電子情報通信学会 パターン認識 ·メディア理解研究会 '98.6

多層パーセプトロンによる情報圧縮に適した二つのエネルギー関数

高橋隆史 †, 徳永隆治 ‡
† 筑波大学大学院博士課程工学研究科
〒 305-8573 筑波大学電子 ·情報工学系カオス研究室
e-mail: ume@chaos.is.tsukuba.ac.jp
‡ 筑波大学電子 ·情報工学系

二乗誤差をエネルギー関数とする従来の多層パーセプトロンの学習では、中間層素子が獲得するパラメータの寄与順が定まらないため、そのデータ表現は、主成分分析のような線形手法に比べて優れているとは言い難かった。本研究では、非線形パーセプトロンの恒等写像学習において、寄与順にならんだデータ表現の獲得を可能にする二つのエネルギー関数を検討する。これらのエネルギー関数は、特徴抽出とデータ圧縮という二種類の目的のそれぞれに適したデータ表現を構成し、主成分分析を上回る性能を得られることが示される。

多層パーセプトロン、特徴抽出、データ圧縮、主成分分析

Energy Functions for Efficient Nonlinear Dimensionality Reduction by Multi Layer Perceptrons

Takashi TAKAHASHI†, Ryuji TOKUNAGA‡
† Doctoral Program in Engineering, University of Tsukuba
Ibaraki 305–8573, Japan
e-mail: ume@chaos.is.tsukuba.ac.jp
‡ Institute of Information Sciences and Electronics, University of Tsukuba

This paper investigates two simple energy functions which are valid for two different purposes of dimensionality reduction: feature extraction and data compression. These energy functions enable nonlinear perceptrons to organize data representations whose parameters, namely, outputs of the bottleneck layer units, are arranged in the order of their importance. The efficacy of these energy functions is shown by numerical experiments in comparison with conventional squared error functions and Principal Component Analysis.

multi layer perceptron, feature extraction, data compression, principal component analysis

1　まえがき

情報圧縮あるいは次元圧縮は、広範な工学問題に適用される重要な手続きの一つである。その目的は、大まかに以下の二つに分類することができる。

- 特徴抽出 — 後に行う認識などの処理を容易にするため、または与えられたデータ集合の構造を解析するために、元のデータよりも低次元でその特徴を良く表すパラメータを取り出す。
- データ圧縮 — 圧縮効率を上げるため、個々のデータを最小限の数のパラメータで近似し、全体としてデータの記述に必要な平均パラメータ数を減らす。

どちらの場合も、情報圧縮の結果を利用する上での自由度を考えると、得られるデータ表現はパラメータが重要度の順に組織化されたものであることが望ましい。このようなデータ表現を構成することによって、例えば特徴抽出においては、必要に応じて重要度の高いパラメータのみを取り出すことが可能となる。また、データ圧縮においては、要求される圧縮レベルに応じてデータ量を調節する、あるいは上位のパラメータほど多くのビットを割り当てて効率的な量子化を実現する、といったことが可能となる。本論文では、このようなデータ表現を順序データ表現と呼ぶ。

順序データ表現を獲得できる手法の例としては、線形写像によって入力データの恒等写像を近似し、情報圧縮を実現する主成分分析(またはKarhunen-Loève(KL) 変換) があげられる。主成分分析の獲得するデータ表現においては、主成分と呼ばれるパラメータの恒等写像近似への寄与の大きさを分散の大きさによって評価し、それらの重要度の順を定めることができる。他方、多層パーセプトロンを用いて恒等写像学習すなわち入力データ自身を教師とする学習を行なうと、主成分分析と同様の情報圧縮を実現する事ができる。多層パーセプトロンの恒等写像近似能力については、3 層パーセプトロンによって主成分分析と等しい近似誤差が得られる事が証明されている [1, 2, 3]。また、非線形素子を含むより多層のパーセプトロンを用いると、同じ中間層素子数でより誤差の小さいデータ表現が得られる事も知られている [4, 5]。これらのパーセプトロンにおいては、各中間層素子の出力をデータ表現のパラメータと考えることができる。しかし、二乗誤差をエネルギー関数とする従来の誤差逆伝搬学習では順序データ表現が得られないため、多層パーセプトロンによる情報圧縮は有用であるとは言いがたい。そこで本論文では、特徴抽出とデータ圧縮という二つの目的のそれぞれに対して、主成分分析よりも効率の良い順序データ表現を多層パーセプトロンに自己組織させるためのエネルギー関数を検討する。

2　順序データ表現を自己組織するためのエネルギー関数

N次元のベクトルから成るデータ集合

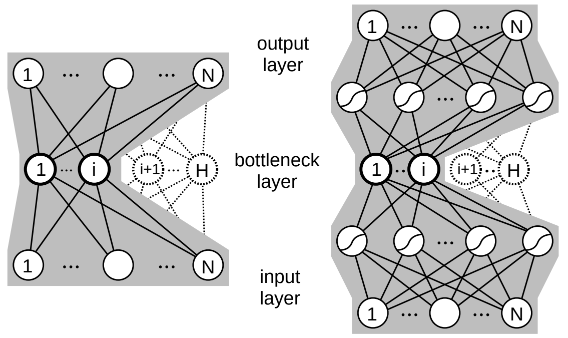
$$\left\{\boldsymbol{x}_t \in R^N\right\}_{t=1}^{T} \tag{1}$$

を変換して $H(< N)$ 個のパラメータで表現する情報圧縮を考える。主成分分析の場合、分散の大きなものから順に H個の主成分がこのパラメータとなる。多層パーセプトロンの場合、入力層および出力層の素子数を N、ボトルネックとなる中間層の素子数を Hとして恒等写像学習を行えば、同様の情報圧縮が実現できる。この場合、H個の中間層素子の出力がパラメータとなってデータ表現を構成する。ここで、両者のデータ表現の違いを明確にするため、H 個のパラメータのうち番号の小さな方から i 個だけを使って式 (1) のデータに対する恒等写像を近似する写像を考え、これを $f_i(i = 1, 2, \ldots, H)$ と表記する。主成分分析においては、f_iは上位 i 個の主成分による i 次元の恒等写像近似である。一方、多層パーセプトロンにおいては、中間層の第 $i + 1$ 素子以降の出力を 0 に置き換えたパーセプトロンが f_iに相当する。本論文では、この f_iを第 i 部分パーセプトロンと呼ぶ (図 1参照)。f_iをこのように定義すると、データ$\boldsymbol{x}_t$を f_iで近似したときの二乗誤差

$$e_{i,t} = ||\boldsymbol{x}_t - f_i(\boldsymbol{x}_t)||^2 \tag{2}$$

およびその平均

$$E_i = \frac{1}{T}\sum_{t=1}^{T} e_{i,t} = \frac{1}{T}\sum_{t=1}^{T} ||\boldsymbol{x}_t - f_i(\boldsymbol{x}_t)||^2 \tag{3}$$



three layer perceptron　five layer perceptron

図 1: 3 層及び 5 層のパーセプトロンとその部分パーセプトロン

を定めることができる。ここで、上述の多層パーセプトロンにおける誤差 E_H は、全ての素子を用いたときの誤差であり、最急降下法による誤差逆伝搬学習において従来使用されてきたエネルギー関数に一致することに注意する。

2.1 特徴抽出向きのエネルギー関数

E_H をエネルギー関数として学習したパーセプトロンと主成分分析が獲得するデータ表現の違いを、例をあげて示す。図 2は、$(N, H) = (16, 4)$ なる条件で主成分分析および 5 層非線形パーセプトロンによる情報圧縮を行って得られた誤差 $\{E_i\}_{i=1}^{H}$ を示している (実験に用いたデータの詳細については、3節参照)。パーセプトロンの第 2 層および第 4 層にはそれぞれ 32 個のシグモイド素子を用い、E_H をエネルギー関数として恒等写像学習を行った。主成分分析によって得られるパラメータは順序データ表現を成し、全ての $i = 1, 2, \ldots, H$ に対して、恒等写像近似 f_i は E_i を最小にする最適な線形写像となる。そのため、H次元のデータ表現から取り出すパラメータの数 i を任意に設定しても有効な特徴パラメータを構成することができる。これに対してパーセプトロンの場合、H個の中間層素子を全て用いる限りは同条件の主成分分析よりも小さい誤差が得られるが、$i < H$の部分パーセプトロンの誤差は主成分分析の結果よりもかなり大きくなっている。したがって、そのデータ表現から一部のパラメータのみを取り出しても、有効に利用することができない。

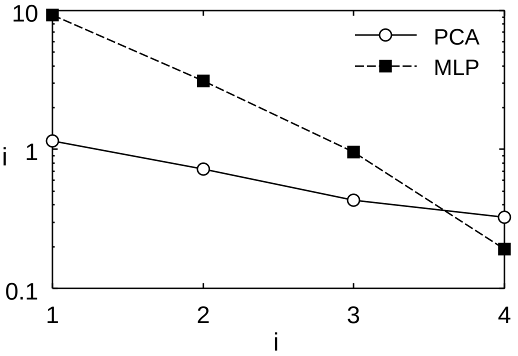


図 2: 主成分分析と多層パーセプトロンのデータ表現の比較 (1) : i–E_i特性

多層パーセプトロンのデータ表現に関するこの問題は、学習に用いるエネルギー関数 E_H に中間層素子を順序付けるための制約がないことに起因する。多層パーセプトロンの学習においても特徴抽出に有効な順序データ表現を自己組織できるようにするためには、全ての E_i を同時に最小化するように学習を行うことが考えられる。そこで、E_H に代えて E_iの線形和をエネルギー関数とする学習の効果を検討する [6, 7]。エネルギー関数 Fを次式で定義する。

$$F = \sum_{i=1}^{H} \beta_i E_i \tag{4}$$

ただし、β_iは正定数とする。このようなエネルギー関数を用いると、番号の小さな中間層素子は素子数の少ない部分パーセプトロンで恒等写像近似を実現しようとする。その結果、番号の小さな中間層素子ほど重要な特徴を抽出し、順序データ表現が自己組織されると期待できる。特に、線形パーセプトロンの恒等写像学習にこのエネルギー関数を適用した場合については、その特性が理論的に解析されており、中間層素子のパラメータが入力データの主成分に一致することが証明されている [6]。本論文では、このエネルギー関数 Fを重畳型エネルギー関数と呼ぶ。

2.2　データ圧縮向きのエネルギー関数

一方データ圧縮においては、全てのデータを同数のパラメータで表現するよりも、個々のデータを必要最小限のパラメータ数で記述することが望まれる。例えば、圧縮による近似誤差の許容値があらかじめ定まっている場合、各データについて許容値を満足する最小の数のパラメータのみを格納するようにすれば、効率的なデータ圧縮が実現できる。このような観点から、以下の手順で求められる平均必要パラメータ数とそれに対応した平均誤差を評価基準として主成分分析と多層パーセプトロンのデータ表現を比較する。

1. 誤差の許容値 $Z \in R^1$を定める。
2. 各 t について、式 (2) の誤差 $e_{i,t}$ $(i = 0, 1, \ldots, H)$ を計算する。ただし、$e_{0,t} \equiv ||\boldsymbol{x}_t - \mathrm{E}[\boldsymbol{x}_t]||^2$とする。
3. 各 t について、次式より必要パラメータ数を決定する。

$$I(t, Z) = \begin{cases} H & (\text{すべての } i \text{ で } e_{i,t} \geq Z \text{ の場合}) \\ e_{i,t} < Z\text{を満たす最小の } i & (\text{上記以外}) \end{cases} \tag{5}$$

4. 平均必要パラメータ数

$$I_Z = \frac{1}{T}\sum_{t=1}^{T} I(t, Z) \tag{6}$$

とそれに関連した平均二乗誤差

$$G_Z = \frac{1}{T}\sum_{t=1}^{T} ||\boldsymbol{x}_t - f_{I(t,Z)}(\boldsymbol{x}_t)||^2 \tag{7}$$

を計算する。

寄与の大きなパラメータが上位に集中している場合、恒等写像近似が容易なデータに対しては、誤差を増加させずに $I(t, Z)$ を Hよりもかなり小さくできると考えられる。したがって、誤差 G_Zを E_Hと同程度に保ったままで平均パラメータ数 I_Z を小さくすることができるならば、パラメータの記述に必要なデータ量を削減して効率的なデータ圧縮が実現可能と言える。

図 3は、2.1節に示した主成分分析と多層パーセプトロンのデータ表現を用い、I_Zと対応する G_Z を求めた結果を示す。この場合も順序データ表現を成す主成分分析の方が効率が良く、$E_H(= G_0)$ と比較して、誤差を増加させずにパラメータ数を大幅に削減することができる。一方、通常のエネルギー関数を用いて学習したパーセプトロンは順序データ表現を成さないため、誤差を増加させずにパラメータの記憶に必要なデータ量を削減することはできない。

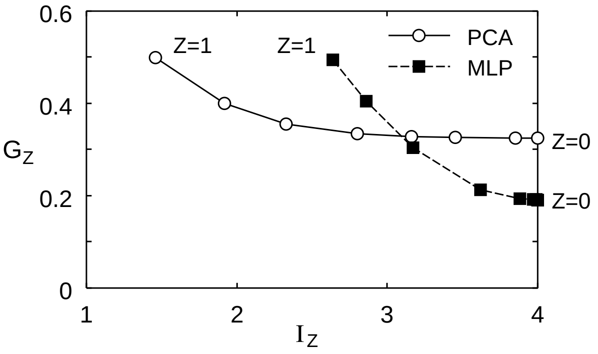


図 3: 主成分分析と多層パーセプトロンのデータ表現の比較 (2)：I_Z–G_Z 特性

多層パーセプトロンでデータ圧縮に有効な順序データ表現を得るため、式 (7) を形式的なエネルギー関数とみなし、これを直接最小化する学習を考える。式 (7) は $Z = 0$ で通常のエネルギー関数 E_Hに一致するが、Zの値を適当に与えれば、この許容値 Zを満たす最小の素子数の部分パーセプトロンのみが学習を行うことになる。その結果、G_Zをエネルギー関数とするパーセプトロンは、各データを必要最小限の中間層素子数で近似するようになると期待できる。本論文では、G_Zを適応選択型エネルギー関数と呼ぶ。

3　実験

上述の二つのエネルギー関数を用いた学習の有効性を示すため、5 層非線形パーセプトロンによる画像情報圧縮の実験を行う。対象画像には、SIDBA 標準画像 “Girl” (256 × 256 画素、8 ビットグレイスケール) を選択した。これを 8×8 画素のブロックに分割して 64 次元のデータ $\{\boldsymbol{v}_t \in R^{64}\}_{t=1}^{1024}$を生成し、その上位 16 個の主成分をパーセプトロン

への入力 $\{\boldsymbol{x}_t \in R^{16}\}_{t=1}^{1024}$ とする。パーセプトロンの素子数は入力層から順に 16–32–H–32–16(ただし $H=4$ または 8) とし、第 2 層および第 4 層にはシグモイド型の非線形素子を、その他の層には線形素子を用いた。これらのパーセプトロンに対して以下の三種類のエネルギー関数を適用し、恒等写像学習を行う。

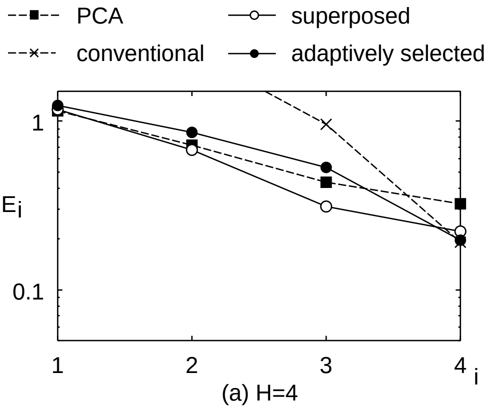
- 従来のエネルギー関数 E_H
- 重畳型エネルギー関数 F
  定数 $\{\beta_i\}$ は $\beta_i \propto i^2$ および $\sum_{i=1}^{H} \beta_i = 1$ より決定する[1]。
- 適応選択型エネルギー関数 G_Z
  許容値は $Z = \frac{1}{T}\sum_{t=1}^{T} e_{H,t} = E_H$ とし、Zの値も学習中に浮動させる[2]。

図 4は、学習後のパーセプトロンにおける誤差 $\{E_i\}_{i=1}^{H}$を示す。すべての中間層素子を用いる場合には、どのエネルギー関数を用いて学習したパーセプトロンも同条件の主成分分析による誤差を下回っている。重畳型で得られた誤差 E_Hは他の二種のパーセプトロンの誤差 E_Hよりも多少大きいが、$i < H$の部分パーセプトロンについては、重畳型の誤差が最小となっている。この場合、すべての i について同次元の主成分分析の誤差を下回る順序データ表現が得られており、主成分分析と等しい誤差の特徴量をより少数のパラメータで記述することが可能となっている。

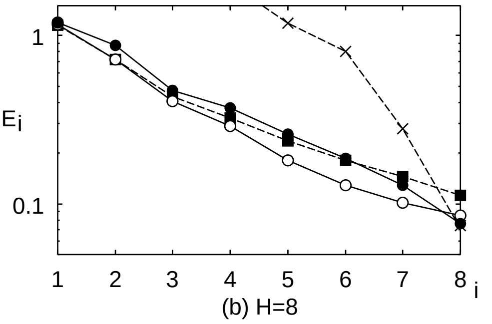
次に、学習後のパーセプトロンを用いて 2.2節に述べた方法で I_Zと G_Zを計算する。図 5は、

$$Z \in \{1.0, 0.5, 0.25, 0.1, 0.05, 0.025, 0.01, 0\} \tag{8}$$

として計算された (I_Z, G_Z) の値を示す。従来のエネルギー関数で学習したパーセプトロンでは、誤差を増加させずに平均必要パラメータ数を減らすことができないのに対して、重畳型と適応選択型の場合、誤差をほとんど増やさずにパラメータ数を半分近くまで減らすことができる。したがって、これらの獲得したデータ表現はデータ圧縮に適したものになっており、任意のZの値に対して主成分分析よりも高い圧縮性能を得ることが可能である。



(a) H=4



(b) H=8

図 4: 多層パーセプトロンによる画像データの情報圧縮 (1) : i–E_i特性

4 議論

本論文では、多層パーセプトロンの恒等写像学習によって順序データ表現を自己組織する二つのエネルギー関数を検討した。多層パーセプトロンによる情報圧縮を特徴抽出に利用する場合には、重要な特徴を表すパラメータのみ取り出せるようなデータ表現を構成することが望まれる。したがって、すべての部分パーセプトロンの誤差 E_iを主成分分析よりも小さな値に収束させることのできる重畳型のエネルギー関数が有効である。一方、データ圧縮に利用する場合には、パラメータの記述に必要なデータ量が平均として削減されれば良い。この場合、図 5が示すように、重畳型と適応選択型のどちらも効率的なデータ表現を自己組織する

---

[1] 各中間層素子に結合する重みの間で実質的な学習係数に大きな差がつくことを避けるため、このような値を選択した [7]。

[2] 無論 Zを適当な値に固定して学習しても良いが、結果にほとんど差が見られないため、このような方法を選択した。

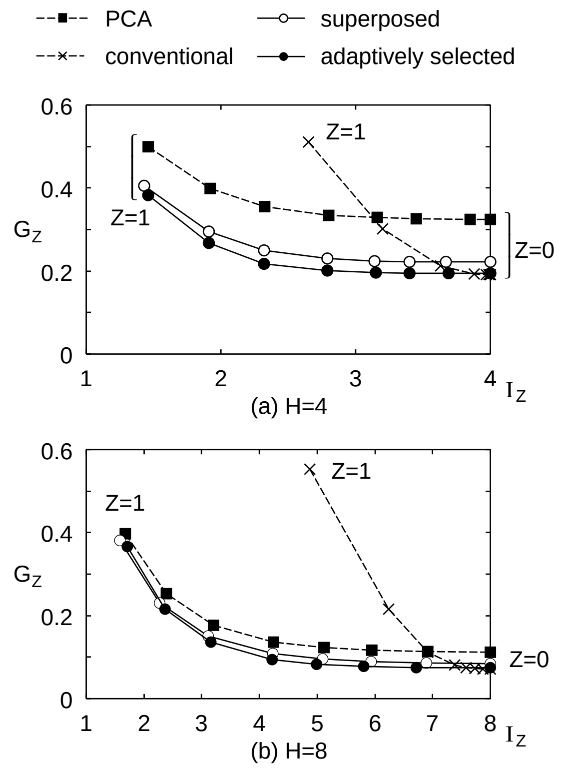


図 5: 多層パーセプトロンによる画像データの情報圧縮 (2) : I_Z–G_Z 特性

ことができる。しかし実際の応用上は、学習に要する計算コストが問題となる。この点で両者を比較するため、3節の実験に用いたパーセプトロン ($H=8$) において、一つのデータに対する一回の学習当たり必要な乗算回数を、出力の計算 (前進) と誤差の逆伝搬および重みの修正 (後退) の二つに分けて算出した。結果を表 1に示す。この例では、重畳型は従来の学習のおよそ $H/2$ 倍の計算量を要するのに対して、適応選択型は従来のものよりも少ない計算量で計算可能となっている。したがって、データ圧縮への利用においては、適応選択型のエネルギー関数が適していると考えられる。

表 1: 計算量の比較

	従来型	重畳型	適応選択型
前進	1600 (回)	5152	1344
後退	2624	11808	2112
計	4224	16960	3456
比率	1	4.02	0.82

参考文献

[1] Bourlard, H. and Kamp, Y., “Auto-association by multilayer perceptrons and singular value decomposition,” Biol. Cybern., 1988.

[2] Baldi, P. and Hornik, K., “Neural networks and principal component analysis: Learning from examples without local minima,” Neural Networks, vol.2, pp.53–58, 1989.

[3] 船橋賢一, “3 層ニューラルネットワークによる恒等写像の近似的実現についての理論的考察,” 信学論 (A), vol.J73-A, no.1, 1990.

[4] Usui, S., Nakauchi, S. and Nakano, M., “Internal color representation acquired by a five-layer neural network,” Proc. Int. Conf. Artificial Neural Networks, vol.1, pp.867–872, 1991.

[5] DeMers, D. and Cottrell, G., “Non-linear dimensionality reduction,” in Advances in Neural Information Processing Systems 5, eds. S. J. Hanson, J. D. Cowan and C. L. Giles, pp.580–587, 1993.

[6] 高橋隆史, 徳永隆治, “KL 変換を実現する 3 層線形パーセプトロンの教師付き学習則 — Baldi–Hornik の定理の拡張 —,” 信学論 (D-II), vol.J80-D-II, no.5, pp.1267–1275, 1997.

[7] 高橋隆史, 徳永隆治, “重畳エネルギー関数による多層パーセプトロンの冗長性削減,” 信学論 (D-II), vol.J80-D-II, no.9, pp.2532–2540, 1997.