

Test Feature Classifier による逐次パターン分類

北大院 ○坂田幸辰 金子俊一 田中孝之

要旨

ノンパラメトリックなパターン分類法として考案されたテスト特徴識別器の逐次学習に関する特性について検討する。この特性により、学習型識別器の作成に関して様々な効果が期待できる。本論文では、テスト特徴に基づく逐次学習型識別器のアルゴリズムを考案し、その有効性を確認している。

1. はじめに

我々は従来より、特徴空間における部分組合せ特徴量の中から、単純な非重複条件を満たす素な部分組合せ特徴量を用いる識別器として、テスト特徴識別器 (Test Feature Classifier, 或いは TFC) を提案し、数々の応用問題に適用してその有効性を確認してきた [1][2]。本研究では TFC に基づく逐次学習アルゴリズムを提案する。TFC においては、多くの部分組合せ特徴量の非重複条件を調べて登録する手続きが学習処理に相当するが、やや計算コストを要していた。これを逐次化することによって、効率的な学習処理を構成することが可能となる。また、時間経過と共に新しいデータが追加されるような実応用分野においては追加されたデータを逐次的に扱う機能をもつ識別器、つまり逐次学習アルゴリズムは必須のものである。本研究の目的は、従来著者達が提案してきたノンパラメトリックなパターン識別器であるテスト特徴識別器に基づく、独自の逐次パターン学習アルゴリズムを提案し、その実応用における可能性について検討することである。

2. テスト特徴法

図 1 は TFC の基本的枠組を示す。TFC はある条件を満たす部分組合せ特徴量による局所識別を累積することによってパターンを分類するものであり、小さな訓練パターン集合を用いても高い識別率を得ようとするものである。クラス間に重複のない訓練パターン集合を与える。任意の組合せ部分特徴空間においてクラス間に重複が生じない (同一の特徴記述が一つもない) とき、すなわちすべての訓練パターンがクラスごとに区別可能であるとき、非重複条件を満たすと呼ぶ。全特徴空間は非重複条件を満たす。この条件を満たす組合せ部分特徴量をテスト特徴量 (以下、TF) と呼ぶ。この中で、いわゆる素なテスト特徴量をプライムテスト (以下、PTF) と呼ぶ。図上に示す PTF は模式例である。すなわち、PTF はその中のいかなる特徴量を除いても TF でなくなるような TF である。複数の PTF を抽出・登録して識別器を構成する。未知パターンが与えられたときには、各 PTF に対して未知パターンを投影し、同じく投影されたすべての訓練パターンとの一致あるいは近接度を評価する。各クラスごとの評価値をそのクラスのスコアとする。全登録 PTF に関してスコアを算出し、各クラスごとの累計スコアを求める。最大累計スコアをもつクラスを分類クラスとする。

TFC の特徴をまとめる。

- ノンパラメトリックな学習型分類法である。
- 一致に基づく投票方式の TFC においては、訓練パターンの誤分類がないことが保証されている。
- 抽出される TF 或いは PTF は局所的な識別手がかりを与える部分空間であり、他の様々な識別原理と併用することができる。

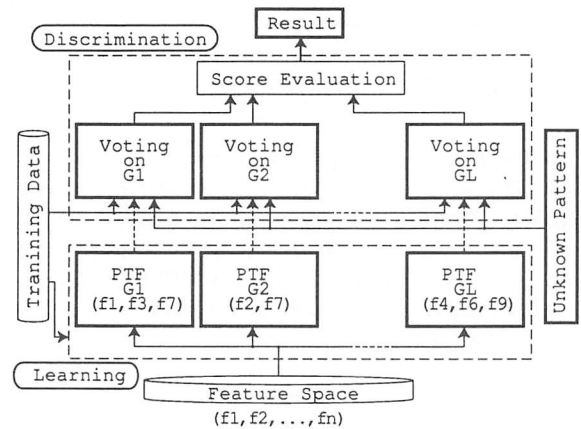


図.1 テスト特徴法の基本的枠組み

- 比較的少ない訓練パターン数で識別器を構成することが可能である。

3. 逐次学習型識別器

データの追加によって識別器を更新する際に最初から計算しなおすのではなく、追加データを逐次的に扱っていく識別器を逐次学習型識別器と定義する。すなわち、新しい PTF 集合の算出に更新前の PTF 集合を利用する識別器のことである。

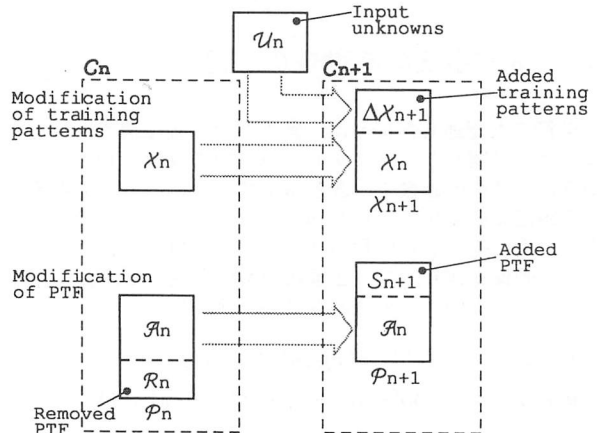


図.2 逐次学習型識別器

n 回目における識別器を C_n , 識別器内の訓練パターン集合を \mathcal{X}_n , TF 集合を \mathcal{T}_n , PTF 集合を \mathcal{P}_n , 非 TF 集合を \mathcal{N}_n と定義する。全特徴空間を \mathcal{F} とするとき、 $\mathcal{F} = \mathcal{T}_n + \mathcal{N}_n, \mathcal{P}_n \subseteq \mathcal{T}_n$ が成り立つ。 $\{\mathcal{T}_n | T_1, T_2, \dots, T_l\}, \{\mathcal{P}_n | P_1, P_2, \dots, P_m\}$.

逐次学習のアルゴリズムを次のように定義する。

1. [初期化]
 $n=0$. \mathcal{P}_0 を算出し, 識別器 C_0 を作成する.
2. [分類]
 識別器 C_n をもちいて未知パタンを分類していく. その後, 結果をみて追加パタン x を決定する.
3. [学習パタン集合の更新]
 追加パタン x を追加し, $\mathcal{X}_{n+1} = \mathcal{X}_n + \{\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k\}$ とする.
4. [プライムテスト特徴の削除]
 全ての $P_m \in \mathcal{P}_n$ に対して, PTF であるかどうかの判定を行い, PTF の資格を失った組み合わせ特徴量 P_m を \mathcal{P}_n から削除していく. $\mathcal{A}_n = \mathcal{P}_n - \{P_m\}$.
5. [新しいプライムテスト特徴の追加]
 前ステップにおいて, 削除された組み合わせ特徴量に対して, それに代わる新たな PTF を求める.
6. [終了判定]
 識別器 $C_{n+1} = (\mathcal{X}_{n+1}, \mathcal{P}_{n+1})$ の性能が十分であるときは学習を終了し, そうでなければ n を更新して 2. へ.

4. データの経時変化

日々新しいデータが追加されそれらを有効に利用することによりできるだけ高性能な識別器を実現したいような実応用分野においては, プロセスの経時変化によるデータの質的変動を考慮することが必要である.

経時変化するデータに対する逐次 TFC の特性を調べるために, 簡単なシミュレーションデータを使って実験をおこなった.

4.1 シミュレーションデータ

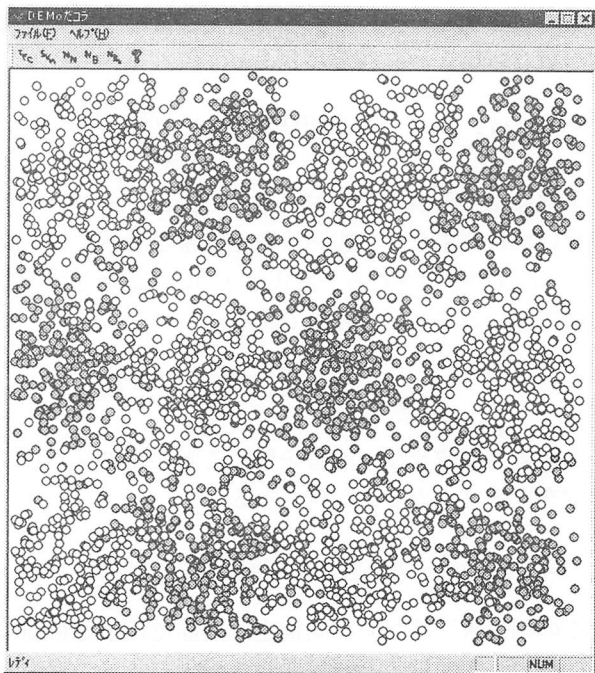


図.3 特徴空間における分布の例

データは2クラス8次元問題とし, 経時変化をシミュレーションするために3種類のデータを作成した. データ2はデータ1の特徴量7, 8を変化させたデータ, データ3はデータ2の特徴量5, 6を変化させたデータとなっている.

データは時間と共に1→2→3の順に変化していくものとする. シミュレーションデータの特徴空間での分布の例を図3に示す.

4.2 実験

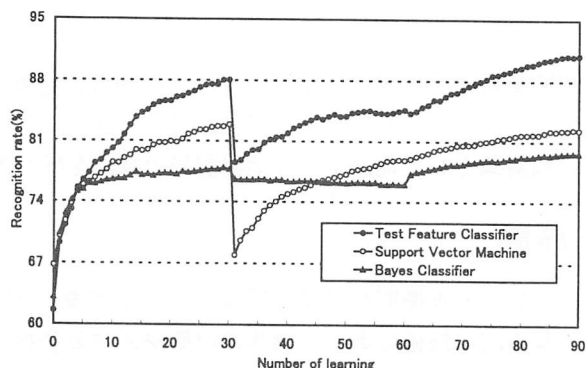


図.4 経時変化するデータに対する各手法の特性

C_0 はデータ1から各クラス5個づつ, 計10個の訓練パタンから構成した. 次に各クラス5個づつのデータを追加し C_1 とする. 同様の操作を $n=30$ までおこない, $n=31$ から $n=60$ までは追加データをデータ2へ, $n=61$ から $n=90$ まではデータ3へ変更する. 識別率の評価は $n=0$ から $n=30$ まではデータ1の, $n=31$ から $n=60$ まではデータ2の, $n=61$ から $n=90$ まではデータ3の各クラス250個から成る未識別データ集合を識別しておこなった. また, 逐次 TFC の特性を検証するために他の手法との比較もおこなった (比較手法は Bayes Classifier, Support Vector Machine). 結果を図4に示す.

TFC, SVM がデータの特性的変化に反応して識別率に変化が現れるのに対して BC にはほとんど変化が見られない. 全特徴空間での識別境界を考える TFC, SVM に対して, 1つ1つの特徴量から結果を推定する BC は特徴量間の相関を識別に利用することができないことが原因であると考えられる. また, 特性は似ているが全体的な識別率は TFC のほうが SVM よりも良好な結果を示している. これは TFC が全特徴空間のみを利用して識別するのではなく, 識別に有効な組み合わせ特徴量を選択的に利用しているため, 特徴量の一部が変化した場合でもある程度対応できるためである.

5. まとめ

経時変化するデータに対する逐次 TFC の特性の検証をシミュレーションデータを使った実験によっておこなった. また, TFC が SVM, BC に対して優れた性能を発揮することを確認した. 今後の課題としては, データの経時変化に対する有効な対応法の検討が挙げられる.

参考文献

1. イットコン, 金子俊一, 五十嵐悟, ラキシア ワクタン グ: "テスト特徴法の多値パタン認識への拡張とその実験的評価", T.IEE Japan, Vol.120-C, No.11, Nov, 2000
2. 坂田幸辰, 金子俊一, 高木祐治, 奥田浩人: "テスト特徴法に基づく逐次パタン学習と欠陥画像分類への応用", T.IEE Japan, Vol.124, No.2, Nov, 2004