ロジスティック回帰を用いた手書き文字・記号の判別

冨田千尋　(指導教員：吉田裕亮)

1　はじめに

パターン認識とは, いくつかの概念に分類できる観測データが存在する時, 観測されたパターンをそれらの概念のうちのひとつに対応させることである.

一般的に識別器のひとつとしてニューラルネットワークが使用されている. ニューラルネットワークとは機械学習によりデータを識別するひとつの手法である.

高次元データのパターン認識は計算量が多く時間がかかる. そこで本研究では主成分分析 (PCA: Principal Component Analysis) を用いて, データを一旦低次元に縮約し, そのデータを最も単純なニューラルネットワークモデルのひとつであるロジスティック回帰によって機械学習を行うことで, 従来より短い時間での認識を実現させる.

2　PCA(主成分分析)

主成分分析 (以下 PCA) とは, 多次元データの情報を, その総合力や特性を保ちながらより低い次元に縮約させる方法である.

X を $p \times n$ のデータ行列とし, X の縦成分 (変数) ごとに平均と標準偏差を求めて標準化し, その行列を X_0 とする. そして, 相関行列 R を求める.

$$R = \frac{1}{p}(X_0)^t X_0$$

さらに, 固有方程式を解き, n 個の固有値と, 各々の固有値に対応する固有ベクトル V を求める. 固有値は元データの情報保持率を表す. 標準化されたデータ行列 X_0 を,

$$X_0 V = X^*$$

と変換する. X^* の第 1 列目は第 1 主成分, X^* の第 2 列目は第 2 主成分と呼ばれる.

3　ニューラルネットワーク (1 素子)

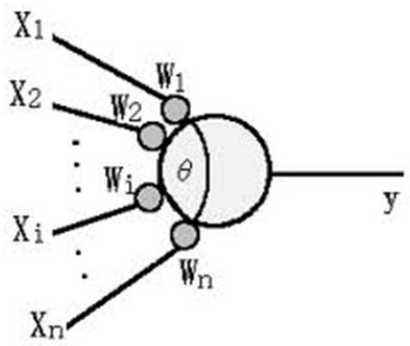

図 1: 単純パーセプトロンモデル

図 1 はニューラルネットワークの単純モデルのひとつである. i 番目の入力信号を $X_i(i = 1, 2, \ldots, n)$, それぞれの重みを $W_i(i = 1, 2, \ldots, n)$ とすると, 入力信号の総和は

$$\sum_{i=1}^{n} X_i W_i$$

となる. 入力信号を受け取ったニューロンは, その入力値が一定の閾値 θ を超えると他のニューロンに信号を出力する. したがって出力 y は

$$y = f(\sum_{i=1}^{n} X_i W_i - \theta)$$

となる. ここで関数 f は伝達関数といい, 2 値モデルと連続値モデルがある. 本研究では連続値モデル, シグモイド関数

$$f(u) = \frac{1}{1 + \exp(-u)}$$

を用いる.

4　ロジスティック回帰モデル

ある現象の発生する確率 p を, その現象の生起を説明するために観察された変数群 $x = (x_1, \ldots, x_r)$ で説明するとき,

$$p(x) = \Pr\{\text{ 発生 } | x_1, x_2, \ldots, x_r\} = \frac{1}{1 + \exp(-Z)}$$

で表すモデルをロジスティック回帰モデルという. ただし,

$$Z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r$$

で β_i は各変量 x_i の重みである. このモデルは他入力 1 出力のニューラルネットワークモデルに相当する.

5　提案手法

本研究では, ニューラルネットワークの 1 つであるロジスティック回帰モデルを用いて手書き文字・記号の判別を以下のように行うことを提案する.

1. 情報量の大きい多次元データを PCA により低次元に縮約する. 1 以上の固有値とそれに対応する固有ベクトルを判別データとして用いる.
2. 次元を縮約した PCA 成分を入力データとしてロジスティック回帰を行うことにより非線形の 2 群判別を行う.

6　数値実験

6.1 手書き数字データ

ここでは, 公開されているアメリカ合衆国郵便公社が業務で得た実際の手書き数字データを使用する. 数字 10 種類の手書きデータを各 100 個用意する. 画像は 16 × 16 画素の 256 次元にして使用する. 各画素値は 256 階調の離散値である. 各文字のデータ 100 個に適当に別の文字データを混合させた計 200 個のデータを作成する. PCA により 256 次元データを実験結果に基づきそれぞれ縮約する. 本研究では, 固有値が 1 以上である成分の個数を縮約次元とする. 数字 0〜9 における縮約次元は, 43, 39, 46, 46, 47, 48, 43, 43, 47, 45 であった. このように作成した PCA データをロジスティック回帰により, あるデータ n であるかそうでないかの 2 群に判別する. 学習に用いたデータを判別しようとしたときの正答率は, 以下の表のようになった.

n	正答率	n	正答率
0	0.985	5	0.975
1	0.980	6	0.995
2	0.965	7	0.985
3	0.985	8	0.930
4	0.985	9	0.976

表 1: 数字学習データ判別の正答率

どの学習データの判別においても高い正答率を得ることができた. 平均正答率は 0.976 であった.

次に, 学習に用いたデータとは異なるデータをテストデータとして 100 個ずつ用意し, 判別を行い汎化性能を調べた. 結果は以下の表となった.

n	正答率	n	正答率
0	0.970	5	0.920
1	0.950	6	0.970
2	0.900	7	0.950
3	0.930	8	0.900
4	0.970	9	0.930

表 2: 数字テストデータ判別の正答率

平均正答率は 0.939 であり, 学習データ判別の正答率と比較すると低い数値であるが, かなり高い判別率を得られた. ちなみに, 誤判別された画像データは以下のようなものであった.

左から 図 2: 2 を 2 と判別できなかった例
図 3: 5 を 5 と判別できなかった例
図 4: 9 を 9 と判別できなかった例

6.2 手書き記号データ

次に, 手書き数字データの実験と同様に, 手書き記号データの判別実験を行った. 本研究では音楽記号 (2 分音符, 4 分音符, 8 分音符, 16 分音符, 2 分休符, 4 分休符, 8 分休符, ト音記号, ヘ音記号) を用いた. これらの縮約次元は, 32, 35, 32, 34, 33, 37, 35, 38, 35 であった. 学習に用いたデータを判別しようとしたときの正答率, 及びテストデータを判別しようとしたときの正答率は以下の表のようになった.

n	学習	汎化性
2 分音符	0.967	0.920
4 分音符	0.967	0.950
8 分音符	0.956	0.930
16 分音符	0.961	0.930
2 分休符	0.989	0.970
4 分休符	0.989	0.960
8 分休符	0.994	0.980
ト音記号	1.000	0.970
ヘ音記号	0.994	0.930

表 3: 記号データ判別の正答率

平均正答率はそれぞれ 0.980, 0.949 であり, これもまた高い正答率を得ることができた. 誤判別された画像データは以下のようなものであった.

左から
図 5: 4 分音符を 4 分音符と判別できなかった例
図 6: 8 分音符を 8 分音符と判別できなかった例
図 7: 16 分音符を 16 分音符と判別できなかった例

7　まとめと今後の課題

PCA とロジスティック回帰を利用することで, 手書きの文字・記号の判別をすることができた. この手法は, 従来の誤差伝搬法を用いたニューラルネットワークによる判別法に比べ, かなり単純な計算法で判別を行うことができるため, 有効であると思われる. さらに, 学習に用いるデータ数を増やすことで判別における正答率を高められると考えられる. また, 今回判別を行った数字や音楽記号以外のデータにおいてもこの手法が有効であるか確認したい.

参考文献

1. 丹後俊郎, 山岡和枝, 高木晴良, ロジスティック回帰分析 SAS を利用した統計解析の実際, 朝倉書店 (1996)
2. 石村貞夫, すぐわかる多変量解析, 東京図書 (1992)