

距離評価に基づく認識のための次元圧縮

佐藤 美沙紀[†] 平岡 和幸^{††} 三島 健稔^{††}埼玉大学大学院理工学研究科[†]埼玉大学工学部情報システム工学科^{††}

1. はじめに

コンピュータが現実世界のものを認識する場合、現実世界で観測された数値をベクトルの形で入力としてパターン認識を行うこととなる。現実世界で観測された数値ベクトルは多くの成分を持つ。しかし、そのような高次元の入力を用いると、計算量や記憶領域、学習に必要なデータ数が膨大となる問題が発生する。[1]

これらの問題を避けるためには、パターン認識の入力データに対し、認識に有効と期待される成分のみを見出す次元圧縮と呼ばれる前処理を行い、入力とすることが有効である。次元圧縮は、パターン認識の学習に使われるようなクラス情報(データの認識結果の正解を示す情報)を持ったデータ(訓練データ)により学習を行う。

現在、パターン認識のための次元圧縮で広く用いられている手法に、主成分分析(PCA: Principal Component Analysis)、線形判別分析(LDA: Linear Discriminate Analysis)がある。しかし、これらの手法には欠点がある。まず、PCAは訓練データの持つクラス情報を利用しない。また、LDAはクラス数により圧縮後の次元数が制限される。さらに、PCAは全データ、LDAは各クラスのデータを1組の平均と分散行列で記述する。そのため、認識すべきデータが複雑な分布をしている場合、パターン認識に有効な次元圧縮が得られないことがある。

これらの欠点を克服するために、異なるクラスに属するデータ間の距離に着目した。パターン認識に有効な次元圧縮を得るということは、圧縮後の空間でパターン認識を行いやすくするということを意味する。そのためには、異なるクラスのデータが離れて分布していることが重要であると考えられる。そこで、本研究では、2クラスの認識問題を扱い、以下に述べる3通りの評価基準に基づいた次元圧縮を検討した。

2. データ対の距離評価に基づく方法

訓練データとして、各クラス $c=1, 2$ に属する d_1 次元のデータベクトル $\mathbf{x}$ をそれぞれ N_c 個与える。このデータから d_2 次元の圧縮データ $\mathbf{y}$ への圧縮は圧縮行列 A を乗ずる($\mathbf{y}=A^T\mathbf{x}$)ことで行われる。本研究の目的は、 $\mathbf{y}$ に基づくパターン認識の精度が高くなる望ましい A を見出すことである。

この圧縮行列 A を定めるため、異なるクラスに属する圧縮した空間でのデータ間のユークリッド距離 L に基づく評価値 $\varphi(L)$ を用いる。全データ対に対して $\varphi(L)$ を求め、その平均値 $P(A)$ を A の評価とする。

$$\varphi(L) = \exp\left(-\frac{L^2}{\omega^2}\right)$$

$$P(A) = \frac{1}{N_1 N_2} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \varphi(L(\mathbf{y}_1(i), \mathbf{y}_2(j)))$$

ω は定数であり、距離基準に応じて定める。

$\varphi(L)$ は距離 L が大きくなるほど小さくなる。

この特性をふまえ、 $P(A)$ が最小となる A を用いて次元圧縮を行う。

この手法をパターン認識の前処理に用いると、PCA、LDA を用いた場合より高い正答率が得られたが、計算量の問題があった。 A を求めるために全データ対に対して $\varphi(L)$ を求めるため、訓練データ数 n に対して、計算量が $O(n^2)$ となる問題である。そこで、次に前述の特性を活かしつつ、計算量の削減を試みる。

3. 正規分布対の距離評価に基づく方法

計算量の問題を緩和する方法として、データ対の削減が考えられる。そこで、訓練データに正規混合分布(GMM: Gaussian Mixture Model)をあてはめ、異なるクラスに属する正規分布間の相違を定量化して「分布間の距離」の定義とする。この距離 L に対する $\varphi(L)$ を用いて、先ほどと同様に $P(A)$ を定める。

これにより、距離評価を行う対は減少する。 n 個のデータが k 個($n > k$)の正規分布の混合分布であてはめられた場合には、計算量は $O(k^2)$ となり、 $n > k$ より計算量が大幅に削減される。具体的には、クラス c の d_1 次元の訓練データに k_c 個の正規分布の重み付き足し合わせでなる密度関数

$$p_c(\mathbf{x}) = \sum_{i=1}^{k_c} e_c(i) G(\mathbf{x}; \mathbf{m}_c(i), S_c(i))$$

をあてはめる。 $G(\mathbf{x}; \mathbf{m}, S)$ は平均ベクトル $\mathbf{m}$ 、分散行列 S の多次元正規分布の密度関数であり、 $e_c(i)$ は $e_c(i) > 0$ 、 $\sum_{i=1}^{k_c} e_{c,i} = 1$ を満たす各々の正規分布の重み係数である。GMM のパラメータ推定は EM アルゴリズムで行う。[2]

確率密度関数 $p_c(\mathbf{x})$ は次元圧縮 $\mathbf{y}=A^T\mathbf{x}$ により、

d_1 次元の正規分布 $G(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$ から d_2 次元の正規分布 $G(\mathbf{y}; \boldsymbol{\mu}, \Sigma)$ (但し、 $\boldsymbol{\mu} = A^T \mathbf{m}$ 、 $\Sigma = A^T S A$) に変換される。

クラス c の GMM 密度関数に用いられる正規分布 $G(\mathbf{x}; \boldsymbol{\mu}_c(i), S_c(i))$ をそれぞれ重みつきデータのようにみなし、異なるクラスに属する2つの正規分布間の距離を圧縮した空間で考える。

この距離 L に対する $\varphi(L)$ を全ての対でそれぞれ求め、その $P(A)$ が最小となる A を圧縮行列として採用する。正規分布間の距離の測り方は、次の2つを検討する。

(ア) KL 情報量による距離基準

第一の距離基準として、KL 情報量 (Kullback-Leibler Divergence) を利用する。KL 情報量は2つの確率密度関数 $p(x)$ 、 $q(x)$ に対して、

$$D(p \parallel q) = \int p(x) \log \left(\frac{p(x)}{q(x)} \right) dx$$

と定義される。 D は2つの分布が一致する場合に0をとり、異なるほど大きな値をとる。そこで、 $L = \sqrt{D(p \parallel q)}$ を距離基準とする。

(イ) 楕円面接点による距離基準

KL 情報量による距離基準では、「平均が等しく、分散の形状が極端に異なる場合」、「平均が離れていて、分布の形状が似通っている場合」、いずれも距離が大きくなる。しかし、パターン認識においては前者より後者が有用であると考えられる。そこで、KL 情報量とは異なる第二の距離基準として、正規分布に従う確率密度関数の等値点がなす楕円面に着目した。

正規分布の等値点がなす楕円面は、

$$(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) = r^2$$

と表される。 r は定数である。同次元の正規分布では、この楕円面の内側に標本が属する確率は r のみで定まり、平均や分散に依存しない。この性質をふまえ、共通の r により、2つの正規分布に従う楕円面を考える。2つの楕円面が接する唯一の r を求め、これを分布間の距離 L として用いる。

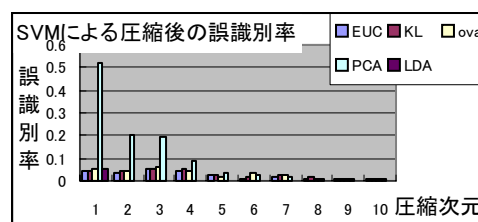
これにより、KL 情報量とは異なる観点から、分布の相違具合を捕らえることができると、期待する。

4. 実験

ここでは、まず PCA、LDA との比較実験を示し、次に3通りの提案した手法について、計算量を検討する。

提案した手法の評価値の最適化には、

Nelder-Mead シンプレックス法を用いた。また、識別には SVM[3] を用い、実験データとして 0、Q の 16 次元特徴量[4]を用いた。



誤識別率に関しては1次元では LDA よりわずかに低く、PCA とは低次元では大きく上回る性能を示せたと考えられる。

計算時間については、1次元への圧縮行列の最適化には、全データ対を評価する手法では $5.6 \times 10^5 \text{sec}$ 、それに対し、KL 情報量による評価では $3.8 \times 10^2 \text{sec}$ 、楕円面を用いた評価では $2.6 \times 10^3 \text{sec}$ とデータを GMM に近似するための計算時間であった $1.9 \times 10^3 \text{sec}$ を含めて考えても、大幅に削減された。

5. まとめ

パターン認識に有効な次元圧縮手法として、距離評価による手法を提案した。まず、訓練データを圧縮し、ユークリッド距離を基に評価する方法を提案した。PCA、LDA と比較し、特に低次元で、よい識別率が得られたが、これには計算量の問題があった。そこで、訓練データを GMM にあてはめ、圧縮後の分布間の距離を評価することで、計算量の削減を図った。この手法では、計算量を削減し、かつ、先の手法に近い結果が得られた。

提案した手法と PCA による識別率が近くなる次元では、PCA に劣ることもあった。これは、最適化に問題があると考えられる。

参考文献

- [1] 坂野鋭, 山田敬嗣, “怪奇!!次元の呪い—識別問題、パターン認識、データマイニングの初心者のために—”, 情報処理 43 巻 5, 6 号, 2002.
- [2] 渡辺澄夫, “データ学習アルゴリズム”, 共立出版, 2001.
- [3] Cawley, G. C, “MATLAB Support Vector Machine Toolbox (v0.54 β)” [<http://theoal.sys.uea.ac.uk/~gcc/svm/toolbox>], University of East Anglia, School of Information Systems, Norwich, Norfolk, U. K. NR47TJ, 2000.
- [4] Murphy, P. M., & Aha, D. W. (1994). UCI Repository of machine learning databases [<http://www.ics.uci.edu/~mllearn/MLRepository.html>]. Irvine, CA: University of California, Department of Information and Computer Science.