統計利用 BP 属性データ分類*

1J－7

北島伸克†
NEC ヒューマンメディア研究所‡
e-mail: kitajima@hml.cl.nec.co.jp

1　はじめに

昨今期待が高まっているデータマイニングの重要な要素技術に顧客データ等属性データの分類がある．従来より，属性データ分類は線形多変量解析，決定木 [2] およびニューラルネット [3] 等で行われてきた．この中で，線形多変量解析による分類は，最も広く実用的に用いられているものの，複雑な分布を持つ属性データを扱うためには性能的限界が明らかである．その限界を超えるためには，非線形的に拡張したデータ分類を行う必要がある．本研究では，誤差逆伝播 (Back Propagation, BP)[4] 法を非線形分類手法として用いて，性能向上を目指した．しかし，従来の BP データ分類では，対象データに対応した適切な構造決定方法が確立されていない点が問題であった．

本研究は，数量化 2 類による事前分析の結果を利用してニューラルネットの構造決定を行い，BP 分類手法の性能の向上を目指した．与信審査のデータを用いた実験で，全くの未知のデータの分類結果を比較した結果，推定の困難な不良データの分類性能に関して提案手法は C4.5[2] および数量化 2 類の性能を上回り，提案手法の効果を確認することが出来た．

2　中間層ユニット基準数

階層型ニューラルネット (図 1) の出力，例えば出力層 (第 3 層) 第 k ユニット ($k = 1, 2, \ldots, n$，ただし n は出力層ユニット数) の出力値 $output_{3,k}$ は，ユニットの活性化関数 f を用いて，

$$output_{3,k} = f(input_{3,k}) = \frac{1}{1 + e^{-\frac{input_{3,k} + \theta_k}{T}}} \quad (1)$$

となる (θ_k: 第 k ユニットの閾値, T: 温度)．また，出力層の第 k ユニットの入力値 $input_{3,k}$ は，次式で表される

$$input_{3,k} = \sum_{j=1,\ldots,m} w_{j,k} \times output_{2,j} \quad (2)$$

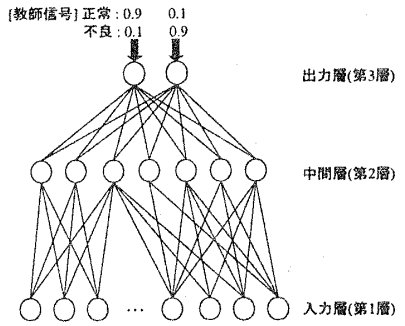

図 1: 本実験で用いた階層型ニューラルネット

(m:　中間層ユニット数，$w_{j,k}$:　第 j 中間層ユニット ($j = 1, 2, \ldots, m$) と第 k 出力層ユニットの間の結合の重み，$output_{2,j}$: 中間層 (第 2 層) の出力値)．つまり，階層型ニューラルネットの出力は，中間層ユニットの出力を出力層ユニットの活性化関数で非線形的に統合することによってデータ空間を表現している．この場合，中間層ユニットは分類結果に直接影響を与える要因を表していると考えることができる．よって，中間層ユニット数を影響度の大きい項目数に一致させれば，対象データ空間を十分モデリングできると考えた．よって本研究では，「分類結果に対する影響度が大きい属性項目数」を中間層ユニットの基準数とした．また，分類結果に対する影響度が小さい項目を入力すると汎化能力を落し，却って分類性能の低下につながると考えて上記属性項目のみをニューラルネットの入力項目とした．

3　入力項目と中間層ユニット基準数の決定

対象データに対して数量化 2 類を実施した結果得られる各項目のスコアのレンジは各項目の分類結果に対する影響度と考えることが出来る．入力属性項目を分類結果に対する影響度の大きい順番で選択して，選択した入力属性項目の影響度の和が全影響度の合計中の一定の割合を超えたところで入力属性項目数を決定した．2節で示した通り，その入力属性項目数を中間層ユニット基準数とした．

*A New BP Attributive Data Classification Method Using Multivariate Analysis

†Nobukatsu Kitajima
‡Human Media Research Laboratories, NEC Corporation

表 1: 与信審査用データの正常 / 不良個数の内訳

正常	不良	合計
6754	1656	8410

表 2: 未知データの正常 / 不良個数の内訳

正常	不良	全体
1877	456	2333

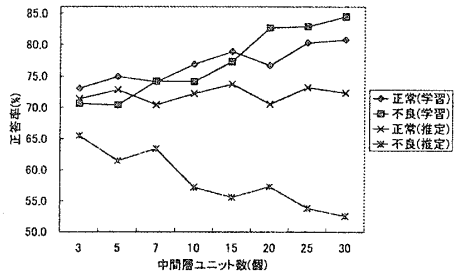

図 2: 中間層ユニット数に対する正答率

4 シミュレーション実験

本節では本構造決定手法の有効性を実証するために, 与信審査 (融資等を行う際に対象顧客を審査する処理) 用のデータを用いて行った 10-fold クロスバリデーション実験および未知データの推定結果を示す.

4.1 学習データ

本実験の入力項目は, 3節で示した方法で, 約 20 個の属性項目の中から 7 項目を選択した. また, 学習データセット中に存在する正常 (正常に取引中), 不良 (返済遅れ等の異常発生) データの内訳は表 1であり, 正常データは不良データの約 4 倍の個数存在する. 正常, 不良各データの個数差による学習への影響を避けるために, 正常および不良のデータを同数ずつ学習するように, 各クロスバリデーションセットの学習データ中の不良データのみ複数回ずつ学習した.

4.2 実験

4.1に記した通り, 3節で示した方法で本実験の中間層ユニットの基準個数は 7 個と決定した. ここで, 提案手法の有効性を検証するため, 3, 5, 7, 10, 15, 20, 25, 30 個の中間層ユニットで比較実験を行った (図 2). 図 2で推定データに注目すると, 中間層ユニット数の増加に伴って, 正常データの正答率 (実際に正常であったデータの中で正しく正常と推定できたデータの比率) はほぼ変動しないが, 不良データの正答率は中間層ユニット数が 10 個以上で大きく下がった. これは, 中間層ユニット数を提案手法で決定した 7 個よりも増加させるとモデルの汎化能力が低下していることを意味しており, 7 個が基準数として優れていると言える. 次に本手法と他の分類手法によるクロスバリデーションで利用していない未知データ (表 2) の分類結果の比較を表 3に示す. 提案手法は, C4.5 と比較して正常, 不良とも優れた正答率を示した. 数量化 2 類は 4.1節で示した学習データセット中の不良データを増加させた場合の実験が出来なかったため, 個数の多い正常データに偏った判定結果となっており, 全体の正答率が高くなっているが, 判別が難しい不良データの正答率を向上させるという目標は提案手法で達成できた.

表 3: 提案手法, C4.5, 数量化 2 類の未知データの推定正答率の比較 (* 学習データセットの正常, 不良比率がオリジナルのまま)

	正常 (%)	不良 (%)	全体 (%)
提案手法	75.8	56.1	71.6
C4.5	74.9	50.0	70.0
数量化 2 類 (*)	83.6	47.1	76.5

5 まとめ

BP 属性データ分類において, 数量化 2 類による事前分析を行うことによって, 対象データに適合した構造決定行う方法を提案した. 実験により, 提案した構造決定方法の有効性を確認した. 他手法と比較した結果, 提案手法によって, 推定が困難である不良データの正答率を向上させることが出来た.

参考文献

[1] Hayashi, C.: Ann. Inst. Statist. Math, Vol. 3, pp. 69-98, 1952.

[2] Quinlan, J. R. : C4.5, Morgan Kaufmann Publishers, Inc, 1993.

[3] Richeson, L.: International Journal of Applied Expert Systems, 2, no. 2, pp. 116-130, 1994.

[4] Rumelhart, D. E., et. al: PDP, The MIT Press, 1, pp. 318-362, 1986.