

I-50 パターンカテゴリ閾値表現法とパターン認識法

Pattern category threshold expression method and pattern recognition method

劉 紹明†

Shaoming Liu

1. まえがき

本研究では、パターンカテゴリの表現法及びパターンの比較方法に着目して、パターンカテゴリの分布によく近似できるカテゴリ閾値表現法、及びその表現法に基づくパターン間の比較尺度を提案する。また、13種類のマルチ印刷文字サンプル(471万)を用いて提案方法の機能検証実験を行う。

2. 従来技術と問題点

パターンカテゴリの分布を如何に表現するかはパターン認識分野における大きな問題である⁽¹⁾。従来技術における基本的な方法として、カテゴリに属している学習サンプルの平均ベクトルでカテゴリを代表する方法がある。パターンを比較する尺度はこれまで数多く提案されており、代表的なものとして、シテイブロック距離⁽²⁾、ユークリッド距離⁽²⁾、重み付きユークリッド距離⁽²⁾、投影距離⁽³⁾などが挙げられる。従来技術には、次の2つの特徴がある。

- ① 代表パターン(複数可)でパターンカテゴリを表す。
- ② パターンと代表パターン間の距離(類似度)を用いて入力された未知パターンを比較する。

従来技術では、標準パターンの認識範囲が比較的大きく、よく重なるので、認識精度に影響を与えるという問題が生じる。

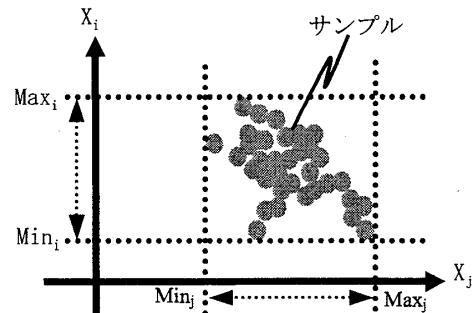
3. 問題を解決するアプローチ

従来技術の問題点を解決するために、本研究では、人がものを想起するとき、想起された情報からどのような情報が脳に記憶されているかを仮定し、その仮定を用いた新しいパターンカテゴリの表現方法及び認識方法を提案する。

人間の記憶は、記憶対象の概念、名前とともに事象の各特徴及び特徴量の範囲も記憶していると考えられる。例えば、“リンゴ”を思い出すとき、“色は赤い、黄色い或いは青いなどがあり、黒ではないこと；味は甘い、甘酢っぱいなどがあり、辛くはないこと；重さが150グラム位～450グラム位；”などが自然に思い出される。つまり、人間は学習するとき、学習対象の各特徴量を取って、各特徴及び特徴量の変化範囲を記憶していると考えられる。勿論、人間は連想という機能を持っているので、未学習のリンゴも認識できる。これは、未学習のリンゴが学習したリンゴに似ているからである。上記の分析により次のパターンカテゴリ閾値表現法を提案する。

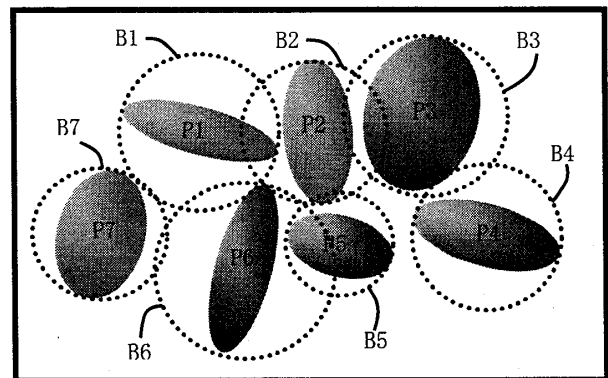
3.1 パターンカテゴリの閾値表現法

上記の仮説を用いて、パターンカテゴリを次のように表現できる。カテゴリ内に属しているすべての学習サンプル

図1 X_i と X_j 次元の位置範囲

から、特徴量の各次元毎に、次元の値の最小値及び最大値を求め、求められた各次元の最小値と最大値ベクトルを該パターンカテゴリの代表とする。図1はパターンカテゴリに属しているすべての学習サンプルが i 次元 X_i と j 次元 X_j に現れる位置範囲を示している。 Min_i と Max_i はそれぞれ X_i 次元の最小値と最大値であり、 Min_j と Max_j はそれぞれ X_j 次元の最小値と最大値である。

パターンカテゴリ代表の作成方法から分かるように、パターンカテゴリ代表は、カテゴリに属しているすべての学習サンプルを含む外接多次元長方形である。この外接長方形は多次元空間に該パターンカテゴリに属しているすべての学習サンプルが各次元毎に現れる位置範囲を表し、該パターンの認識範囲である。外接長方形は比較的パターンの学習サンプルの分布に近いので、認識範囲が重なっているパターンカテゴリの数を大幅に削減することができる。例えば、図2に示す7つのパターン $P1, P2, \dots, P7$ について、従来の技術により、 $P1 \sim P7$ の認識範囲は図2(a)に示している点線 $B1 \sim B7$ である。 $B1$ は $B2$ 及び $B6$ と、 $B2$ は $B1, B3, B5$ 及び $B6$ と、 $B3$ は $B2$ 及び $B4$ と、 $B6$ は $B1, B2, B5, B7$ と重なっている。しかし、本研究の手法により、パターンの認識範囲は図2(b)に示す $B1' \sim B7'$ である。図から分かるように、 $B1', B2', \dots, B7'$ は相互に重なっていない。



(a)

† 富士ゼロックス(株) ITメディア研究所
Information Media Lab., Fuji Xerox Co., Ltd.

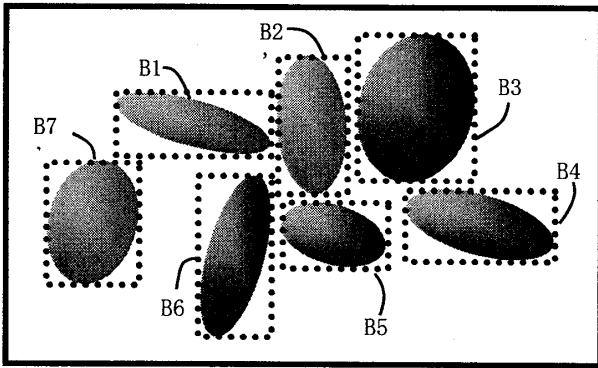


図2 (a)従来技術の認識範囲, (b)提案技術の認識範囲

3.2 パターンの比較方法

本節では、提案したパターンカテゴリ閾値表現法を用いてパターン認識を行うときに、パターンの比較尺度を提案し、提案した類似度の計算法を示す。

類似度

入力された未知パターン $X=(x_1, x_2, \dots, x_n)$ と認識辞書に格納しているパターンカテゴリ p の代表 $C(p)=[c_{\min}(p, 1), c_{\max}(p, 1)], \dots, [c_{\min}(p, n), c_{\max}(p, n)]$ の間の類似度 $S(X, C(p))$ は次式で計算される(今後、閾値類似度と呼び)。ここで、 $[c_{\min}(p, i), c_{\max}(p, i)]$ はパターンカテゴリ代表の i 次元のデータである。

$$S(X, C(p)) = \frac{\sum_{i=1}^n f(x_i, c_{\min}(p, i), c_{\max}(p, i))}{n}$$

ここで、

$$f(a, b, c) = \begin{cases} 1, & \text{if } b \leq a \leq c; \\ 0, & \text{if } a < b \text{ or } a > c; \end{cases}$$

関数 $f()$ の定義から分かるように、入力された未知パターン X の i 次元の値 x_i はカテゴリ代表の i 次元の値の変化範囲に入ると、類似度がすこし高くなる。逆に、入力された未知パターン X の i 次元の値 x_i はカテゴリ代表の i 次元の値の変化範囲以外に入ると、類似度がすこし低くなる。すべての次元に対して、 $f()=1$ なら、類似度=1 であるので、カテゴリに属しているすべての学習サンプルと該パターンのカテゴリ代表間の類似度は同じであり、“1”である。認識するとき、未知パターン X がパターンカテゴリ代表で示すパターン P の認識範囲に入ると、 $S(X, C(p))=1$ になり、パターン P が認識の結果として出力される。これは従来技術で実現できなかった部分である。

4. 活字文字サンプルを用いた機能確認実験

実験対象

実験に使用した文字カテゴリの範囲は J I S 漢字第一水準、記号など合わせて 3455 個文字であり、文字フォントは石井中明朝体など合計 13 種類である。

特徴量

本実験では、411 次元の複合特徴量(124 次元のペリフェラル特徴量+62 次元のストローク特徴量+225 次元のメッシュ特徴量)で文字パターンを表現する。

機能検証実験 1 とその結果

提案したカテゴリ表現法及びカテゴリの中心値で表現する方法を用いて、カテゴリの認識範囲の平均重なり具合を求めた。表 1 は機能検証実験 1 の結果である。

$$\text{平均重なり数} = \frac{\sum (\text{文字カテゴリ } p \text{ の認識範囲と重なっている文字カテゴリ数})}{\text{総文字カテゴリ数}}$$

表 1 認識範囲の重なり数

	類似度 or 距離	平均重なり数
提案した閾値表現法	提案した類似度	113.16
従来の中心値表現方法	シテイブロック距離	3417.08
従来の中心値表現方法	ユークリッド距離	3448.29
従来の中心値表現方法	重み付きユークリッド	3095.52
従来の中心値表現方法	投影距離	2877.96

機能検証実験 2 と実験結果

本研究で提案した認識方法の有効性を検証するために、提案した閾値類似度及び従来の距離尺度を用いて、学習サンプル(249 万)と未学習のテストサンプル(222 万)の認識実験を行った。実験方法として入力された文字パターンを全文字カテゴリ代表と比較して正解を求める総当たり法を採用した。表 2 は実験の結果を表している。ここで、認識率は正しく認識されたサンプル数のテストの総サンプル数に対する割合とする。また、実験に使用したコンピュータはパーソナルコンピュータ(CPU : Pentium 450, OS : NT4.0)であり、使用した言語は VC++ であった。

表 2 認識結果

	学習	未学習	時間
シテイブロック	88.4%	87.6%	34ms
ユークリッド	88.8%	87.8%	20ms
重み付きユークリッド	93.6%	92.4%	92ms
投影距離	97.8%	97.6%	88ms
閾値類似度	98.4%	97.4%	17ms

5. まとめ

本研究では、パターンカテゴリの閾値表現法及びパターンとカテゴリ代表間の類似度を提案し、活字文字サンプルを用いて機能検証実験を行った。実験結果から分かるように、投影距離認識法では、認識率は高いが、認識時間が長いという欠点がある。逆に、ユークリッド距離認識法では、認識時間は短いが、認識率が低いという欠点がある。従来技術と比較すると、提案手法の認識率は投影距離認識法の認識率とほぼ同じであるし、認識時間はユークリッド距離認識法の認識時間とほぼ同じである。パターンの分布による閾値の設定方法は今後の研究課題である。

6. 参考文献

[1]. 小川 英光：“パターン認識・理解の新たな展開—挑戦すべき課題—”，電子情報通信学会，1994。
 [2]. 安居，長尾：“画像の処理と認識”，昭晃堂，1992。
 [3]. 池田，田中，元岡：“手書き文字認識における投影距離法”，情処論，vol.24, no.1, pp.106-112, 1983。