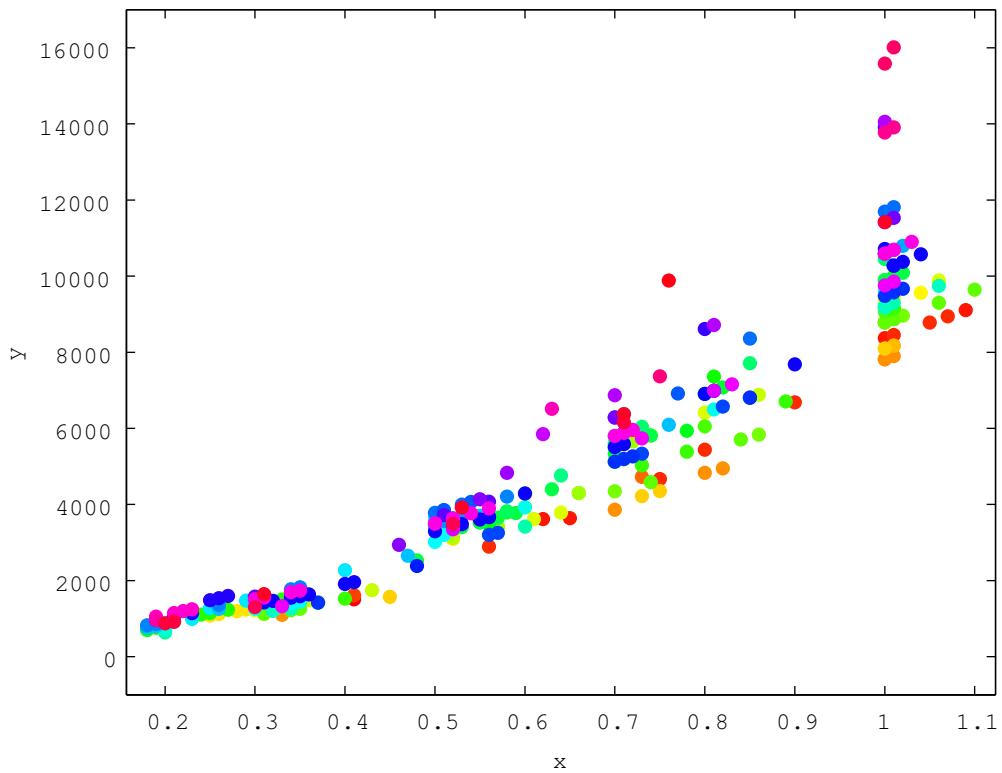


図 7.1: ダイヤモンドの価格データ

実験方法

重回帰分析，数量化理論一類，RF5 法，RF6.4 法，3 層パーセプトロン，HME，及び nc-MR 法にデータを適用し，得られた回帰式について考察する．なお，3 層パーセプトロンでは，量的変数のみを用いる MLP_x と，質的変数も用いる MLP_{xq} の両方を行う．また，量的変数が 1 つしかないので，RF5 法や RF6.4 法では BCW 法による重み共有は行わない．

多層パーセプトロンを用いるモデルにおいては，隠れユニット数を $J = 1, \dots, 5, R = 1, \dots, 5$ として BIC 及び 10-fold 交差検証法によって最適なモデルを選択する．また，RF6.4 法及び nc-MR 法では，ルール数を $I = 1, \dots, 10$ とした．

HME では，HME の構成的学習アルゴリズム [39] で学習を行い，エキスパート

数(回帰ルール数)を $I = 1, 2, \dots, 10$ と変化させた。ゲート変数として量的変数を用いた HME_x と、質的変数を用いた HME_q の両方を行う。

パーセプトロンの学習におけるパラメータの初期値は、人工データを用いた実験と同じである。ただし、MLP では全ての重みの初期値を $(-1, 1)$ の一様乱数とした。学習は勾配ベクトルのどの成分も 10^{-5} 未満となったとき、または学習の繰り返し回数が 20,000 回を超えたとき停止させた。

なお、学習の際には量的変数に以下のような正規化を施した。

$$\tilde{x}_j = \frac{x_j}{\max(x_j)}, \quad \tilde{y} = \frac{y - \text{mean}(y)}{\text{std}(y)} \quad (7.1)$$

7.2.2 実験結果

まず、RF5 法、MLP、nc-MR 法において、 J または R の値ごとにパラメータの初期値を変えて 100 回ずつ学習を行った。このとき $\text{BIC}(J)$ の値が最小となった回数を表 7.1 に、 J の値ごとの $\text{BIC}(J)$ の最小値を表 7.2 に示す。また、 J の値ごとにパラメータの初期値を変えて 10 回ずつ交差検証を行ったときの交差検証誤差の最小値を表 7.3 に示す。ただし、ここでは nc-MR 法における中間ユニット数 R を J と表記している。

表 7.1: $\text{BIC}(J)$ が最小となった回数 (RF5 法・MLP・nc-MR 法、ダイヤモンドデータ)

J	1	2	3	4	5	合計
RF5 法	11	42	25	16	6	100
MLPx	100	0	0	0	0	100
MLPxq	0	83	17	0	0	100
nc-MR 法	0	4	91	5	0	100

表 7.2: $\text{BIC}(J)$ の最小値 (RF5 法・MLP・nc-MR 法、ダイヤモンドデータ)

J	1	2	3	4	5
RF5 法	517.4	522.6	519.7	523.2	528.9
MLPx	519.5	520.9	524.9	533.5	542.1
MLPxq	215.3	79.9	77.6	91.0	115.7
nc-MR 法	477.5	319.3	305.0	315.5	336.6

表 7.3: $MSE_{CV}(J)$ ($\times 10^6$) の最小値 (RF5 法・MLP・nc-MR 法, ダイヤモンドデータ)

J	1	2	3	4	5
RF5 法	0.9765	0.9552	1.0132	1.1902	1.3005
MLPx	0.9639	0.9698	0.9699	1.0151	1.0336
MLPxq	0.6445	0.5393	0.4833	0.5022	0.5185
nc-MR 法	0.7327	0.5274	0.4978	0.5837	0.5998

同様に, RF6.4 法において, J, R の値ごとにパラメータの初期値を変えて 100 回ずつ学習を行ったときの $BIC(J, R)$ の値が最小となった回数, $BIC(J, R)$ の最小値, 及び $MSE_{CV}(J, R)$ の最小値をそれぞれ表 7.4 ~ 7.6 に示す.

表 7.4: $BIC(J, R)$ が最小となった回数 (RF6.4 法, ダイヤモンドデータ)

models	J = 1	J = 2	J = 3	J = 4	J = 5
R = 1	0	0	0	0	0
R = 2	0	0	0	0	0
R = 3	25	40	21	2	0
R = 4	0	0	4	0	8
R = 5	0	0	0	0	0

表 7.5: $BIC(J, R)$ の最小値 (RF6.4 法, ダイヤモンドデータ)

models	J = 1	J = 2	J = 3	J = 4	J = 5
R = 1	439.4	444.7	448.2	453.6	459.1
R = 2	108.0	97.0	102.5	111.2	115.8
R = 3	61.6	62.1	66.6	78.6	94.5
R = 4	77.3	82.8	84.3	98.6	112.5
R = 5	92.5	97.5	112.7	115.8	137.2

表 7.6: $MSE_{CV}(J, R)$ ($\times 10^6$) の最小値 (RF6.4 法, ダイヤモンドデータ)

models	J = 1	J = 2	J = 3	J = 4	J = 5
R = 1	5.5266	5.5137	5.5088	5.4823	5.4831
R = 2	0.5858	0.6057	0.5895	0.5828	0.5756
R = 3	0.4598	0.4690	0.5924	0.6246	0.6754
R = 4	0.4619	0.4778	0.5253	0.6026	0.5968
R = 5	0.4594	0.5596	0.5398	0.5724	0.7204

RF6.4法及びnc-MR法において， I の値ごとにパラメータの初期値を変えて100回ずつ学習を行ったときの $BIC(I)$ の値が最小となった回数， $BIC(I)$ の最小値，及び $MSE_{CV}(I)$ の最小値をそれぞれ表7.7～7.9に示す．ただし，RF6.4法及びnc-MR法では，表7.3, 7.6において交差検証誤差が最小になるときの R, J の値を用いた．また，HME_x及びHME_qでは， $I = 1$ のときは通常の重回帰分析を行った．

表 7.7: $BIC(I)$ が最小となった回数 (RF6.4法・HME・nc-MR法，ダイヤモンドデータ)

I	1	2	3	4	5	6	7	8	9	10	合計
RF6.4 法	0	0	0	0	0	0	89	10	1	0	100
HME _x	0	0	63	11	13	3	1	4	2	3	100
HME _q	0	0	0	0	0	10	11	16	22	41	100
nc-MR 法	0	0	0	0	0	2	2	9	53	34	100

表 7.8: $BIC(I)$ の最小値 (RF6.4法・HME・nc-MR法，ダイヤモンドデータ)

I	1	2	3	4	5
RF6.4 法	-1732.0	-1832.5	-1898.2	-1980.5	-2034.6
HME _x	104.5	-1.4	-10.0	-38.3	-35.1
HME _q	104.5	-5.3	-32.9	-68.4	-91.1
nc-MR 法	-314.0	-332.9	-356.1	-357.3	-357.7
	6	7	8	9	10
	-2087.3	-2100.5	-2123.4	-2144.8	-2134.7
	-31.1	-26.3	-24.1	-20.1	-18.3
	-109.5	-119.6	-124.5	-131.8	-146.8
	-359.6	-365.6	-365.3	-367.6	-375.5

表 7.9: $MSE_{CV}(I)$ ($\times 10^6$) の最小値 (RF6.4法・HME・nc-MR法，ダイヤモンドデータ)

I	1	2	3	4	5
RF6.4 法	1.3906	0.9889	0.7213	0.6882	0.6131
HME _x	1.2477	1.0624	0.9810	1.0025	0.9962
HME _q	1.2477	0.9324	0.8111	0.7591	0.6355
nc-MR 法	1.4243	1.0533	0.9291	0.7582	0.6426
	6	7	8	9	10
	0.4729	0.3942	0.4172	0.4481	0.4494
	1.0945	1.0858	1.1003	1.0752	1.1525
	0.5198	0.5297	0.4967	0.5110	0.5192
	0.5100	0.4798	0.4678	0.4818	0.4790

全データを用いて、重回帰分析、数量化理論一類、RF5法、RF6.4法、及びnc-MR法によって得られた回帰式を以下に示す。ただし、RF5法、MLPx、RF6.4法、nc-MR法では、表7.3、7.6、7.9において交差検証誤差が最小になるときの R, J, I の値を用いた。また、数量化理論一類では、各質的変数 q_1, q_2, q_3 の中で最も値が低いカテゴリに対するパラメータ値が0となるようにした。

重回帰分析

$$y = 11599x - 2298 \quad (7.2)$$

数量化理論一類 $y = 12766x - 4936$

$$\begin{aligned} &+3313q_{1,D} + 1874q_{1,E} + 1471q_{1,F} + 1136q_{1,G} + 566q_{1,H} \\ &+1792q_{2,IF} + 1103q_{2,VVS1} + 601q_{2,VVS2} + 317q_{2,VS1} \\ &+141q_{3,IGI} + 15q_{3,HRD} \end{aligned} \quad (7.3)$$

RF5法

$$y = 20367 + 12084x^{1.4178} - 22760x^{0.0589} \quad (7.4)$$

RF6.4法

$$\left\{ \begin{array}{ll} \text{if rule1 then } y = 15532x^{1.8372} + 279 \\ \text{if rule2 then } y = 13649x^{1.8372} + 351 \\ \text{if rule3 then } y = 11827x^{1.8372} + 363 \\ \text{if rule4 then } y = 10831x^{1.8372} + 432 \\ \text{if rule5 then } y = 9877x^{1.8372} + 347 \\ \text{if rule6 then } y = 9074x^{1.8372} + 181 \\ \text{if rule7 then } y = 8231x^{1.8372} - 74 \end{array} \right. \quad (7.5)$$

表 7.10: 回帰ルールの条件部 (RF6.4法, ダイヤモンドデータ)

	D	E	F	G	H	I
IF	rule1	rule2	rule2	rule4	rule5	rule6
VVS1	rule1	rule3	rule4	rule5	rule6	rule7
VVS2	rule3	rule4	rule5	rule5	rule6	rule7
VS1	rule4	rule5	rule5	rule6	rule6	rule7
VS2	rule4	rule5	rule6	rule6	rule7	rule7

nc-MR 法

$$\left\{ \begin{array}{l} \text{if rule1 then } y = 23696x - 7992 \\ \text{if rule2 then } y = 19642x - 6326 \\ \text{if rule3 then } y = 15499x - 4357 \\ \text{if rule4 then } y = 14696x - 4590 \\ \text{if rule5 then } y = 12910x - 4120 \\ \text{if rule6 then } y = 11950x - 2797 \\ \text{if rule7 then } y = 10002x - 2595 \\ \text{if rule8 then } y = 9595x - 1210 \end{array} \right. \quad (7.6)$$

表 7.11: 回帰ルールの条件部 (nc-MR 法 , ダイヤモンドデータ)

	D	E	F	G	H	I
GIA	rule1	rule3	rule3	rule6	rule6	rule7
IGI	rule8	rule8	rule8	rule8	rule6	rule7
HRD	rule2	rule4	rule4	rule5	rule5	rule5

重回帰分析 , RF5 法 , RF6.4 法 , MLPx , nc-MR 法で得られた回帰式のグラフをそれぞれ図 7.2 ~ 7.6 に示す . ただし RF6.4 法 , nc-MR 法では , それぞれの回帰式に属するサンプルの色を , その回帰式の線と同じ色にした .

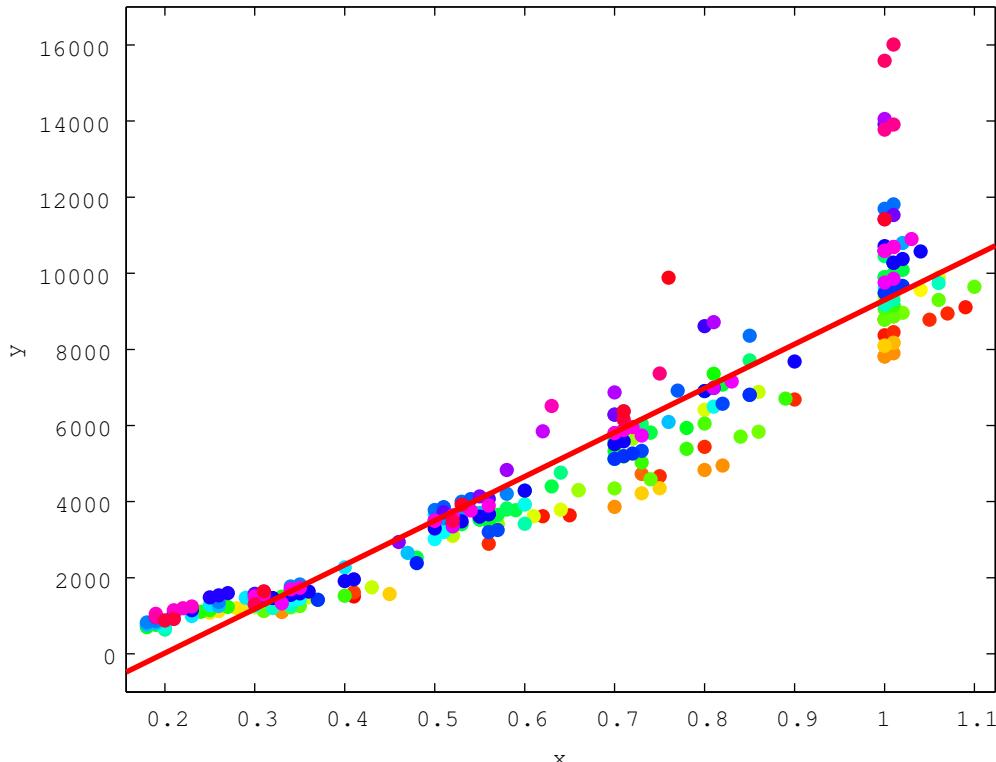


図 7.2: 回帰式のグラフ (重回帰分析 , ダイヤモンドデータ)

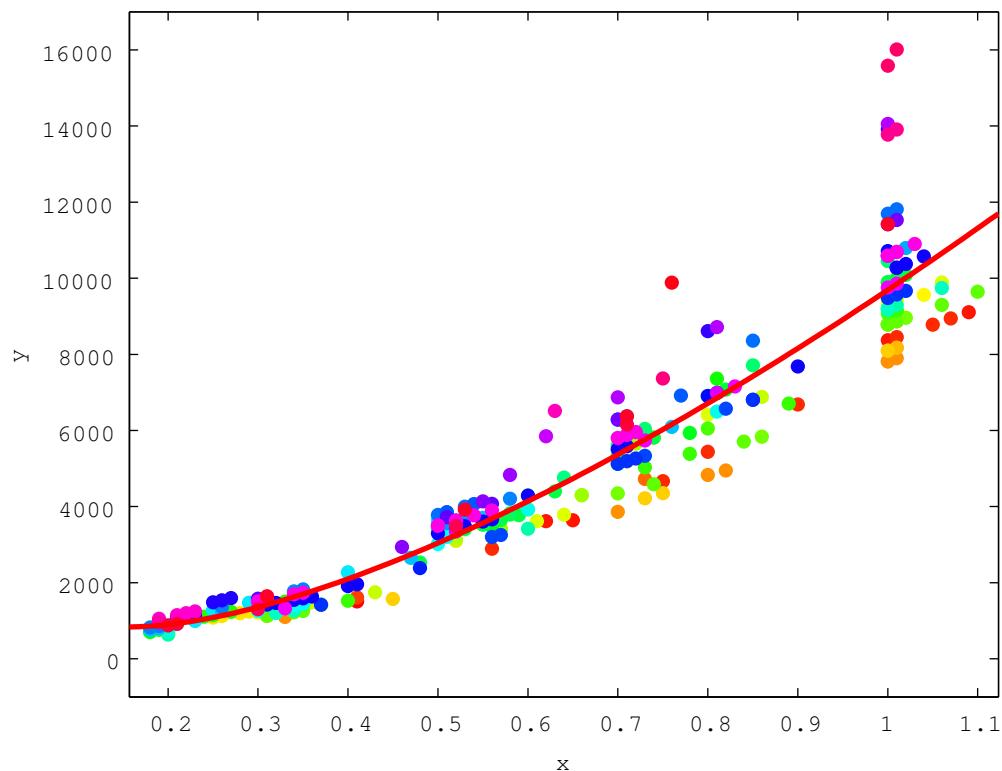


図 7.3: 回帰式のグラフ (RF5法, ダイヤモンドデータ)

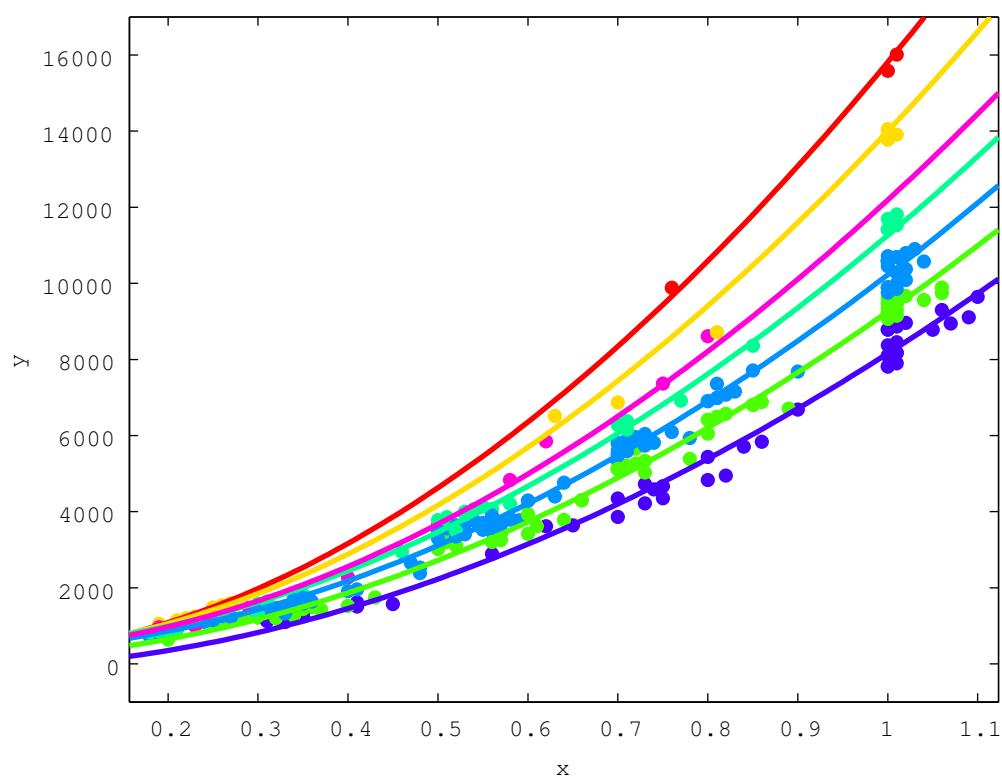


図 7.4: 回帰式のグラフ (RF6.4法, ダイヤモンドデータ)

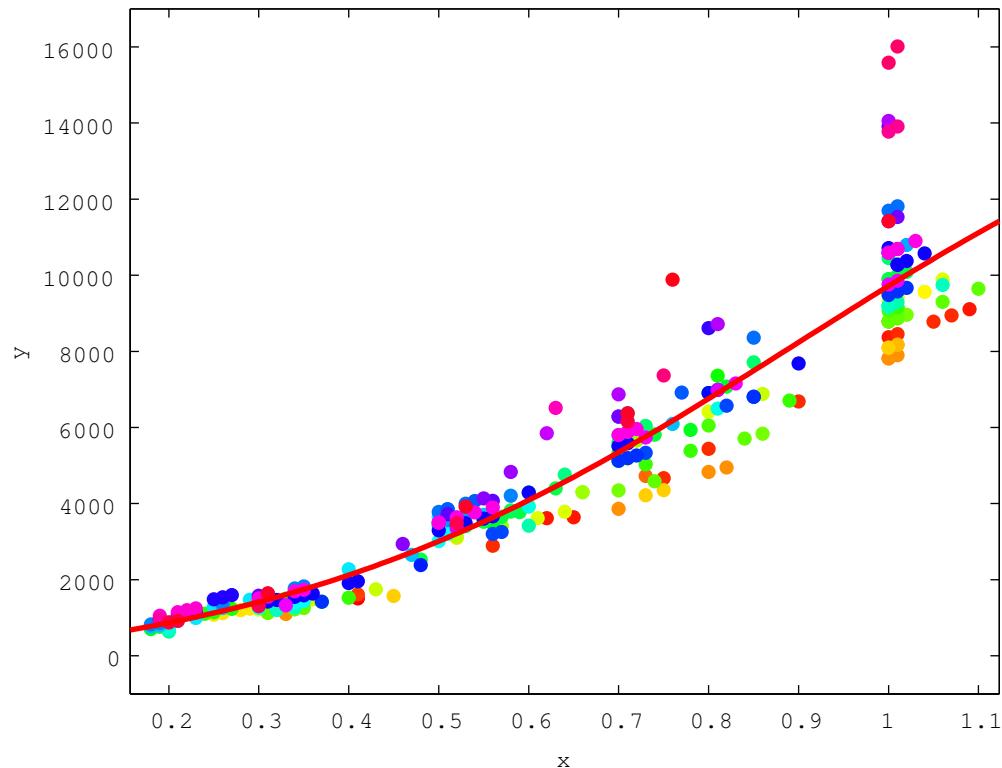


図 7.5: 回帰式のグラフ (MLPx, ダイヤモンドデータ)

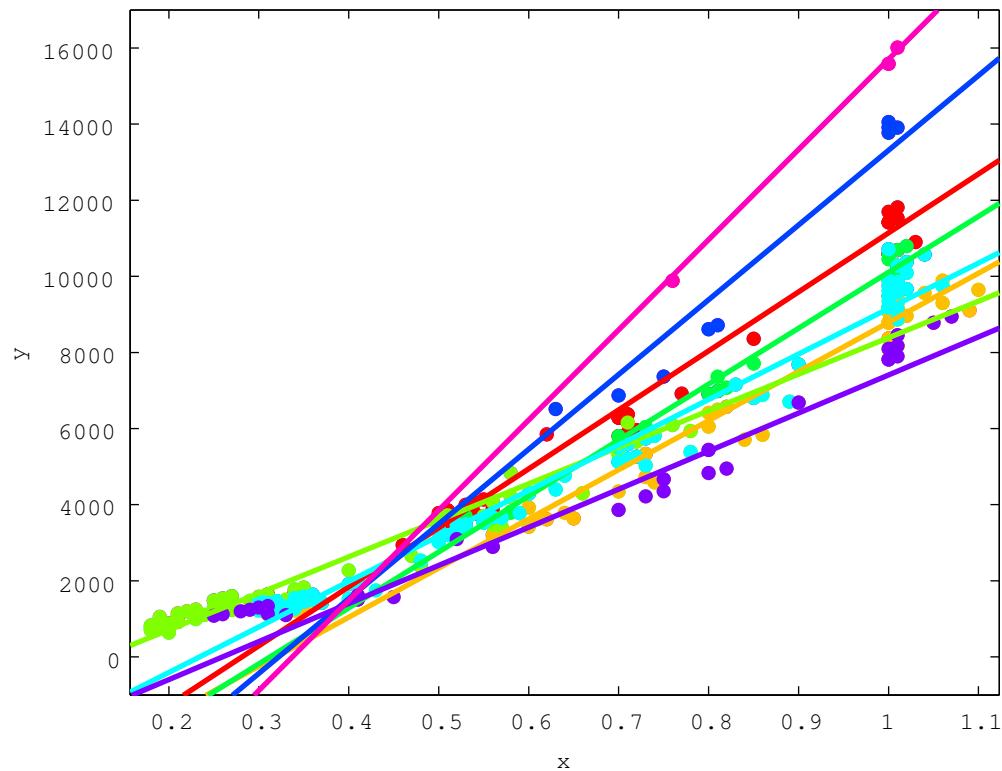


図 7.6: 回帰式のグラフ (nc-MR法, ダイヤモンドデータ)

それぞれの回帰モデルにおける交差検証誤差の比較を図 7.7 に示す。上の 4 つのモデルは量的変数のみを用いるモデルで、下の 5 つは量的変数と質的変数を両方用いるモデルである。

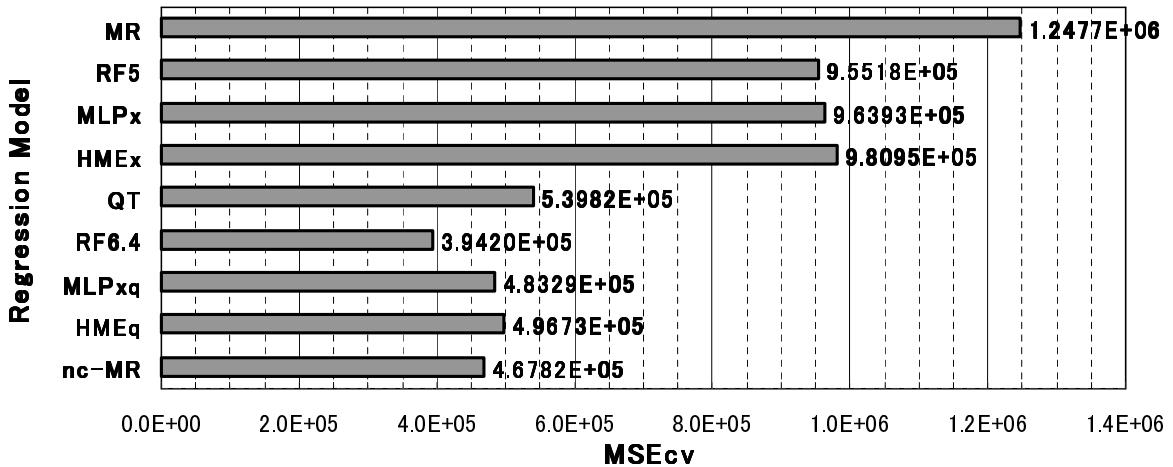


図 7.7: 交差検証誤差の比較 (ダイヤモンドデータ)

7.2.3 考察

まず表 7.1 ~ 7.6 での最適中間ユニット数の選択では、MLPx と nc-MR 法が 3 つの表全てで同じ J の値を最適中間ユニット数として選択できている。RF5 法、MLPxq, RF6.4 法では 3 つの表全てで同じ値を選択できているわけではないが、交差検証誤差が最小となる J, R の値を、BIC が最小となった回数または最小値の表のどちらかでは選択できている。全体的に最適中間ユニット数としては小さめの値が選択されているが、これはデータの入力が 1 变数しかなく、式があまり複雑でなくても十分にデータの挙動を説明できるからであると考えられる。

表 7.7 ~ 7.9 での最適ルール数の選択では、交差検証法に比べて、BIC で少し大きめの I の値を選択していることがわかる。汎化性能の面でも、また式の可読性の面でも、このデータに関しては交差検証誤差が最小となる I を最適ルール数として選択した方が良い。

次に、各モデルごとに得られた式とそのグラフについて考察する。まず、量的変数のみを用いるモデルである重回帰分析、RF5 法、MLPx では、 y の平均値付近をグラフが通っていることがわかる。特に、非線形である RF5 法と MLPx ではほとんどの形のグラフとなった。図 7.1 からわかるように、 x の値が大きいところでは、 x が同じ値でも質的変数によって y の値が大きく異なっているため、質的変数を用いないモデルでは、明らかに良い性能が得られない。図 7.7 の交差検証誤差

の比較でも、量的変数のみを用いるモデルの性能が悪いことがわかる。

質的変数も用いるモデルについては、図7.7より、量的変数のみを用いるモデルに比べて高い汎化性能を得ていることがわかる。RF6.4法、nc-MR法で得られた回帰ルールの条件部を表す表7.10、7.11は共に2次元の表となっているが、これはRF6.4法では q_3 が、nc-MR法では q_2 が条件部に現れなかつたためである。

q_1, q_2 は順序のある質的変数となっており、そのような順序尺度を持つ変数をその順序も考慮して扱う方法もあるが、今回は順序を無視してモデルに適用した。しかし、数量化理論一類では q_1, q_2 に対応するパラメータ値を見ると、その順序通りのパラメータ値を得られていることがわかる。更にRF6.4法、nc-MR法でも同様に、 q_1, q_2 の順序がルールに現れていることを確認できる。まずRF6.4法では、回帰式(7.5)が多項式の定数項ではない部分の係数値の順に並べてある。そして、表7.10は、 q_1 では左から、 q_2 では上からグレードの高い順になっているが、対応するルール番号を見ると、どの行または列も、ルール番号の降順に並んでいることがわかる。すなわち、グレードの高い質的変数に対応する回帰式の係数値が大きくなるような条件部となっている。なお、定数項の値は非常に小さく、ほとんど効いていないことがわかる。

nc-MR法でも同様に、回帰式(7.6)は多項式の定数項ではない部分の係数値の順に並べてあり、表7.11は左から q_1 のグレードが高い順になっている。そして、ほとんど部分においてルール番号が左から降順に並んでいることがわかる。rule8だけはそのようになっておらず、グレードの高いところに現れてしまっているが、これは定数項の値が他のルールに比べて大きい値となっていて、それに伴って y も大きな値を取れるからである。図7.6ではrule8は黄緑色の線に対応しており、その様子がよくわかる。

図7.4、7.6より、RF6.4法及びnc-MR法は x が大きい部分の y の広がりをうまくカバーしていることがわかる。そのため、図7.7でも高い汎化性能が得られている。しかし、図7.4と7.6では明らかに7.4の方が、うまくデータの挙動を表現できている。このダイヤモンドデータに対しては、RF6.4法が最も有効であるといえる。

7.3 プロ野球選手の年俸データ

多変量データへの適用の例として、説明変数の内容がわかりやすいプロ野球選手の年俸データを用いた検証を行う。

7.3.1 実験における諸設定

実験データ

プロ野球選手の年俸データ²は、2005年度のプロ野球選手（野手）の成績と年俸に関するデータで、6つの量的説明変数 x_1, \dots, x_6 、2つの質的説明変数 q_1, q_2 及び目的変数 y から成る。量的説明変数としては、 x_1 は年齢、 x_2 はプロでのプレイ年数、 x_3 は前年度年俸、 x_4 は安打数、 x_5 は本塁打数、 x_6 は盗塁数をそれぞれ表す。質的説明変数としては、 q_1 は12の球団（T, D, B, S, G, C, M, H, L, O, F, E）を、 q_2 は守備位置（捕手、内野手、外野手）を表す。目的変数 y は選手の年俸を表す。総打席数が10未満の選手データ及び欠損値のあるデータを除外し、211個のサンプルを用いた ($N = 211$)。

実験方法

パラメータ等の設定は、ダイヤモンドデータを用いた実験と同じである。ただし、RF5法、RF6.4法、MLPに対しては、BCW法による重み共有を適用した。なお、学習の際には量的変数に式(7.1)の正規化を施した。

7.3.2 実験結果

RF5法、MLP、nc-MR法において、 $BIC(J)$ の値が最小となった回数を表7.12に、 J の値ごとの $BIC(J)$ の最小値を表7.13に示す。交差検証誤差の最小値を表7.14に示す。同様に RF6.4法において、 $BIC(J, R)$ の値が最小となった回数、 $BIC(J, R)$ の最小値、及び $MSE_{CV}(J, R)$ の最小値をそれぞれ表7.15~7.17に示す。

表 7.12: $BIC(J)$ が最小となった回数 (RF5法・MLP・nc-MR法、野球データ)

J	1	2	3	4	5	合計
RF5 法	1	4	18	70	7	100
MLPx	0	2	10	57	31	100
MLPxq	0	0	9	43	48	100
nc-MR 法	11	42	30	9	8	100

²データは2005年度プロ野球選手名鑑より入手した。

表 7.13: $BIC(J)$ の最小値 (RF5 法・MLP・nc-MR 法, 野球データ)

J	1	2	3	4	5
RF5 法	329.5	330.7	302.1	263.2	276.8
MLPx	312.6	301.1	257.0	230.6	218.0
MLPxq	339.4	286.4	229.4	207.5	194.5
nc-MR 法	300.8	281.9	262.5	278.5	287.6

表 7.14: $MSE_{CV}(J) (\times 10^7)$ の最小値 (RF5 法・MLP・nc-MR 法, 野球データ)

J	1	2	3	4	5
RF5 法	1.4390	0.9125	0.8711	0.8832	0.9274
MLPx	1.3591	1.0248	0.9833	0.7864	0.8195
MLPxq	1.4018	1.0721	0.8425	0.8543	0.7931
nc-MR 法	1.1901	0.8631	0.8642	1.2331	1.3207

表 7.15: $BIC(J, R)$ が最小となった回数 (RF6.4 法, 野球データ)

models	J = 1	J = 2	J = 3	J = 4	J = 5
R = 1	0	0	0	0	0
R = 2	0	0	0	20	48
R = 3	0	0	0	3	5
R = 4	0	0	0	0	9
R = 5	0	0	0	0	15

表 7.16: $BIC(J, R)$ の最小値 (RF6.4 法, 野球データ)

models	J = 1	J = 2	J = 3	J = 4	J = 5
R = 1	340.3	263.9	339.6	315.0	345.5
R = 2	492.9	336.4	295.0	208.2	233.1
R = 3	527.1	353.3	266.0	226.2	237.1
R = 4	566.4	512.7	298.3	255.3	238.7
R = 5	608.0	609.3	350.4	240.8	246.6

表 7.17: $MSE_{CV}(J, R) (\times 10^7)$ の最小値 (RF6.4 法, 野球データ)

models	J = 1	J = 2	J = 3	J = 4	J = 5
R = 1	1.7908	1.6294	1.3711	1.3027	1.2904
R = 2	1.7309	1.5525	1.2844	1.1937	1.2070
R = 3	1.7729	1.2793	0.9648	0.9511	0.9285
R = 4	1.6232	1.1847	0.8479	0.8617	0.8599
R = 5	1.9273	1.0440	0.8629	0.8532	0.8682

次に RF5 法 , RF6.4 法 , MLP において , BCW 法による共有重み構造でのニューラルネットの学習を行った . パラメータの初期値を変えて 100 回学習を行った後に , BCW 法を適用した . このとき $BIC(G)$ の値が最小となった回数を表 7.18 に , G の値ごとの $BIC(G)$ の最小値を表 7.19 に , G の値ごとの $MSE_{CV}(G)$ の最小値を表 7.20 にそれぞれ示す .

表 7.18: $BIC(G)$ が最小となった回数 (RF5 法 · RF6.4 法 · MLP , 野球データ)

G	2	3	4	5	6	7	8	9	10	11	12	合計
RF5 法	0	0	0	0	0	3	59	26	12	0	0	100
RF6.4 法	0	0	0	0	0	0	1	5	63	17	14	100
MLPx	0	0	0	0	0	0	0	14	10	54	22	100
MLPxq	0	0	0	0	0	0	0	3	72	17	8	100

表 7.19: $BIC(G)$ の最小値 (RF5 法 · RF6.4 法 · MLP , 野球データ)

G	2	3	4	5	6
RF5 法	342.9	295.9	294.6	293.8	295.7
RF6.4 法	331.7	300.4	286.4	262.5	244.2
MLPx	361.9	322.8	314.3	300.1	289.4
MLPxq	368.8	352.7	338.1	309.9	288.9
	7	8	9	10	11
	295.8	294.7	297.1	299.7	302.2
	235.9	231.4	210.47	191.3	183.1
	262.5	249.8	222.2	217.8	206.0
	261.7	226.5	212.8	193.6	186.3
	12				
					306.1
					187.2
					193.4
					185.2

表 7.20: $MSE_{CV}(G)$ ($\times 10^7$) の最小値 (RF5 法 · RF6.4 法 · MLP , 野球データ)

G	2	3	4	5	6
RF5 法	6.8711	3.2122	1.7045	1.2843	0.9637
RF6.4 法	8.6285	6.1612	2.1884	1.9277	1.5403
MLPx	4.1864	3.5101	1.4245	1.1223	0.9943
MLPxq	3.7931	2.0274	1.7364	1.5284	0.9374
	7	8	9	10	11
	0.8602	0.8694	0.8742	0.8761	0.8702
	1.1427	0.9628	0.8451	0.8463	0.8472
	0.7911	0.7821	0.7702	0.7692	0.7586
	0.8830	0.8321	0.7862	0.7771	0.7729
	12				
					0.8752
					0.8520
					0.7649
					0.7752

RF6.4+BCW 法及び nc-MR 法において, I の値ごとにパラメータの初期値を変えて 100 回ずつ学習を行ったときの $BIC(I)$ の値が最小となった回数, $BIC(I)$ の最小値, 及び $MSE_{CV}(I)$ の最小値をそれぞれ表 7.21 ~ 7.23 に示す. ただし, RF6.4+BCW 法及び nc-MR 法では, 表 7.14, 7.17 において交差検証誤差が最小になるときの R, J の値を用いた. また, HMEx 及び HMEq では, $I = 1$ のときは通常の重回帰分析を行った.

表 7.21: $BIC(I)$ が最小となった回数 (RF6.4+BCW 法・HME・nc-MR 法, 野球データ)

I	1	2	3	4	5	6	7	8	9	10	合計
RF6.4 法	0	0	0	0	0	7	63	16	10	4	100
HMEx	0	0	0	0	0	0	1	59	30	10	100
HMEq	0	0	0	0	0	0	0	0	71	29	100
nc-MR 法	0	0	0	0	0	0	6	51	14	29	100

表 7.22: $BIC(I)$ の最小値 (RF6.4+BCW 法・HME・nc-MR 法, 野球データ)

I	1	2	3	4	5
RF6.4 法	-590.4	-591.5	-594.2	-613.9	-621.5
HMEx	50.2	-14.3	-97.3	-98.2	-105.3
HMEq	50.2	-4.4	-88.3	-84.8	-99.2
nc-MR 法	-370.4	-548.8	-623.1	-671.8	-692.6
	6	7	8	9	10
	-610.5	-626.7	-603.4	-594.9	-597.2
	-135.5	-176.3	-183.2	-191.2	-189.4
	-113.7	-142.3	-177.0	-190.4	-184.2
	-698.0	-705.1	-697.2	-697.4	-718.7

表 7.23: $MSE_{CV}(I) (\times 10^7)$ の最小値 (RF6.4+BCW 法・HME・nc-MR 法, 野球データ)

I	1	2	3	4	5
RF6.4 法	2.3283	2.0267	1.6873	0.8523	0.8241
HMEx	0.8481	0.8502	0.8491	0.8605	0.8521
HMEq	0.8481	0.8404	0.8295	0.8277	0.8281
nc-MR 法	1.2741	1.0315	0.9155	0.8650	0.8677
	6	7	8	9	10
	0.8298	0.8332	0.8316	0.8389	0.8351
	0.8462	0.8224	0.8300	0.8316	0.8453
	0.8272	0.8302	0.8231	0.8142	0.8191
	0.8409	0.7868	0.7692	0.7987	0.7871

全データを用いて、重回帰分析、数量化理論一類、RF5+BCW法、RF6.4+BCW法、MLPx+BCW法、及びnc-MR法によって得られた、正規化された変数に対する回帰式を以下に示す。ただし、RF5法、MLPx、RF6.4法、nc-MR法では、表7.14, 7.17, 7.20, 7.23において交差検証誤差が最小になるときの R, J, I の値を用いた。また、数量化理論一類では、各質的変数 q_1, q_2 の中で最も値が低いカテゴリに対するパラメータ値が0となるようにした。

重回帰分析

$$y = -0.971 + 0.275x_1 - 0.061x_2 + 4.545x_3 - 0.139x_4 + 1.204x_5 + 0.836x_6 \quad (7.7)$$

数量化理論一類

$$\begin{aligned} y = & -1.036 - 0.093x_1 - 0.134x_2 - 0.057x_3 - 0.053x_4 + 0.048x_5 - 0.042x_6 \\ & + 0.107q_{1,T} + 0.161q_{1,D} + 0.146q_{1,G} + 0.102q_{1,C} + 0.122q_{1,B} + 0.145q_{1,H} \\ & + 0.040q_{1,E} + 0.005q_{1,F} + 0.332q_{1,L} + 0.062q_{1,M} + 4.646q_{1,O} \\ & + 1.179q_{2,\text{捕手}} + 0.785q_{3,\text{内野手}} \end{aligned} \quad (7.8)$$

RF5+BCW法

$$\begin{aligned} y = & -0.8925 + 3.2089(x_1x_3x_4x_5x_6)^{0.1886} \\ & + 2.0124x_1^{-1.1787}x_2^{0.1886}x_3^{1.0668} \\ & + 0.8996x_1^{0.1886}x_3^{1.0668}x_4^{-2.0971}x_5^{2.4101}x_6^{-0.2071} \end{aligned} \quad (7.9)$$

RF6.4+BCW法

$$\left\{ \begin{array}{ll} \text{if rule1 then } & y = -0.7346 + 0.7798h_1 + 3.0430h_2 + 1.2276h_3 \\ \text{if rule2 then } & y = -0.7905 + 5.2092h_1 + 3.7466h_2 - 0.4753h_3 \\ \text{if rule3 then } & y = -0.7567 + 0.2109h_1 + 4.3373h_2 + 0.6892h_3 \\ \text{if rule4 then } & y = -0.7410 + 1.6820h_1 + 4.0711h_2 + 0.3478h_3 \\ \text{if rule5 then } & y = -0.7628 - 0.5724h_1 + 2.8896h_2 + 1.6317h_3 \end{array} \right. \quad (7.10)$$

$$h_1 = x_1^{-0.1011}x_2^{0.6271}x_3^{-0.3678}x_4^{-2.6930}x_5^{3.6555},$$

$$h_2 = (x_1x_3x_4)^{0.4517},$$

$$h_3 = x_1^{-2.6930}(x_2x_6)^{-0.1011}x_3^{1.8900}x_4^{-0.4621}$$

表 7.24: 回帰ルールの条件部 (RF6.4+BCW法, 野球データ)

	T	D	B	S	G	C	M	H	L	O	F	E
捕手	rule1	rule3	rule3	rule3	rule3	rule3	rule3	rule4	rule5	rule1	rule2	rule5
内野手	rule1	rule4	rule1	rule4	rule4	rule1	rule4	rule4	rule5	rule1	rule2	rule1
外野手	rule1	rule3	rule3	rule3	rule3	rule3	rule3	rule4	rule5	rule1	rule2	rule5

nc-MR 法

$$\left\{
 \begin{array}{l}
 \text{if rule1 then} \\
 \quad y = -0.992 + 0.207x_1 - 0.042x_2 + 4.693x_3 + 0.105x_4 + 0.996x_5 + 0.635x_6 \\
 \text{if rule2 then} \\
 \quad y = -0.814 + 0.164x_1 - 0.130x_2 + 4.243x_3 + 0.085x_4 + 0.533x_5 + 0.528x_6 \\
 \text{if rule3 then} \\
 \quad y = -0.981 + 0.222x_1 + 0.253x_2 + 3.431x_3 + 0.108x_4 + 1.856x_5 + 0.606x_6 \\
 \text{if rule4 then} \\
 \quad y = -1.164 + 0.248x_1 + 0.044x_2 + 5.124x_3 + 0.125x_4 + 1.444x_5 + 0.738x_6 \quad (7.11) \\
 \text{if rule5 then} \\
 \quad y = -0.874 + 0.184x_1 - 0.012x_2 + 4.031x_3 + 0.093x_4 + 0.952x_5 + 0.558x_6 \\
 \text{if rule6 then} \\
 \quad y = -0.892 + 0.180x_1 - 0.142x_2 + 4.652x_3 + 0.094x_4 + 0.584x_5 + 0.579x_6 \\
 \text{if rule7 then} \\
 \quad y = -0.877 + 0.202x_1 + 0.279x_2 + 2.849x_3 + 0.097x_4 + 1.819x_5 + 0.538x_6 \\
 \text{if rule8 then} \\
 \quad y = -1.000 + 0.218x_1 + 0.125x_2 + 4.042x_3 + 0.108x_4 + 1.499x_5 + 0.628x_6
 \end{array}
 \right.$$

表 7.25: 回帰ルールの条件部 (nc-MR 法, 野球データ)

	T	D	B	S	G	C	M	H	L	O	F	E
捕手	rule5	rule5	rule5	rule5	rule2	rule2	rule3	rule5	rule7	rule4	rule8	rule2
内野手	rule1	rule1	rule1	rule6	rule6	rule6	rule8	rule6	rule3	rule4	rule1	rule6
外野手	rule5	rule5	rule5	rule5	rule2	rule2	rule3	rule5	rule7	rule4	rule8	rule2

nc-MR 法によって得られた式 (7.11) の回帰ルール集合における, 各回帰ルールの位置関係を Spring Model で表すと図 7.8 のようになった.

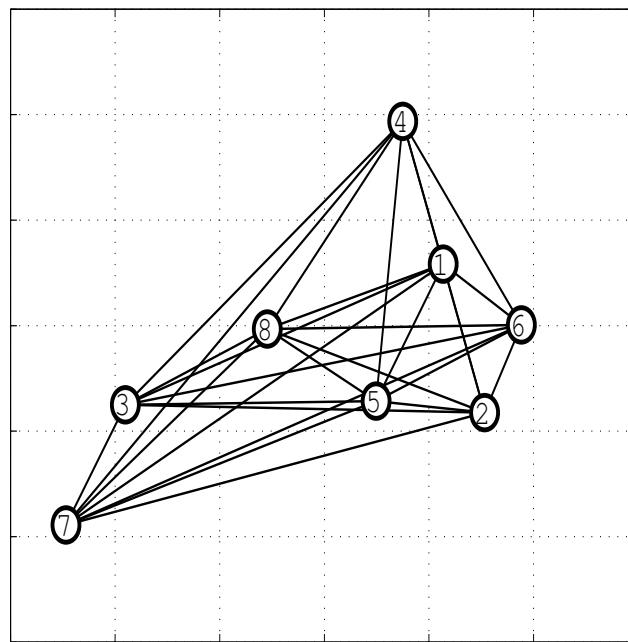


図 7.8: 各回帰ルールの位置関係 (nc-MR 法 , 野球データ)

それぞれの回帰モデルにおける交差検証誤差の比較を図 7.9 に示す . 上の 6 つは量的変数のみを用いるモデルで , 下の 7 つは量的変数と質的変数を両方用いるモデルである .

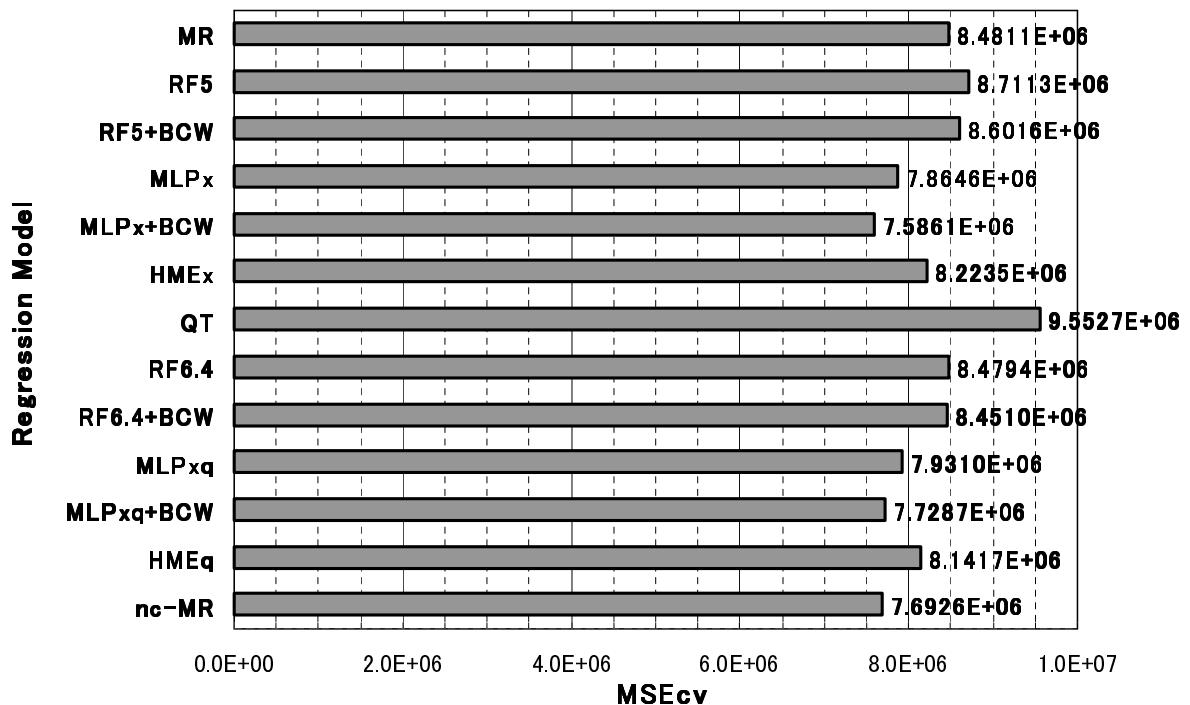


図 7.9: 交差検証誤差の比較 (野球データ)

7.3.3 考察

まず表7.12～7.17の最適中間ユニット数の選択では、ほとんどの回帰モデルで、多少のずれはあるがBICでの選択結果と交差検証誤差での選択結果が近いJの値を示している。しかし、BICの最小値による選択では、BICの最小になった回数や交差検証誤差による選択結果よりも大きめのJの値を選択していることがわかる。

共有重み構造でのクラスタ数G及び回帰ルール数Iの選択においても同様に、BICの最小値による選択方法では他の2つの選択方法に比べて大きめの値を選択していることがわかる。最終出力式の可読性を考えると、これらはなるべく値が小さい方が良い。従って、モデル選択の方法としてBICの最小値を用いるのは好ましくないといえる。

表7.14と7.17、及び7.20を比べると、重み共有をする前に比べて重み共有をした後の方が、交差検証誤差が減少していることがわかる。従って、重み共有によってパラメータの取れる値に制約を与えることで、データに含まれるノイズへのオーバーフィットを減らすことができているといえる。更にRF6.4法とnc-MR法では、係数をベクトル量子化することによっても交差検証誤差を少し減少させることができている。

次に各モデルによって得られた回帰式について考察する。重回帰分析では、係数値の正負を考えると全体的に納得できる結果を得られている。安打数に対するパラメータが負になっているのは好ましくないが、前年度年俸や本塁打数に対するパラメータに対して値が小さく、あまり効いていないと思われる。

RF5+BCW法とRF6.4+BCW法では、重み共有によって比較的簡素な式が得られている。RF6.4法で得られた回帰ルール集合の条件部をみると、どのチームにおいても捕手と外野手が同じルールになるという結果を得られた。

nc-MR法では、係数値の正負を考えるとどの回帰ルールにおいても納得のできる結果を得られている。どの選手も前年度の年俸と大きく変わることは多くないと思われるので、前年度年俸に対するパラメータ値が大きくなっている。各回帰ルールの条件部をみると、RF6.4法で得られた回帰ルール集合と同様に、どのチームにおいても捕手と外野手が同じルールとなっていることがわかる。またチームごとに回帰ルールをみると、{T, D, B}と{G, C, E}及び{S, H}はそれぞれの中で全く同じ回帰ルール群が対応している。チーム〇では守備位置に関わらず同じ回帰ルールとなっており、他のチームよりもより前年度年俸を重視した年俸となっていることがわかる。同じチームでも守備位置によって回帰ルールが異なる

チームでは、具体的にはそのルールはチーム内で、rule1 と 5, rule5 と 6, rule2 と 6, rule3 と 8, rule3 と 7, rule1 と 8, というように分かれている。しかし、図 7.8 をみると、これらの回帰ルールは全て近い位置にあることがわかる。

各モデルごとの性能を比較すると、図 7.9 より MLPx+BCW が最も良い結果となり、その次に nc-MR 法が続く。しかし、MLPx は式の可読性は良くないので、それも考慮すると mc-MR 法が最もこのデータに適しているといえる。

7.4 様々なデータによる汎化性能の比較

7.4.1 実験における諸設定

実験データ

15種類の現実データ³ を用いて、提案モデルの汎化性能を比較する。実験に用いたデータを表 7.26 にまとめる。

表 7.26: 実験に用いた現実データ

データ名	目的変数の内容	N	J	L _k
automobile	自動車価格	194	13	20
b-carotene	ベータカロチンの量	315	10	2, 3, 3
ball-bearing	ボールベアリングの寿命	210	2	3, 4
bb-america	アメリカのプロ野球選手の年俸	337	12	2, 2, 2, 2
boston	ボストンの不動産価格	486	12	75, 2
cholesterol	コレステロールの量	297	5	2, 4, 2, 3, 2, 3, 3
college	大学の学費	1129	11	50, 2
cpu	cpu の性能	205	6	26
fish	川魚の水銀含有量	171	2	2, 15
iq-brain	子供の知能指数	20	5	2, 2
lung-cancer	肺がん患者の生きた日数	126	3	2, 4, 2
mpg	自動車の燃費	388	5	28
philadelphia	フィラデルフィアの不動産価格	99	3	5
wage	労働者の賃金	534	2	2, 2, 2, 3, 6, 3, 2
yokohama	横浜の不動産価格	558	4	2, 18, 5

³ automobile, boston, cpu データは UCI Machine Learning Repository より入手し, b-carotene, cholesterol, college, iq-brain, lung-cancer, mpg, philadelphia, wage データは StatLib より入手した。 ball-bearing, bb-america データは Jurnal of Statistics Education のホームページより入手し, fish データは Duke Statistics Data Set Collection より入手した。また, yokohama データは神奈川県のオフィシャルページより入手した。

実験方法

10-fold 交差検証法により、重回帰分析、数量化理論一類、多変量多項式回帰（RF5法、RF5+BCW法）、質的条件付き多変量多項式回帰（RF6.4法、RF6.4+BCW法）、3層パーセプトロン（MLPx、MLPxq）、HME（HME_x、HME_q）及びnc-MR法の性能を比較した。パラメータ等の設定は、ダイヤモンドデータ及び野球データを用いた実験と同じである。ただし、データによっては質的変数の数が多いものも含まれるので、RF6.4法及びnc-MR法におけるルール数は $I = 1, \dots, 20$ とした。なお、学習の際には量的変数に式(7.1)の正規化を施した。重回帰と数量化以外では、それぞれの J, R, G, I の値ごとに、パラメータの初期値を変えてそれぞれ10回ずつ学習を行った。

7.4.2 実験結果

各回帰モデルごとの、交差検証誤差の最小値によって選択された J^*, R^*, G^*, I^* の値を示す。ただしRF6.4法においては、 I_n^* は共有重み構造無しの場合の最適ルール数を表し、 I_s^* はBCW法を適用した場合の最適ルール数を表す。また、HMEにおいては、 I_x^* はHME_xの最適ルール数を表し、 I_q^* はHME_qの最適ルール数を表す。

表 7.27: 各回帰モデルにおける J^*, R^*, G^*, I^* (現実データ)

	RF5		RF6.4				NNx		NNxq		HME		nc-MR		
	J*	G*	J*	R*	G*	I _n *	I _s *	J*	G*	J*	G*	I _x *	I _q *	R*	I*
automobile	3	5	3	5	8	10	10	2	7	1	7	4	8	1	11
b-carotene	2	7	1	4	4	6	7	2	6	2	8	4	6	3	6
ball-bearing	4	5	4	3	5	4	4	4	4	3	6	5	2	2	2
bb-america	3	8	2	5	7	4	4	1	6	2	7	3	7	3	6
boston	1	4	2	2	12	10	10	2	11	3	12	8	8	2	20
cholesterol	2	6	3	3	6	7	8	5	8	4	10	3	6	2	10
college	2	10	2	3	9	9	9	4	12	2	12	6	8	4	16
cpu	3	5	3	3	6	6	6	3	6	3	6	3	4	2	4
fish	4	5	4	4	6	7	7	2	3	2	6	4	4	3	5
iq-brain	4	6	5	3	5	3	4	2	5	1	5	4	3	3	3
lung-cancer	4	4	3	5	5	4	4	4	5	3	5	7	6	5	7
mpg	2	6	2	4	6	5	5	3	6	4	11	6	8	3	12
philadelphia	5	7	3	5	3	3	3	5	7	4	7	5	5	3	3
wage	5	5	4	3	5	8	8	4	6	2	7	6	7	4	9
yokohama	3	9	4	2	5	7	6	3	8	3	9	8	8	4	11

各データごとの、回帰モデルによる汎化性能を比較する。 J, R, G, I を表7.27の値としたときの交差検証誤差の比較結果を、図7.10～7.24に示す。

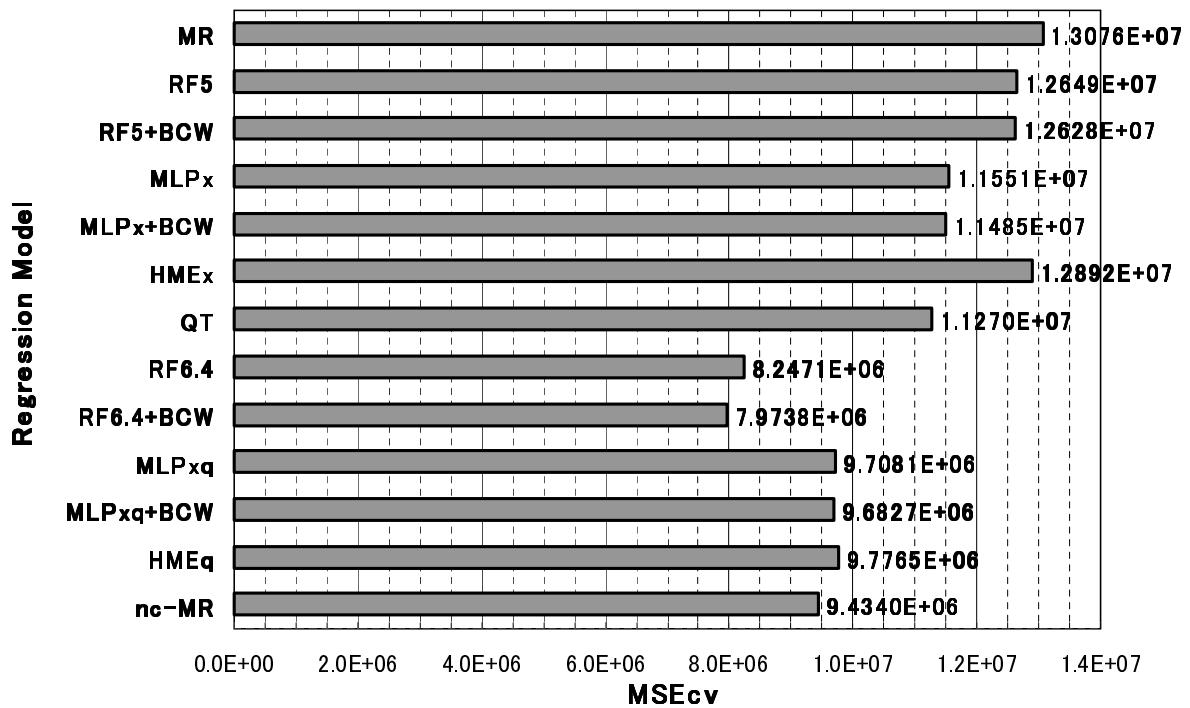


図 7.10: 交差検証誤差の比較 (automobile データ)

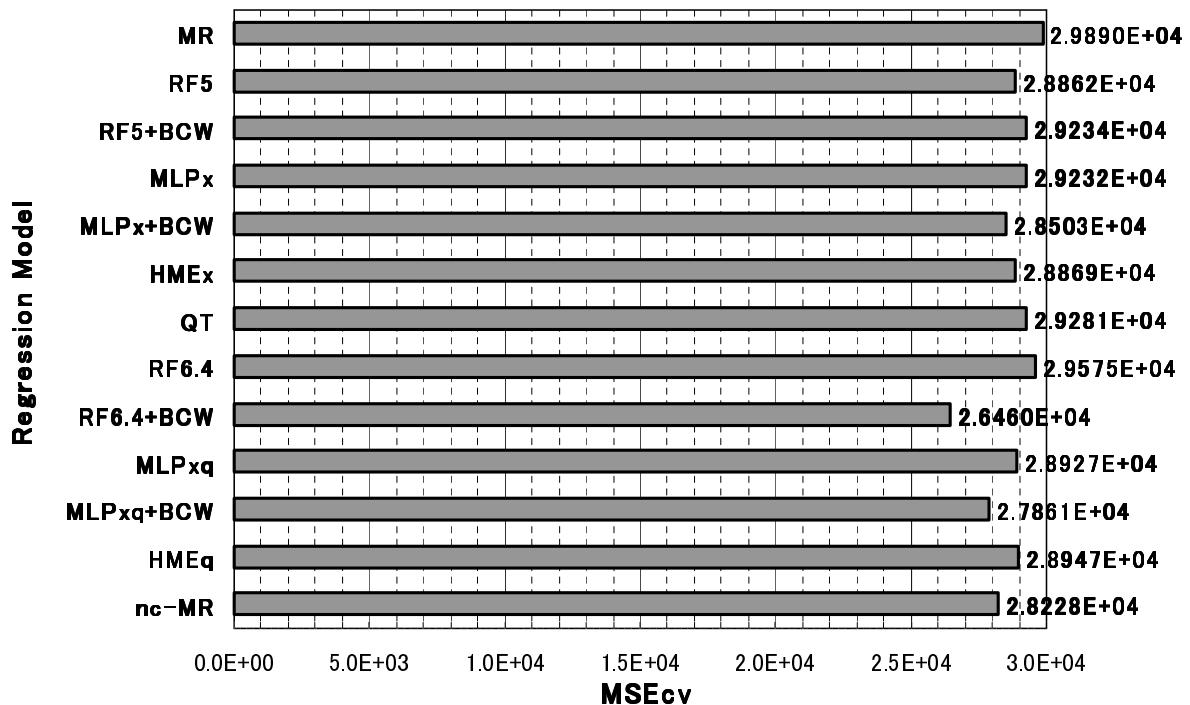


図 7.11: 交差検証誤差の比較 (b-carotene データ)

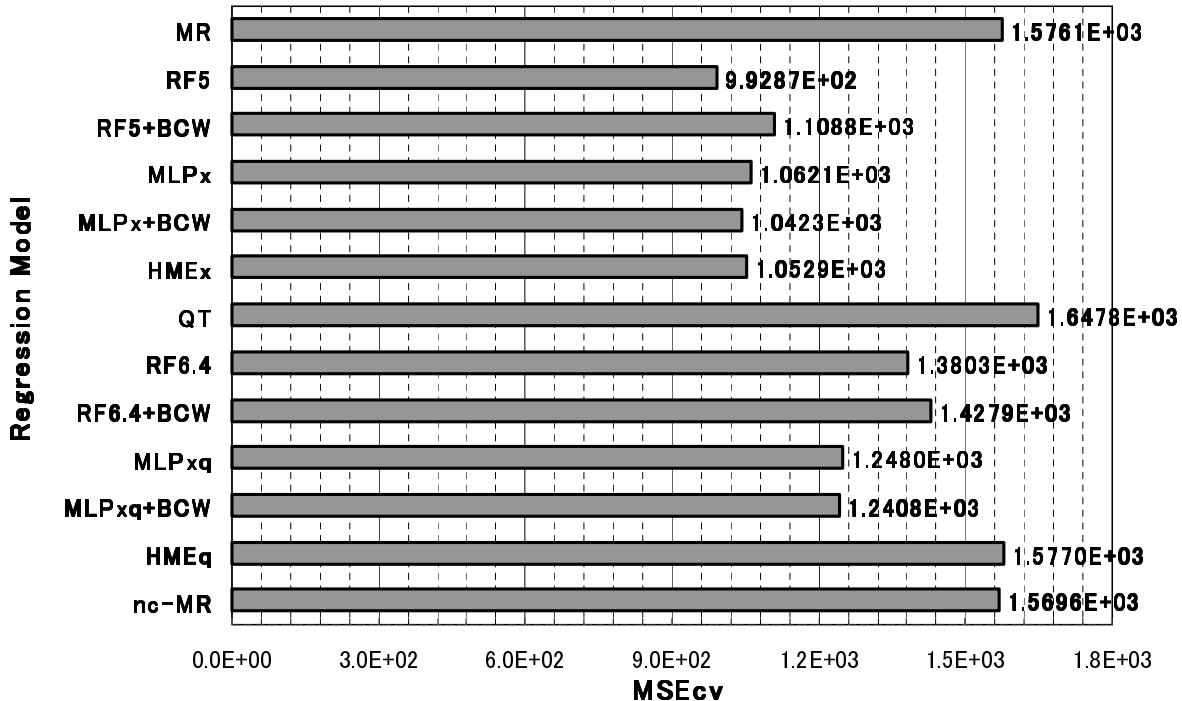


図 7.12: 交差検証誤差の比較 (ball-bearing データ)

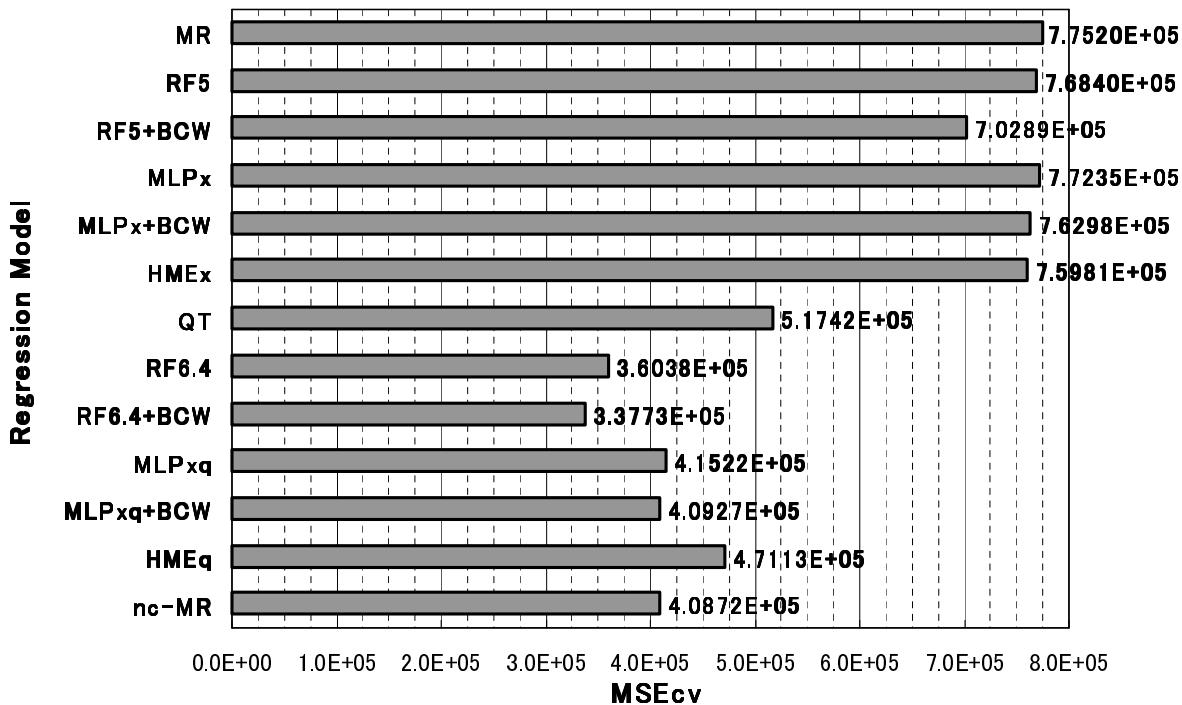


図 7.13: 交差検証誤差の比較 (bb-america データ)

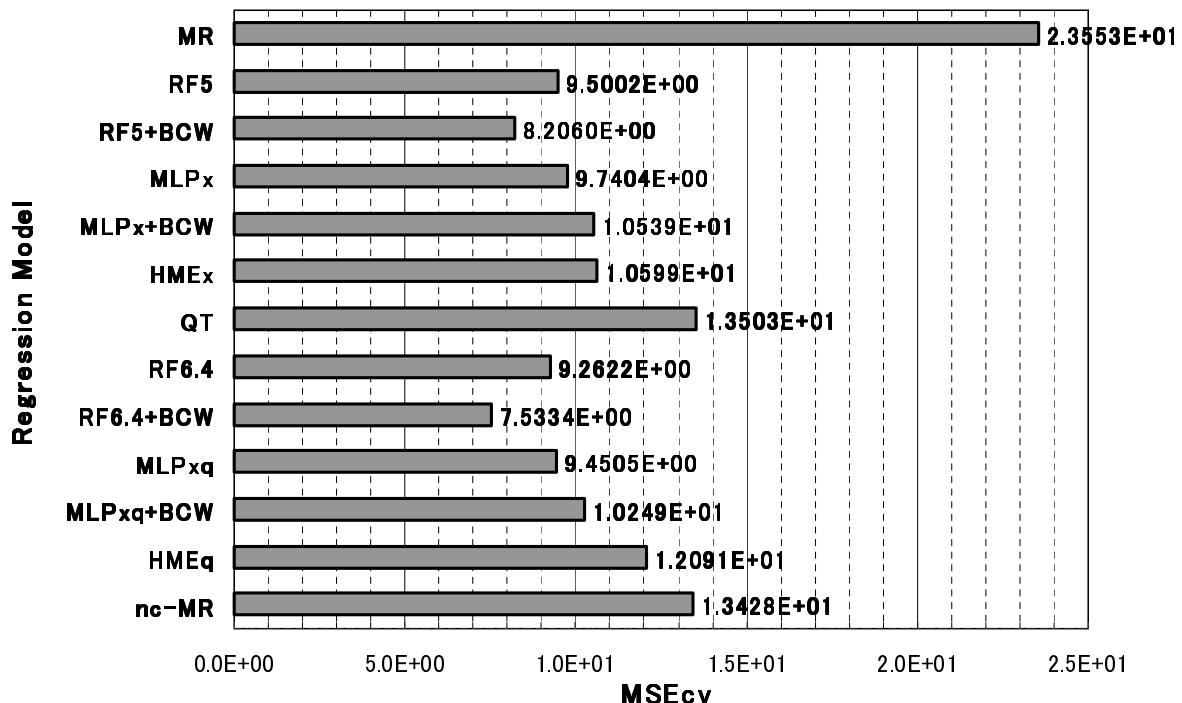


図 7.14: 交差検証誤差の比較 (boston データ)

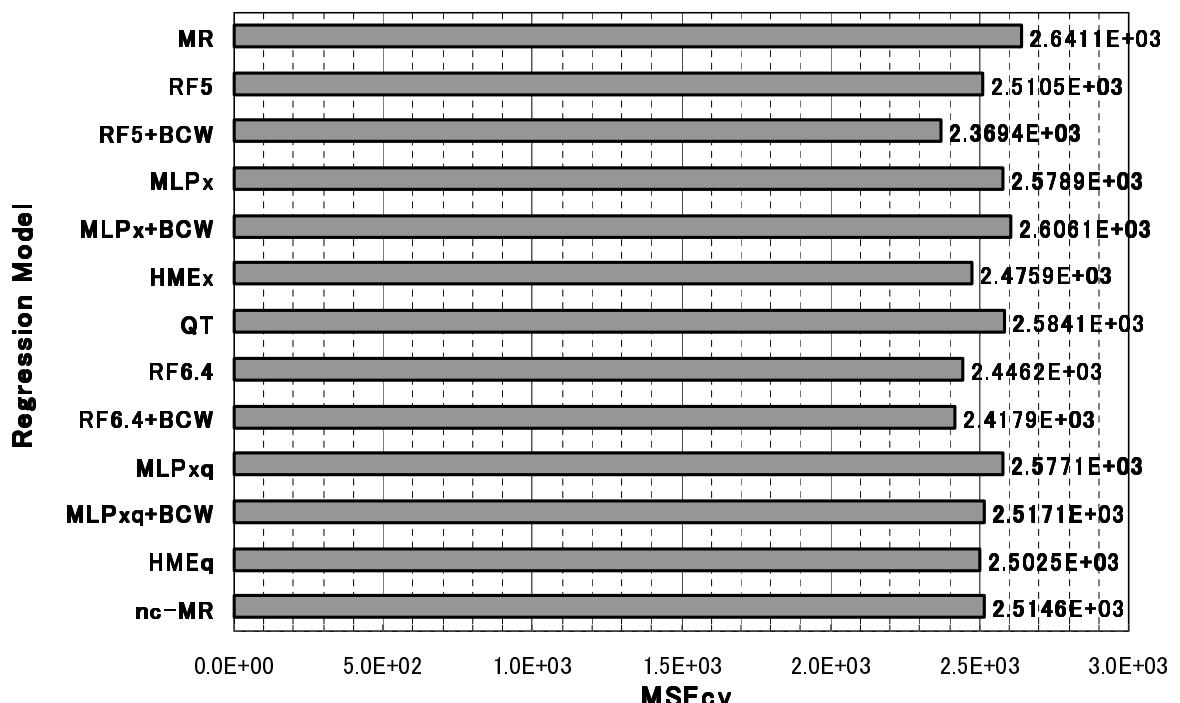


図 7.15: 交差検証誤差の比較 (cholesterol データ)

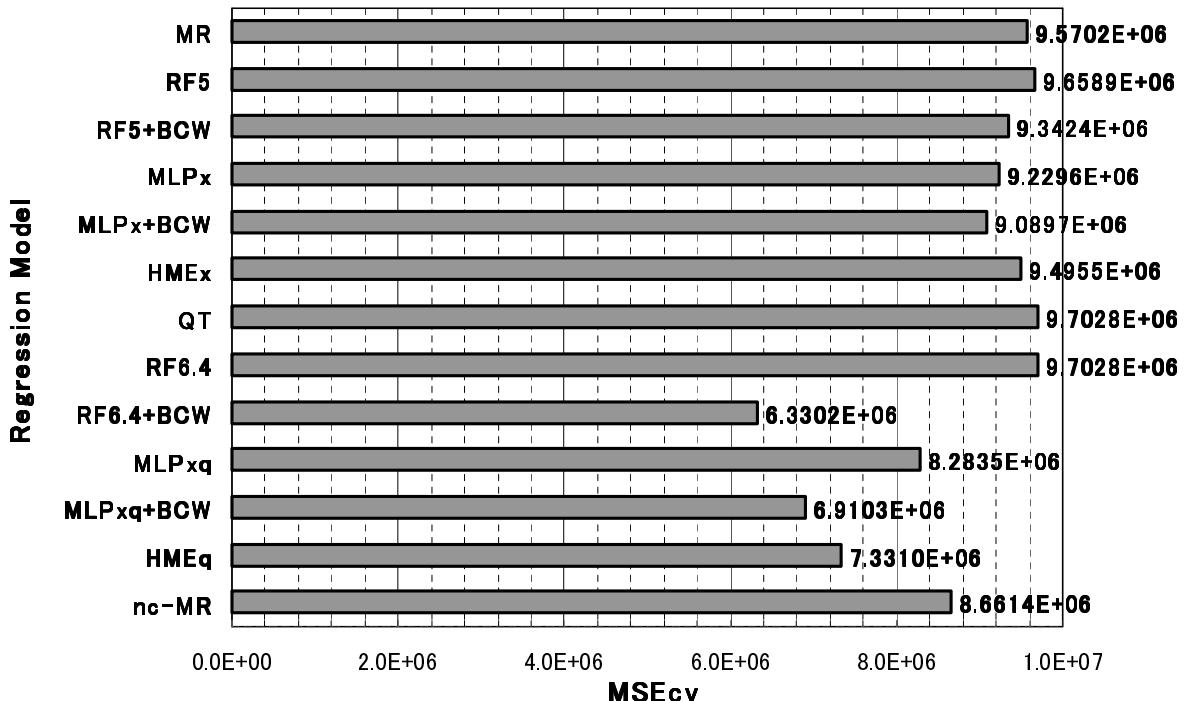


図 7.16: 交差検証誤差の比較 (college データ)

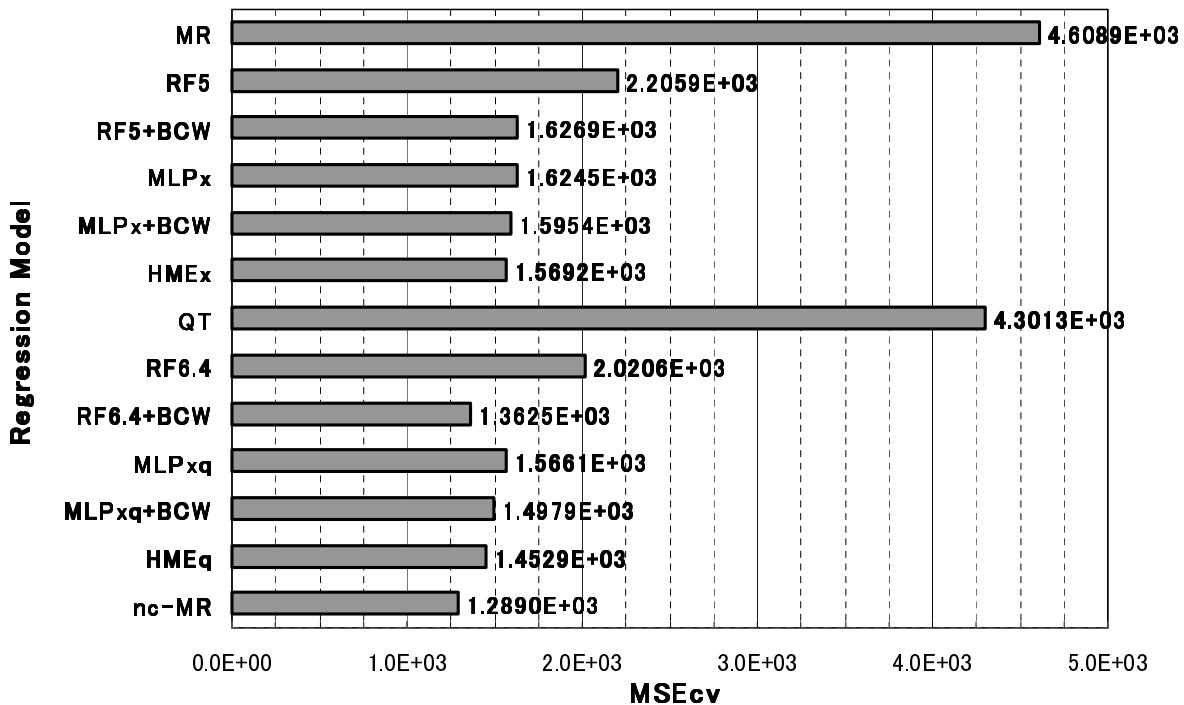


図 7.17: 交差検証誤差の比較 (cpu データ)

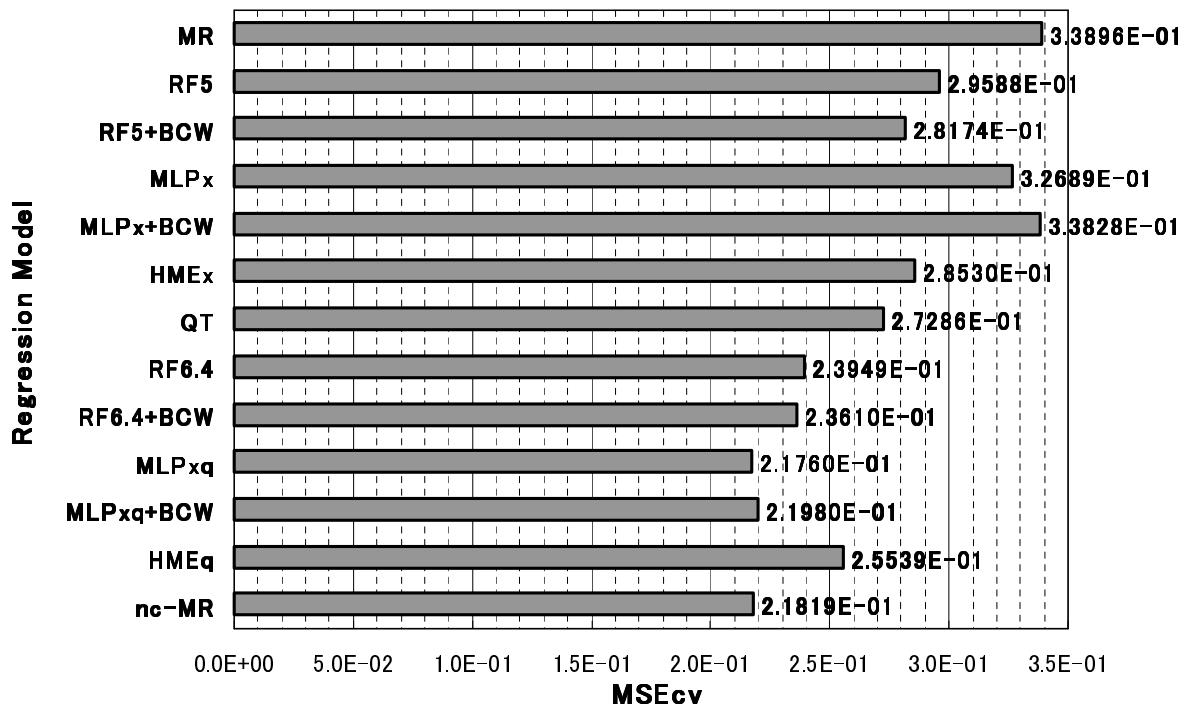


図 7.18: 交差検証誤差の比較 (fish データ)

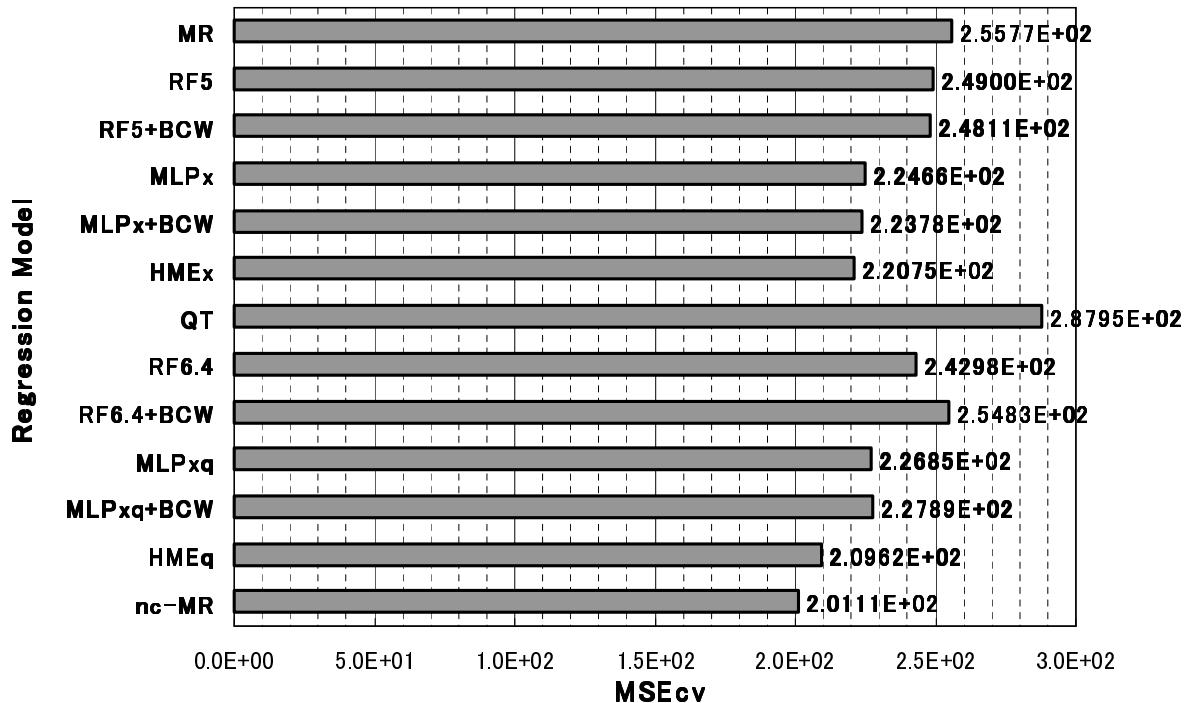


図 7.19: 交差検証誤差の比較 (iq-brain データ)

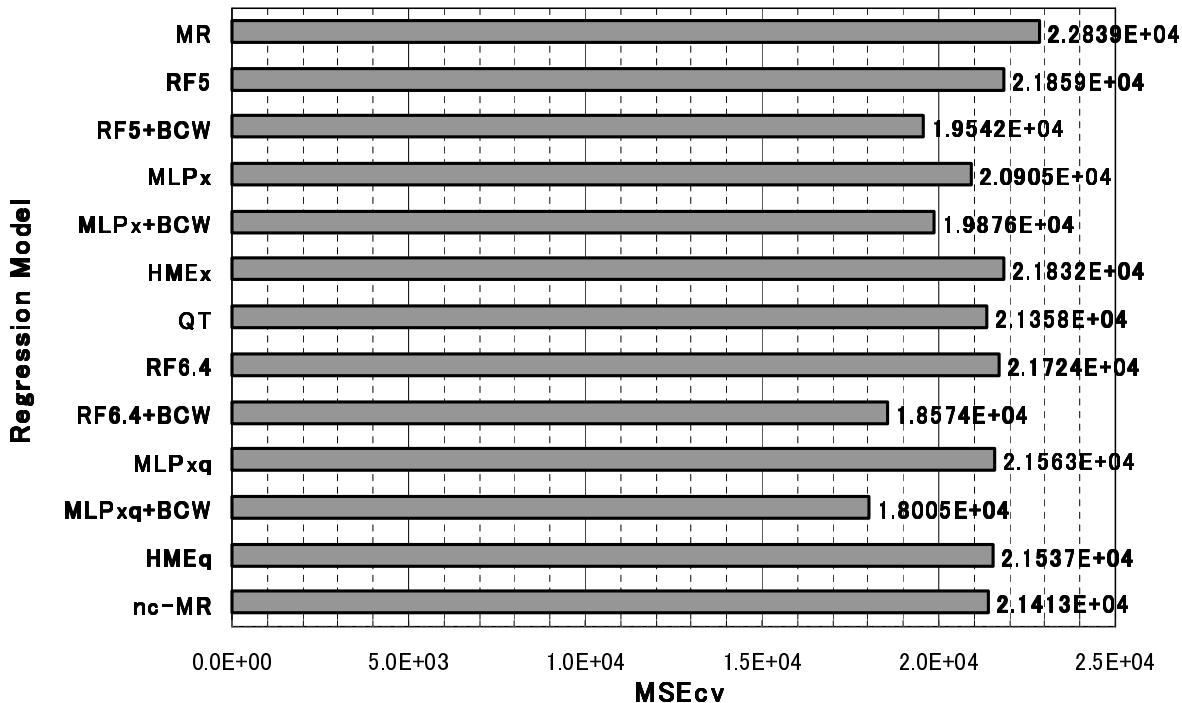


図 7.20: 交差検証誤差の比較 (lung-cancer データ)

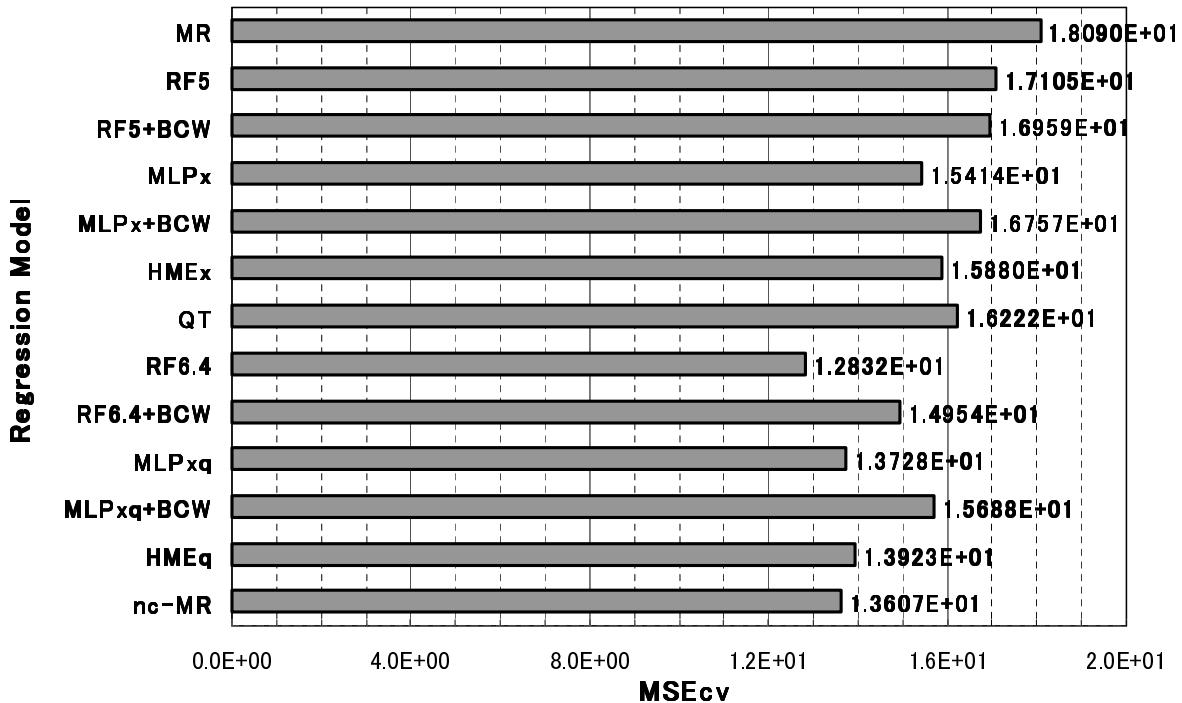


図 7.21: 交差検証誤差の比較 (mpg データ)

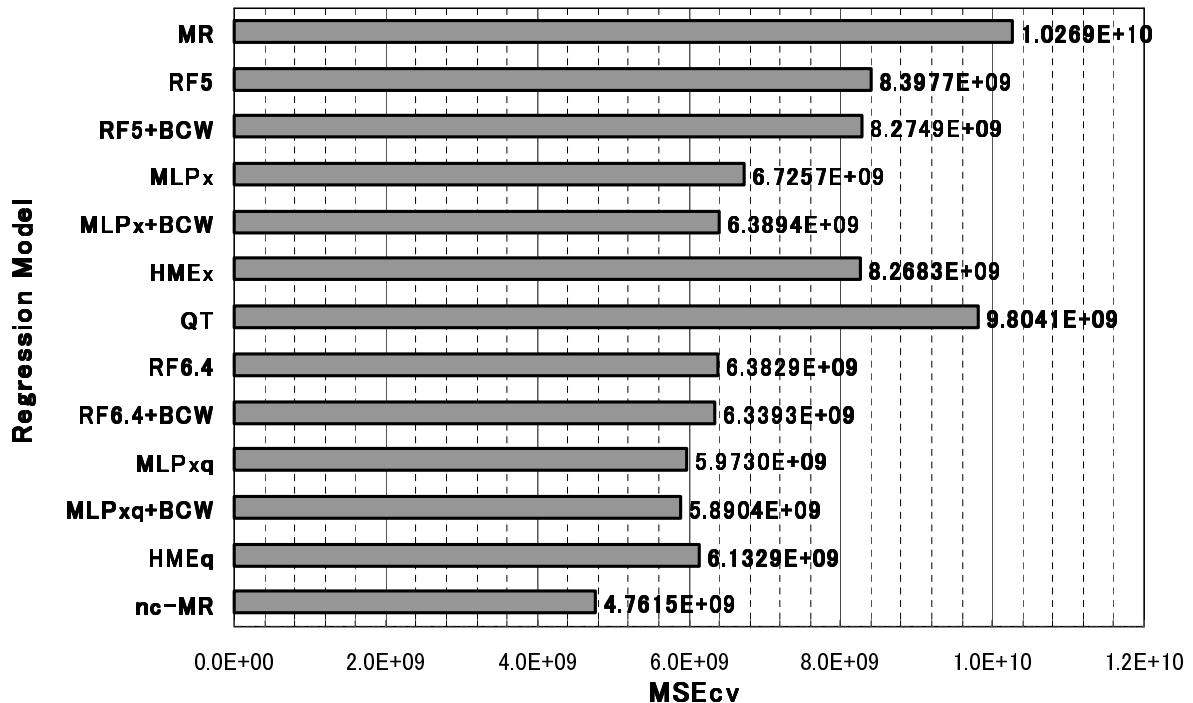


図 7.22: 交差検証誤差の比較 (philadelphia データ)

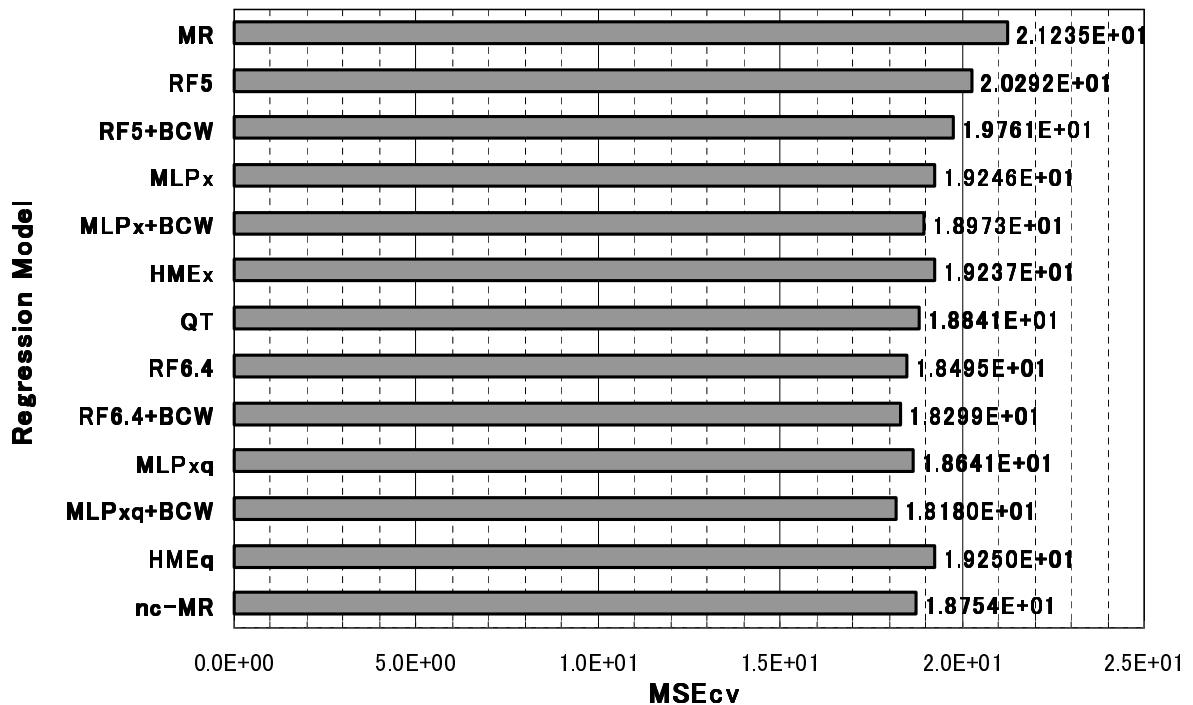


図 7.23: 交差検証誤差の比較 (wage データ)

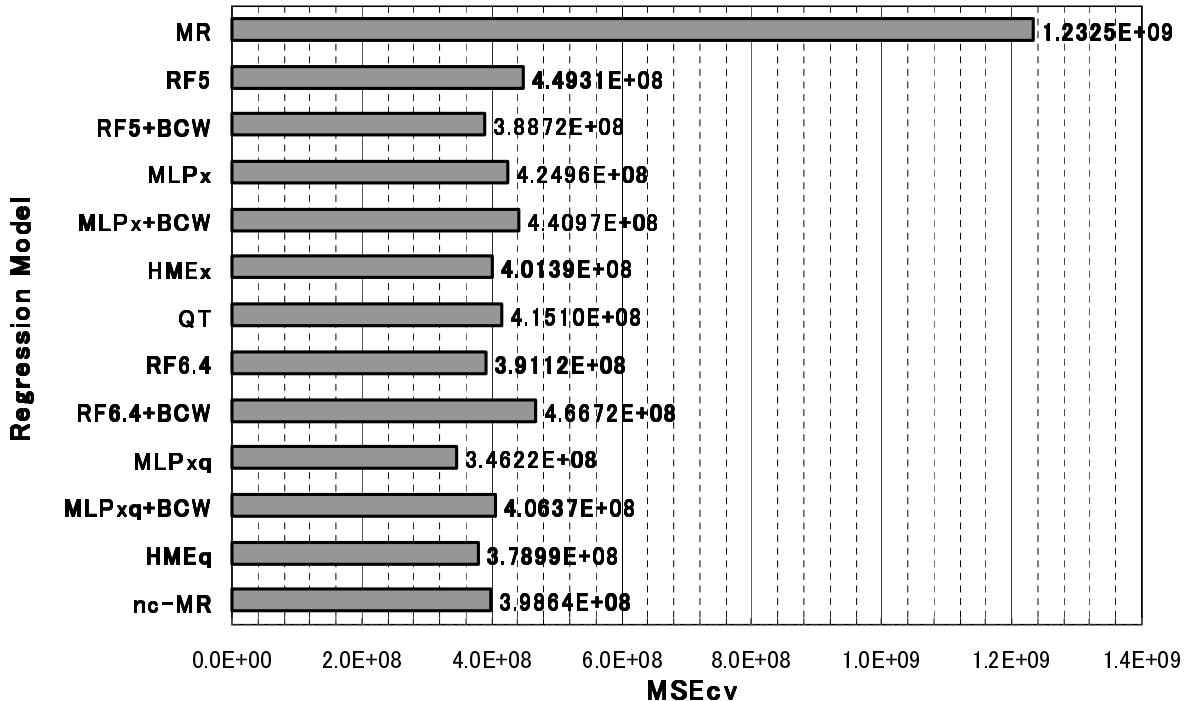


図 7.24: 交差検証誤差の比較 (yokohama データ)

RF5+BCW 法 , RF6.4+BCW 法 , MLPxq , HMEq , 及び nc-MR 法で学習に要した CPU 時間を表 7.28 に示す . 表の値は , それぞれの回帰モデルにおいて中間ユニット数を 1 から 5 まで変化させてパーセプトロンを学習したときの総計算時間で , 10 回の試行の平均値となっている . ただし BCW 法においては , 重み共有のクラスタ数を 2 から 12 まで変化させて , 双方向クラスタリングを 1 回のみ行った . また RF6.4 法においては J のみを変化させ . R の値は 3 に固定して学習を行った .

7.4.3 考察

図 7.10 ~ 7.24 における交差検証誤差の比較では , データによって各回帰モデルの汎化性能の傾向が大きく異なっていることがわかる . 多くのデータにおいて , BCW 法による重み共有を適用することによって汎化性能が向上していることが確認できる . また , HME と nc-MR 法は共に区分線形モデルであるが , 多くのデータにおいて nc-MR 法は HME に比べて良い汎化性能を得られていた .

各データごとに , 単純に最も性能が良かった回帰モデルをみると , RF5 系 (RF5 法 , RF5+BCW 法) では 2 つ , RF6 系 (RF6.4 法 , RF6.4+BCW 法) では 6 つ , MLP 系 (MLPx , MLPxq) では 4 つ , 区分線形系 (HME , nc-MR 法) では 3 つのデータで , 最も良い汎化性能が得られている . しかし , データによっては , 回帰モデルに

表 7.28: CPU 時間の比較 (秒)

データ名	RF5 +BCW	RF6.4 +BCW	MLPxq	HMEq	nc-MR
automobile	20.726	31.903	9.375	15.387	15.190
b-carotene	12.093	22.495	5.112	11.972	10.146
ball-bearing	24.958	31.359	11.428	18.039	16.728
bb-america	10.728	24.916	4.912	12.782	11.861
boston	159.281	221.064	74.258	133.212	117.956
cholesterol	7.184	8.423	3.214	8.092	6.606
college	199.039	441.711	111.494	209.937	200.084
cpu	11.400	20.814	5.580	7.950	12.590
fish	6.534	19.871	4.817	14.527	13.921
iq-brain	0.823	1.877	0.253	0.427	0.404
lung-cancer	6.322	8.719	3.427	3.842	3.951
mpg	12.518	18.982	6.662	15.343	10.237
philadelphia	5.827	6.194	2.487	3.118	3.102
wage	8.617	13.721	4.294	6.712	5.912
yokohama	11.622	29.440	5.114	9.586	16.424

ごとに多少は汎化性能に違いがあっても、ほとんど性能が変わらないようなデータもある。そこで、データを次のように大きく 6 つのグループに分けて、それぞれのグループごとに結果を考察する。

まず、タイプ I は、重回帰で十分と思われるデータで、b-catotene, cholesterol, wage データがこれに該当する。これらのデータでは、各方法において交差検証誤差の値に大差がない。従って量的変数だけを用いる線形回帰モデルで十分であり、質的変数が目的変数の予測をするのにあまり役立っていないといえる。また、非線形モデルや区分線形モデルを用いても回帰性能を大きく向上させることはできない。

タイプ II は、数量化で十分と思われるデータで、yokohama データがこれに該当する。このデータでは、重回帰では回帰性能が良くないが、数量化で一気に回帰性能が向上している。しかし、非線形モデルである MLP や区分的線形モデルである HME, nc-MR 法では数量化とあまり汎化性能が変わらない。すなわち、このタイプのデータでは量的変数に加えて質的変数を用いた線形回帰モデルで十分な汎化性能を期待できる。

タイプ III は、量的変数のみの非線形モデルが有効であると思われるデータで、

ball-bearing データがこれに該当する。このデータでは、質的変数を用いてもあまり汎化性能が変わらないが、RF5法やMLPxなどの量的変数のみの非線形モデルで良い汎化性能を得られている。式の可読性も考えると、このようなデータに対してはRF5法が有効であるといえる。

タイプIVは、量的変数と質的変数を両方用いた非線形モデルが有効であると思われるデータで、automobile, bb-america, boston, college データがこれに該当する。これらのデータでは、線形モデルや区分線形モデルに比べて、質的変数も用いる非線形モデルであるRF6.4法やMLPxqで良い汎化性能を得られている。式の可読性も考えると、このようなデータに対してはRF6.4法が有効であるといえる。

タイプVは区分線形モデルが有効に働くデータで、cpu, fish, iq-brain, mpg, philadelphia データがこれに該当する。これらのデータでは、nc-MR法やHMEqのような質的変数を用いる区分線形モデルで良い汎化性能を得られている。cpu, fish, mpg データでは、MLPxq や RF6.4 法でも良い結果が得られているが、線形の方が式の可読性が良いので、同様の性能であれば区分線形モデルの方が適しているといえる。全体的に HME より nc-MR 法の方が良い性能が得られているので、このタイプのデータに対しては nc-MR 法が最も有効であるといえる。

最後に、タイプVIは重み共有法が有効に働くデータで、lung-cancer データがこれに該当する。このデータでは、全体的には同様の汎化性能しか得られないが、BCW法による重み共有を適用したモデルで汎化性能が向上している。

表7.28より、多くのデータでnc-MR法はHMEより少ない計算時間で学習できていることがわかる。従って区分線形モデルとしては、nc-MR法はHMEより汎化性能と計算時間の両面で優れているといえる。RF5+BCW法とRF6.4+BCW法では、共有重みのクラスタが変わることに繰返し学習が必要となるので、他のモデルに比べて計算時間はかかるてしまう。しかし、RF6.4+BCW法では多くてもMLPxqの4倍程度の計算時間となっているので、回帰ではリアルタイム処理をあまり必要としないことを考えれば、十分実用に耐えうる範囲であるといえる。

7.5 まとめ

様々な現実データを用いた実験によって、提案モデルが現実データに対してても有効に働くかどうかを検証した。まず入力が1次元であるダイヤモンドの価格データによる検証では、実際にデータをプロットして回帰式のグラフを描画することによって、それぞれのモデルがどのようにデータにフィットしているかを見るこ

とができた。特に RF6.4 法のグラフは、サンプル点の y 軸方向への広がりにうまく対処できていることを確認できた。また、質的変数の順序を考えると、非常に納得のできる回帰式を得ることができたといえる。

野球選手の年俸データによる検証では、特に nc-MR 法で良い汎化性能を得ることができた。また、Spring Model による回帰ルールの位置関係を見ると、同じチームの回帰ルールが近い位置になるという、納得のいく結果を得ることができていた。非線形のモデルに対しては、BCW 法による重み共有がうまく働いて、可読性と汎化性を共に向上させることができていた。

他の現実データを用いた性能比較では、提案モデルで良い汎化性能を得られたデータが多くあった。また、多くのデータで BCW 法によって汎化性能も向上できることを確認した。データをいくつかのグループにわけることによって、どのようなタイプのデータにどの提案モデルが適しているのかを考察した。

第 8 章

あとがき

本研究では、可読性と汎化性の両面に優れた回帰分析を行うことを目的として、多層パーセプトロンを用いた回帰モデルについての研究を行った。そして、新たな3種のモデルを提案し、その学習法を考案した。また、人工データ及び多数の現実データを用いた実験により、その有効性を検証した。

1章では、本研究の背景及び目的について述べた。また、本論文の構成について説明した。

2章では、量的変数のみを用いる多変量多項式回帰を実現する方法として提案されたRF5法について説明した。RF5法は、3層パーセプトロンの学習によってデータにフィットする多変量多項式を推定する。組合せ探索によって多変量多項式を推定する方法では、変量が多い場合には組合せ爆発に陥るなどの問題があったが、RF5法では数値的アプローチを採用することにより、組合せ爆発に陥ることなく多変量多項式を推定する。パーセプトロンの学習には、準ニュートン法を基本枠組みとして、最適ステップ幅を2次近似の最小点として計算するBPQ法を用いる。本章では、BPQ法の基となる準ニュートン法の基本アルゴリズムについて説明した。RF5法では多項式の定数項以外の項数が3層パーセプトロンの中間ユニット数に相当することを示し、モデル選択の方法として交差検証法またはBICを用いて、最適中間ユニット数を選択する方法を説明した。また、汎化性能を向上させるための方法の1つとして、重みの絶対値をできるだけ小さくする制約を加えて学習を行う、正則化法についても説明した。

3章では、量的変数と質的変数の両方を用いる質的条件付き多変量多項式回帰を実現する方法として提案されたRF6.3法と、RF6.3法の問題点を克服するために本研究で提案したRF6.4法について説明した。質的条件付き多変量多項式回帰では、質的変数を用いて自動的かつ適切に部分空間を構成し、各部分空間の中ではそれぞれにフィットする多変量多項式を量的変数を用いて推定する。RF6.3法で

は，3層パーセプトロンの学習とベクトル量子化によるルール復元によってそれを実現する．そのための条件部の数値表現の方法や，ルール復元の方法の詳細について説明した．RF6.3法ではパーセプトロンの質的変数部分の構造が单層であるために複雑な条件部を表現できないことを示し，4層パーセプトロンを用いてより柔軟に条件部を表現できるように改良したRF6.4法を提案した．本章では，その詳細について説明した．また，BICを用いた最適ルール数の決定方法についても説明した．

4章では，量的変数と質的変数の両方を用いる質的条件付き重回帰を実現する方法として本研究で提案したnc-MR法について説明した．nc-MR法では，RF6.4法と同様に質的変数を用いて自動的かつ適切に部分空間を構成し，各部分空間の中ではそれぞれにフィットする線形回帰式を量的変数を用いて推定する．質的条件付き多变量多項式回帰では，可読性よりも汎化性を重視した回帰ルール集合を得ることを目的としているが，質的条件付き重回帰では，各ルールにおける回帰式を線形にすることによって，汎化性よりも可読性を重視した回帰ルール集合を得ることを目的とする．nc-MR法の基本的な枠組みについてはRF6.4法とほぼ同じであるが，ベクトル量子化によるルール復元において，係数ベクトルのサイズがRF6.4法より大きくなることが多いので，それを高速に処理をするための工夫が成されている．その工夫について説明した．また，各回帰ルールの近さを視覚的に捕らえるために，spring modelを用いて回帰ルールの位置関係をグラフとして表示する方法についても説明した．

5章では，多層パーセプトロンの重み共有法について述べ，それを実現する学習法として提案されたBCW法，及び本研究でBCW法に加えた改良について説明した．重み共有法では，パラメータの取りうる値の種類を限定することにより，回帰式の可読性を向上させると共に，ノイズへのオーバーフィットを減らして汎化性能を向上させることを目的とする．また，0に近い共有重みに重み枝刈りを適用することにより，不要変数に対するパラメータを一挙に削除することが可能となる．共有重みの値の決定，及び各重みをどのクラスタに所属させるかを決める重みクラスタリング問題には，多くの局所解が存在すると思われる所以，クラスタ数を増減しながらボトムアップとトップダウンの双方向にクラスタリングを行うBCW法によって，大域最適な解を探索する．本章では，BCW法による双方向のクラスタリングの方法について説明し，ボトムアップ及びトップダウンクラスタリングにおける，誤差の増加量または減少量に基づくクラスタ併合・分割基準の導出についても説明した．また，BCW法の初期クラスタ決定の方法及びクラス

タ分割の方法に改良を加え，RF5法またはRF6.4法に適用した場合の処理手順について説明した。

6章では，人工データに対してRF5法，RF6.4法，nc-MR法，及び重み共有法を適用したRF5+BCW法，RF6.4+BCW法を適用する実験を行い，提案モデルが元の回帰式及び回帰ルール集合を復元できるかどうかを検証した。特にモデル選択の方法として，何回も繰り返して実験を行った場合のBICが最小になった回数，BICの最小値，及び交差検証誤差を用いて，正しいパラメータが得られるかどうかを検証した。データに付加するノイズの大きさを変えて実験を行ったが，データに付加したノイズが大きい場合には，BICの最小値で最適なパラメータを決定すると正解よりも大きい値を選択してしまう場合が僅かにあった。しかし，ほとんどの場合には正しいモデルを選択することができていた。また，BCW法による重み共有によって可読性を向上させるだけでなく，汎化性能も僅かに向上させることができた。そして，元の回帰式及び回帰ルール集合とほぼ等価な式を復元することができた。

7章では，様々な現実データに対して提案モデルを適用する実験を行い，現実データに対しても提案モデルが有効に働くかどうかを検証した。まず，人工データを用いた実験と現実データを用いた実験の違いについて説明した。ダイヤmondの価格データを用いた実験では，データに説明変数が1つしかないため，実際に回帰式のグラフを描画してそれぞれの回帰モデルがどのようにデータにフィットしているのかを，視覚的に確認した。提案モデルは通常は多変量のデータに適用することを想定しているが，单変量のデータに対しても問題なく適用できる。ただし，BCW法による重み共有はあまり意味がないので実験は行わなかった。実験により，ダイヤmondデータに対しては特にRF6.4法が有効であることを確認できた。野球データを用いた実験では，このデータに対しては特にnc-MR法が可読性，汎化性の両面で優れていることを確認できた。また，BCW法による重み共有を適用することによって，重み共有を行わない場合と比較して可読性の良い結果を得ることができていた。更に，汎化性能も僅かに向上させることができていた。他の現実データを用いた性能比較では，データを6つのタイプに分類して検証を行った。そして，提案モデルが有効に働くタイプのデータは多く存在することが確認できた。計算時間の比較では，提案モデルは多層パーセプトロンを用いているので他のモデルに比べて決して早いとはいえないが，回帰分析では通常リアルタイム処理は必要としないことを考えると，十分に実用的な計算時間で処理ができるることを確認した。

以上の実験から、提案モデルによって可読性と汎化性の両方に優れた回帰分析を行うことができる事を示した。データそのものの構造によっては、提案モデルより重回帰や数量化などの既存の線形回帰モデルで十分な場合もあるが、多くの場合に、提案モデルは既存の線形回帰モデルに比べて良い汎化性能を得ることが期待できる。

本研究に関する今後の課題としては、もっと様々なタイプのデータに対して適用できるようにするということが挙げられる。本研究における提案モデルは、説明変数が量的変数と質的変数、目的変数が量的変数であるデータにのみ適用可能であり、例えば時系列性を持つデータや、質的変数に順序がある場合などには、そのまま提案モデルに適用するのではなく、データのそのような特徴をうまく生かす方法を考えるべきで、これは今後の重要な課題の1つであるといえる。また、本研究の提案モデルは、小規模～中規模のデータへの適用を前提としており、大規模データに対しては計算量などの面からまだまだ課題が残る。特に多層パーセプトロンの学習では、他の機械学習の方法に比べると莫大な計算量が必要となる。本研究では計算量の削減についてはあまり考慮していないので、それについては今後の大きな課題であるといえる。

謝辞

本研究を進めるにあたり、多くの方々のご支援とご協力をいただきました。ここに感謝の意を表すとともに、お礼申し上げます。

名古屋工業大学大学院工学研究科情報工学専攻教授 犬塚信博先生には、本論文の主査として多大なご尽力をいただき、多数の有益なご意見、ご指摘を賜りました。心からお礼申し上げます。

名古屋工業大学大学院工学研究科情報工学専攻教授 内匠逸先生、同 世木博久先生には、本論文の副査としてご尽力をいただき、貴重なご意見を賜りました。厚くお礼申し上げます。

中部大学工学部情報工学科教授 中野良平先生には、本研究の機会を与えていただき、研究の進め方から論文の書き方に至るまで、本論文の執筆に関する全ての分野を懇切丁寧にご指導いただきました。また、中部大学へご移籍された後も、ご多忙の中、本論文の学外審査員として細部までご指導いただきました。本研究を成し遂げることができたのは、ひとえに先生のご指導とご鞭撻の賜物です。心から感謝とお礼申し上げます。

名古屋工業大学大学院工学研究科創成シミュレーション工学専攻准教授 竹内一郎先生には、今年一年間、研究についてはもちろんのこと、研究室での生活や学位申請に関しても多大なご支援をいただきました。また、勉強会などを通して、関連分野の幅広い知識をご教授いただき、本論文の細部までご指導いただきました。更に、卒業後の進路についても親身になってご相談に乗っていただきました。心から感謝いたします。

東京工業高等専門学校情報工学科講師 北越大輔先生には、公私に渡って多大なご支援をいただきました。特に、博士後期課程への進学やその後の進路についてなどのご相談に乗っていただき、数々のご助言をいただきました。心から感謝いたします。

静岡県立大学経営情報学部教授 斎藤和巳先生には、共同研究者として本研究についての数々のご指導とご助言をいただきました。厚くお礼申し上げます。

名古屋工業大学 中野・竹内研究室所属の皆様には、研究についてのみならず、研究室での日々の生活においても多大なご協力をいただきました。特に、博士後期課程の石川勇太氏と鳥山昌幸氏には、同じ博士後期課程の学生として、様々なご相談に乗っていただきました。厚くお礼申し上げます。

最後に、今まで暖かく見守り激励してくださった友人、知人の皆様、そして長きにわたる学生生活を支えてくださった家族に深く感謝いたします。

2008年12月

棚橋 裕輔

参考文献

- [1] P. McCullagh, and J. Nelder, Generalized Linear Models, CRC Press, 1989.
- [2] C. M. Bishop, Pattern recognition and machine learning, Springer, 2006.
- [3] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, “Classification and Regression Trees,” CRC Press, 1984.
- [4] A. Karalic, “Employing linear regression in regression tree leaves,” Proc. 10th European Conf. on Artificial Intelligence, pp.440-441, 1992.
- [5] T. Hastie, and R. Tibshirani, Generalized Additive Models, CRC Press, 1990.
- [6] G. Shakhnarovich, T. Darrell, and P. Indyk, Nearest-Neighbor Methods in Learning and Vision: Theory and Practice, MIT Press, 2006.
- [7] M. Minsky, and S. Papert, “Perceptrons,” MIT Press, 1969.
- [8] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” Nature, Vol.323-9, pp.533-536, 1986.
- [9] C. M. Bishop, Neural networks for pattern recognition, Clarendon Press, 1995.
- [10] 中野良平, ニューラル情報処理の基礎数理, 数理工学社, 2005 .
- [11] P. Langley, “Bacon.1: a general discovery system,” Proc. 2nd Biennial Conf. of the Canadian Society for Computational Studies of Intelligence, pp.173-180, 1978.
- [12] 斎藤和巳, 中野良平, “コネクショニストアプローチによる数法則の発見,”情処学論, Vol.37, No.9, pp.1708-1716 , 1996 .
- [13] R. Nakano and K. Saito, “Discovering polynomials to fit multivariate data having numeric and nominal variables,” Progress in Discovery Science, pp.482-493, 2002.
- [14] 斎藤和巳, 中野良平, “重み双方向クラスタリング学習法: 最適な共有重み構造のニューラルネットを求めて,”信学論(D-II), Vol.J88-D-II, No.4, pp.789-799 , 2005.
- [15] 小西貞則, 北川源四郎, 情報量基準, 朝倉出版, 2004 .
- [16] 高橋大輔, 数値計算, 岩波書店, 1996 .
- [17] C. Schaffer, “Bivariate scientific function finding in a sampled, real-data testbed,” Machine Learning, Vol.12, No.1/2/3, pp.167-183, 1993.
- [18] K. Saito and R. Nakano, “Partial BFGS update and efficient step-length calculation for three-layer neural networks,” Neural Computation, 9(1), pp.239-257, 1997.

- [19] 金谷健一 , これなら分かる最適化数学 , 共立出版 , 2005 .
- [20] 斎藤和巳 , 中野良平 , “ 交差検証誤差最小化による重要重みの発見 , ” 信学論 (D-II) , Vol.J84-D-II , No.1 , pp.178-187 , 2001 .
- [21] 永田崇 , 川田敦史 , 山田憲一 , 中野良平 , “MV 正則化を用いたニューラルネット簡素化 , ” 信学技報 , NC2001-121 , pp.117-122 , 2002.
- [22] B. C. Falkenhainer and R. S. Michalski, “Integrating quantitative and qualitative discovery in the abacus system,” Machine Learning: An Artificial Intelligence Approach, vol.3, pp.153-190, 1990.
- [23] B. Nordhausen and P. Langley, “A robust approach to numeric discovery,” Proc. 7th Int. Conf. on Machine Learning, pp.411-418, 1990.
- [24] R. Nakano and K. Saito, “Discovery of a set of nominally conditioned polynomials,” Proc. 2nd Int. Conf. on Discovery Science, LNAI 1721, pp.287-298, 1999.
- [25] J. R. Quinlan, C4.5: programs for machine learning, Morgan Kaufmann, 1993.
- [26] D. Pelleg and A. Moore, “X-menas: Extracting K-menas with efficient estimation of the number of clusters,” Proc. 17th Int. Conf. on Machine Learning, pp.727-734, 2000.
- [27] R.A. Jacobs, M.I. Jordan and A.G. Barto, “Task Decomposition Through Competition in a Modular Connectionist Architecture: The What and Where Vision Tasks,” Cogn. Sci., vol.15, pp.219-250, 1991.
- [28] 浦浜喜一 , 毛井雄一郎 , “ 割り当てを伴うニューラルネット学習 , ” 信学論 (D-II) , Vol.J78-D-II , No.1 , pp.124-130 , 1995.
- [29] 岡田正之 , 浦浜喜一 , “ クラスタ分割関数回帰のこう配学習法 , ” 信学論 (D-II) , Vol.J79-D-II , No.11 , pp.1913-1919 , 1996.
- [30] R.A. Jacobs, M.I. Jordan, S.J. Nowlan and G.E. Hinton, “Adaptive mixtures of local experts,” Neural Computation, vol.3, no.1, pp.79-87, 1991.
- [31] M.I. Jordan and R.A. Jacobs, “Hierarchical mixtures of experts and EM algorithm,” Neural Computation, vol.6, no.2, pp.181-214, 1994.
- [32] T. Kamada and S. Kawai, “An algorithm for drawing general undirected graphs,” Information Processing Letters, vol.31, pp.7-15, 1989.
- [33] T. M. J. Fruchterman and E.M. Reingold, “Graph drawing by force-directed placement,” Software-practice and Experience, vol.21 , pp.1129-1164, 1991.
- [34] Y. LeCun, J. S. Denker and S. A. Solla, “Optimal brain damage,” Advances in Neural Information Processing Systems, vol.2, pp.1598-605, 1990.
- [35] S. J. Nowlan and G. E. Hinton, “Simplifying neural net-works by soft weight sharing,” Neural Computation, vol.4, no.4, pp.473-493, 1992.
- [36] F. Koksal, E. Alpaydin and G. Dundar, “Weight quantiza-tion for multi-layer perceptrons using soft weight sharing,” Proc. Int. Conf. on Artificial Neural Networks, pp.211-216, 2001.

- [37] K. J. Binkley, K. Seehart and M. Hagiwara, "A study of artificial neural network architectures for othello evaluation functions," Trans. of the Japanese Society for Artificial Intelligence 22(5), pp.461-471, 2007.
- [38] S. Chu, "Pricing the C's of diamond stones," Journal of Statistics Education, Vol.9, No.2, 2001.
- [39] 斎藤和巳, 中野良平, "HMEの構成的学習アルゴリズム," 信学論 (D-II) , Vol.J81-D-II , No.2 , pp.404-411 , 1998.

研究業績

学会誌掲載論文（査読有り）

- [1] 棚橋裕輔，中野良平，斎藤和巳，“4層パーセプトロンを用いた質的条件付き重回帰，”電子情報通信学会論文誌，Vol.J91-D-II，No.8，pp.2166-2176，2008。
- [2] **Yusuke Tanahashi**, and Ryohei Nakano, “Bidirectional Clustering of Weights for Finding Succinct Multivariate Polynomials,” International Journal of Computer Science and Network Security, Vol.8, No.5, pp.85-94, 2008.

国際学会発表論文（査読有り）

- [3] **Yusuke Tanahashi**, Kazumi Saito, and Ryohei Nakano, “Piecewise Multivariate Polynomials using a Four-Layer Perceptron,” Proceeding of the 8th International Conference on Knowledge-based Intelligent Information & Engineering Systems (KES '04), pp.602-608, 2004.
- [4] **Yusuke Tanahashi**, Xiang-Fang Chin, Kazumi Saito, and Ryohei Nakano, “Finding a Succinct Multi-layer Perceptron Having Shared Weights,” Proceeding of the 2005 International Joint Conference on Neural Networks (IJCNN '05), pp.1418-1423, 2005.
- [5] **Yusuke Tanahashi**, Kazumi Saito, and Ryohei Nakano, “Model Selection and Weight Sharing of Multi-layer Perceptrons,” Proceeding of the 9th International Conference on Knowledge-based Intelligent Information & Engineering Systems (KES '05), pp.716-722, 2005.
- [6] **Yusuke Tanahashi**, Kazumi Saito, Daisuke Kitakoshi, and Ryohei Nakano, “Finding Nominally Conditioned Multivariate Polynomials using a Four-layer Perceptron Having Shared Weights,” Proceeding of the 10th International Conference on Knowledge-based Intelligent Information & Engineering Systems (KES '06), pp.969-976, 2006.
- [7] **Yusuke Tanahashi**, Daisuke Kitakoshi, and Ryohei Nakano, “Nominally Piecewise Multiple Regression using a Four-layer Perceptron,” Proceeding of the 11th International Conference on Knowledge-based Intelligent Information & Engineering Systems (KES '07), pp.218-226, 2007.
- [8] Ying Yan, **Yusuke Tanahashi**, and Ryohei Nakano, “A set of linear regressions with automatic nominal space partition using a four-layer perceptron,” Proceeding of International Conference on Artificial Intelligence and Applications (ICAIA'07) in the IMECS2007, pp.53-58, 2007.

国内学会発表論文

- [9] 棚橋裕輔, 伊藤貴博, 斎藤和巳, 中野良平, “多層パーセプトロンを用いた区分的多项式回帰,”電子情報通信学会技术研究报告, Vol.103, No.490, pp.7-11, 2003.
- [10] 棚橋裕輔, 北越大輔, 中野良平, “4層パーセプトロンを用いた区分的線形回帰,”日本神經回路学会第16回全国大会講演論文集, JNNS2006-P1-25, pp.60-61, 2006.
- [11] 棚橋裕輔, 燕英, 中野良平, “4層パーセプトロンを用いた名義空間自動分割型重回帰,”電子情報通信学会技术研究报告, Vol.106, No.407, pp.67-72, 2006.

付録

A. BPQ 法における最適探索幅の計算方法

準ニュートン法の効率は step3 で採用する最適探索幅 α_k の計算方法に大きく依存する。以下では、効率の良い α_k の計算方法を導出する。なお、以降では簡単のため添え字 k を省略し、 $E(x; \theta)$, $f(x; \theta)$ をそれぞれ E , f と表記する。特に、 $f(x^\mu; \theta)$ を f^μ と表記する。

step3 では、変数は α だけなので、 $E(x; \theta + \alpha \Delta \theta)$ を単に $g(\alpha)$ で表す。このとき、 $g(\alpha)$ の 2 次近似式は以下となる。

$$g(\alpha) \approx g(0) + g'(0)\alpha + \frac{1}{2}g''(0)\alpha^2 \quad (\text{A.1})$$

ここで、 $g'(0)$ 及び $g''(0)$ の正負によって以下のように場合を分けて考える。

(i) $g'(0) > 0$ のとき

$g'(0) > 0$ のとき、その探索方向で目的関数の値を減少できないので、 $\Delta \theta = -\nabla E$, $H = I$ に設定する。それにより、 $g'(0) < 0$ が保証される。

(ii) $g'(0) < 0$ かつ $g''(0) > 0$ のとき

式 (A.1) の右辺の最小値は以下で与えられる。

$$\alpha = -\frac{g'(0)}{g''(0)} \quad (\text{A.2})$$

(iii) $g'(0) < 0$ かつ $g''(0) \leq 0$ のとき

式 (A.1) の値は負または無限大となるので、Gauss-Newton 法を用いてこの場合に対処する。 $f(x; \theta + \alpha \Delta \theta)$ の 1 次近似は $f^\mu + f'^\mu \alpha$ となり、 $g(\alpha)$ の近似は以下となる。

$$g(\alpha) \approx g(0) + g'(0)\alpha + \frac{1}{2} \sum_{\mu=1}^N (f'^\mu)^2 \alpha^2 \quad (\text{A.3})$$

よって、式 (A.3) の右辺の最小値は以下で与えられる。

$$\alpha = -\frac{g'(0)}{\sum_{\mu=1}^N (f'^\mu)^2} \quad (\text{A.4})$$

明らかに $g'(0) < 0$ のとき、式 (A.4) の値は正となる。この場合、探索は鞍点の近くである可能性があるので、 $k = M$ とし、 \mathbf{H} を単位行列に戻す。

以上の3つの場合分けによって、最適探索幅を計算することができる。しかし多くの場合、各反復で θ の修正量の上限を設定することは有効であるといえる。よって、 $||\alpha \Delta \theta|| > 1.0$ ならば、 α を $1.0/||\Delta \theta||$ とする。この場合、探索は鞍点の近くである可能性があるので、 $k = M$ とし、 \mathbf{H} を単位行列に戻す。

α は近似に基づき計算されるので、目的関数 $g(\alpha)$ の値が常に減少するとは限らない。よって、 $g(\alpha) \geq g(0)$ のときは、 α を縮める処理を導入する。 $\tilde{\alpha}$ を式 (A.2) または (A.4) で計算した値とすれば、条件 $h(0) = g(0)$, $h(\tilde{\alpha}) = g(\tilde{\alpha})$, $h'(0) = g'(0)$ を満たす2次近似式 $h(\alpha)$ は以下となる。

$$g(\alpha) \approx h(\alpha) = g(0) + g'(0)\alpha + \frac{g(\tilde{\alpha}) - g(0) - g'(0)\tilde{\alpha}}{\tilde{\alpha}^2}\alpha^2 \quad (\text{A.5})$$

$g(\tilde{\alpha}) > g(0)$ かつ $g'(0) < 0$ より、 $h(\alpha)$ の最小点は以下で与えられる。

$$\alpha = -\frac{g'(0)\tilde{\alpha}^2}{2(g(\tilde{\alpha}) - g(0) - g'(0)\tilde{\alpha})} \quad (\text{A.6})$$

ここで、式 (A.6) では $0 < \alpha < \tilde{\alpha}$ が保証される。よって、この処理を $g(\alpha) < g(0)$ となるまで繰り返せば、 $g(\alpha) < g(0)$ となる α をつねに求めることができる。

B. RF5法における微分値の計算

RF5法において、BPQ法での学習時に必要となる勾配 ∇E の計算式を以下に示す。

$$\nabla E = \frac{\partial E}{\partial \theta} = \sum_{\mu=1}^N (f^\mu - y^\mu) \frac{\partial f^\mu}{\partial \theta} \quad (\text{B.1})$$

$$\frac{\partial f^\mu}{\partial w_0} = 1 \quad (\text{B.2})$$

$$\frac{\partial f^\mu}{\partial w_j} = m_j \quad (\text{B.3})$$

$$\frac{\partial f^\mu}{\partial w_{jk}} = w_j m_j \ln x_k \quad (\text{B.4})$$

BPQ 法の最適探索幅計算で必要となる $g'(0)$, $g''(0)$ の計算式を以下に示す .

$$g'(0) = \sum_{\mu=1}^N (f^\mu - y^\mu) f'^\mu \quad (\text{B.5})$$

$$g''(0) = \sum_{\mu=1}^N ((f'^\mu)^2 + (f^\mu - y^\mu) f''^\mu) \quad (\text{B.6})$$

$$f'^\mu = \Delta w_0 + \sum_{j=1}^J \left(\Delta w_j m_j^\mu + w_j m_j'^\mu \right) \quad (\text{B.7})$$

$$f''^\mu = \sum_{j=1}^J \left(2 \Delta w_j m_j'^\mu + w_j m_j''^\mu \right) \quad (\text{B.8})$$

$$m_j''^\mu = m_j^\mu \sum_{k=1}^K \Delta w_{jk} \ln x_k^\mu \quad (\text{B.9})$$

$$m_j''^\mu = m_j^\mu \left(\sum_{k=1}^K \Delta w_{jk} \ln x_k^\mu \right)^2 \quad (\text{B.10})$$

ただし , Δw_j , Δw_{jk} はそれぞれ step2 で計算される w_j , w_{jk} の変化量である .

C. RF6.4 法における微分値の計算

RF6.4 法において , BPQ 法での学習時に必要となる勾配 ∇E の計算式を以下に示す .

$$\nabla E = \frac{\partial E}{\partial \Theta} = \sum_{\mu=1}^N (f^\mu - y^\mu) \frac{\partial f^\mu}{\partial \Theta} \quad (\text{C.1})$$

$$\frac{\partial f^\mu}{\partial w_{jk}} = c_j^\mu m_j^\mu x_k^\mu \quad (\text{C.2})$$

$$\frac{\partial f^\mu}{\partial v_{jr}} = \sigma_r^\mu m_j^\mu \quad (\text{C.3})$$

$$\frac{\partial f^\mu}{\partial v_{rkl}} = \sigma_r^\mu (1 - \sigma_r^\mu) q_k^\mu \left(\sum_{j=0}^J v_{jr} m_j^\mu \right) \quad (\text{C.4})$$

ただし , $m_0^\mu = 1$ とする .

BPQ 法の最適探索幅計算で必要となる $g'(0)$, $g''(0)$ の計算式を以下に示す .

$$g'(0) = \sum_{\mu=1}^N (f^\mu - y^\mu) f'^\mu \quad (\text{C.5})$$

$$g''(0) = \sum_{\mu=1}^N ((f'^\mu)^2 + (f^\mu - y^\mu) f''^\mu) \quad (\text{C.6})$$

$$f'^\mu = \sum_{j=0}^J (c_j'^\mu m_j^\mu + c_j^\mu m_j'^\mu) \quad (\text{C.7})$$

$$f''^\mu = \sum_{j=0}^J (c_j''^\mu m_j^\mu + 2c_j'^\mu m_j'^\mu + c_j m_j''^\mu) \quad (\text{C.8})$$

$$c'^\mu = \sum_{r=1}^R (\Delta v_{jr} \sigma_r^\mu + v_{jr} \sigma_r'^\mu) \quad (\text{C.9})$$

$$c''^\mu = \sum_{r=1}^R (2\Delta v_{jr} \sigma_r'^\mu + v_{jr} \sigma_r''^\mu) \quad (\text{C.10})$$

$$\sigma_r'^\mu = \sigma_r^\mu (1 - \sigma_r^\mu) \left(\sum_{k=1}^{K_1} \sum_{l=1}^{L_k} \Delta v_{rkl} q_k^\mu \right) \quad (\text{C.11})$$

$$\sigma_r''^\mu = \sigma_r^\mu (1 - \sigma_r^\mu) (1 - 2\sigma_r^\mu) \left(\sum_{k=1}^{K_1} \sum_{l=1}^{L_k} \Delta v_{rkl} q_k^\mu \right)^2 \quad (\text{C.12})$$

$$m_j'^\mu = m_j^\mu \sum_{k=1}^{K_2} \Delta w_{jk} \ln x_k^\mu \quad (\text{C.13})$$

$$m_j''^\mu = m_j^\mu \left(\sum_{k=1}^{K_2} \Delta w_{jk} \ln x_k^\mu \right)^2 \quad (\text{C.14})$$

ただし , Δw_{jk} , Δv_{jr} , Δv_{rkl} はそれぞれ step2 で計算される w_{jk} , v_{jr} , v_{rkl} の変化量である .
また , $m_0'^\mu = m_0''^\mu = 0$ とする .

D. nc-MR 法における微分値の計算

nc-MR 法において , BPQ 法での学習時に必要となる勾配 ∇E の計算式を以下に示す .

$$\nabla E = \frac{\partial E}{\partial \theta} = \sum_{\mu=1}^N (f^\mu - y^\mu) \frac{\partial f^\mu}{\partial \theta} \quad (\text{D.1})$$

$$\frac{\partial f^\mu}{\partial v_{jr}} = \sigma_r^\mu x_j^\mu \quad (\text{D.2})$$

$$\frac{\partial f^\mu}{\partial v_{r0}} = \sigma_r^\mu (1 - \sigma_r^\mu) \left(\sum_{j=0}^J v_{jr} x_j^\mu \right) \quad (\text{D.3})$$

$$\frac{\partial f^\mu}{\partial v_{rkl}} = \sigma_r^\mu (1 - \sigma_r^\mu) q_k^\mu \left(\sum_{j=0}^J v_{jr} x_j^\mu \right) \quad (\text{D.4})$$

BPQ 法の最適探索幅計算で必要となる $g'(0)$, $g''(0)$ の計算式を以下に示す .

$$g'(0) = \sum_{\mu=1}^N (f^\mu - y^\mu) f'^\mu \quad (\text{D.5})$$

$$g''(0) = \sum_{\mu=1}^N ((f'^\mu)^2 + (f^\mu - y^\mu) f''^\mu) \quad (\text{D.6})$$

$$f'^\mu = \sum_{j=0}^J c'_j{}^\mu x_j^\mu \quad (\text{D.7})$$

$$f''^\mu = \sum_{j=0}^J c''_j{}^\mu x_j^\mu \quad (\text{D.8})$$

$$c'^\mu = \sum_{r=1}^R (\Delta v_{jr} \sigma_r^\mu + v_{jr} \sigma'_r{}^\mu) \quad (\text{D.9})$$

$$c''^\mu = \sum_{r=1}^R (2\Delta v_{jr} \sigma'_r{}^\mu + v_{jr} \sigma''_r{}^\mu) \quad (\text{D.10})$$

$$\sigma'_r{}^\mu = \sigma_r^\mu (1 - \sigma_r^\mu) \left(\Delta v_{r0} + \sum_{k=1}^{K_1} \sum_{l=1}^{L_k-1} \Delta v_{rkl} q_k^\mu \right) \quad (\text{D.11})$$

$$\sigma''_r{}^\mu = \sigma_r^\mu (1 - \sigma_r^\mu) (1 - 2\sigma_r^\mu) \left(\Delta v_{r0} + \sum_{k=1}^{K_1} \sum_{l=1}^{L_k-1} \Delta v_{rkl} q_k^\mu \right)^2 \quad (\text{D.12})$$

ただし , Δv_{jr} , Δv_{r0} , Δv_{rkl} はそれぞれ step2 で計算される v_{jr} , v_{r0} , v_{rkl} の変化量である .

E. Spring Model における微分値の計算

式 (4.18) 及び式 (4.20) の計算で必要となる , E_{spr} の x_m, y_m での微分値を以下に示す .

$$\frac{\partial E_{spr}}{\partial x_m} = \sum_{i \neq m} k_{mi} \left\{ (x_m - x_i) - \frac{l_{mi}(x_m - x_i)}{\{(x_m - x_i)^2 + (y_m - y_i)^2\}^{1/2}} \right\} \quad (\text{E.1})$$

$$\frac{\partial E_{spr}}{\partial y_m} = \sum_{i \neq m} k_{mi} \left\{ (y_m - y_i) - \frac{l_{mi}(y_m - y_i)}{\{(x_m - x_i)^2 + (y_m - y_i)^2\}^{1/2}} \right\} \quad (\text{E.2})$$

$$\frac{\partial^2 E_{spr}}{\partial x_m^2} = \sum_{i \neq m} k_{mi} \left\{ 1 - \frac{l_{mi}(y_m - y_i)^2}{\{(x_m - x_i)^2 + (y_m - y_i)^2\}^{3/2}} \right\} \quad (\text{E.3})$$

$$\frac{\partial^2 E_{spr}}{\partial x_m \partial y_m} = \sum_{i \neq m} k_{mi} \frac{l_{mi}(x_m - x_i)(y_m - y_i)}{\{(x_m - x_i)^2 + (y_m - y_i)^2\}^{3/2}} \quad (\text{E.4})$$

$$\frac{\partial^2 E_{spr}}{\partial y_m^2} = \sum_{i \neq m} k_{mi} \left\{ 1 - \frac{l_{mi}(x_m - x_i)^2}{\{(x_m - x_i)^2 + (y_m - y_i)^2\}^{3/2}} \right\} \quad (\text{E.5})$$

