ISS-P-117

粒子群最適化による 高次元データに適応した特徴選択

渡邉　文洸† 森　康久仁† 松葉　育雄†
†千葉大学大学院融合科学研究科

1. はじめに

情報社会の進歩に伴い高次元(多特徴)データが増加している．高次元データは有用なデータであるが冗長な特徴も多く含まれており，有効な特徴のみを選択する処理が必要である[1]．本研究の目的は粒子群最適化法(Particle Swarm Optimization)を用いて，選択する特徴の数を抑えつつ識別に有効な特徴を選択する方法を提案することである．

2. 粒子群最適化(PSO)による特徴選択

粒子群最適化は目的関数$f(\boldsymbol{x})$が与えられた際に，粒子を複数生成し局所最適解を共有しながら各粒子が探索空間内を動き最適解を求める手法である[2]．更新ステップtのi番目の粒子は速度情報$\boldsymbol{v}_i^{(t)}$と位置情報$\boldsymbol{x}_i^{(t)}$を持つ．粒子群最適化を用いた特徴選択では粒子の位置情報をバイナリ化する[3]．$\boldsymbol{x}_i^{(t)}$は特徴の採用を“1”・不採用を“0”とした特徴集合を表す配列とし特徴空間を探索する．粒子の更新は各粒子内の最適解を記録するParticle Best($\boldsymbol{pb}_i$)と全粒子内の最適解を記録するGlobal Best($\boldsymbol{gb}$)を使用し，式(1)，(2)で更新する．

$$v_{id}^{(t+1)} = \omega \times v_{id}^{(t)} + c_1 r_1 \left(pb_{id}^{(t)} - x_{id}^{(t)}\right) + c_2 r_2 \left(gb_d^{(t)} - x_{id}^{(t)}\right) \quad (1)$$

$$x_{id}^{(t+1)} = \begin{cases} 1, & if\ sigmoid\left(v_{id}^{(t)}\right) > U(0,1) \\ 0, & otherwise \end{cases} \quad (2)$$

$$(i = 1,2,\dots,N),\quad (d = 1,2,\dots.D)$$

ここでNは粒子数であり，Dは次元数である．$\omega, = 0.4$，$c_1 = c_2 = 2$としr_1, r_2は0～1の乱数とした．

3. 提案法による特徴選択

データが高次元になるにつれ特徴空間が拡がることにより粒子の軌跡上の情報のみでは十分な探索が困難となる．本研究では新たに重回帰分析に基づいたMultiple Regression analysis Best($\boldsymbol{mrb}$)を加え$\boldsymbol{pb}_i$, $\boldsymbol{gb}$, $\boldsymbol{mrb}$ の3つのBestで粒子を更新させる．$\boldsymbol{mrb}$は次の方法で生成する．

1. 粒子の軌跡から位置情報を重なりなくN個取得
2. 粒子の分布$\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N]$とその識別率 $\boldsymbol{Y}$から$\boldsymbol{Y} = \boldsymbol{\beta X}$より$\boldsymbol{\beta}$を算出
3. $\boldsymbol{\beta}$は特徴の重要度を表すので式(3)より$\boldsymbol{mrb}$を生成する

$$mrb_d = \begin{cases} 1, & if\ \beta_d > \epsilon \\ 0, & otherwise \end{cases} \quad (3)$$

また更新式を式(4)，(5)とした．

$$v_{id}^{(t+1)} = \omega \times v_{id}^{(t)} + c_1 r_1 \left(pb_{id}^{(t)} - x_{id}^{(t)}\right) + c_2 r_2 \left(gb_d^{(t)} - x_{id}^{(t)}\right) + c_3 r_3 \left(mrb_d^{(t)} - x_{id}^{(t)}\right) + c_4 r_4 \quad (4)$$

$$c_3 = accuracy(\boldsymbol{mrb})\ /accuracy(\boldsymbol{gb})$$

$$x_{id}^{(t+1)} = \begin{cases} 1, & if\ \left(x_{id}^{(t)} + v_{id}^{(t+1)}\right) > 0.5 \\ 0, & otherwise \end{cases} \quad (5)$$

ここでc_4は減衰する外力とし，r_3, r_4は0～1の乱数とした．

4. 実験・結果

提案法の有効性を確認するために4つのデータに対し実験を行った．各実験において粒子数はデータの次元数の1/5とし，選択数Nは粒子数×5とした．また，特徴の検定法と評価法はそれぞれk-fold(k=10)，1-NN法とする．表1，2に実験データと結果を示す．

表1.実験データ

データ	サンプル数	次元数	クラス数
WDBC	569	31	3
Bands	277	39	2
Arrhythmia	447	274	9
CNAE9	1080	856	9

表2.提案法と他の特徴選択の比較

データ	手法	選択特徴数	識別率(%)	データ	手法	選択特徴数	識別率(%)
WDBC	Proposed	7.75	95.2	Arrhythmia	Proposed	31.25	70.1
	PSO	15.75	95.0		PSO	145.00	65.3
	SFS	19.00	93.0		SFS	212.00	66.2
	GA	16.00	93.9		GA	140.75	66.3
Bands	Proposed	9.00	82.0	CNAE9	Proposed	128.00	94.5
	PSO	16.25	79.8		PSO	422.25	89.5
	SFS	2.00	79.1		SFS	615.00	94.7
	GA	18.00	76.4		GA	435.00	94.0

5. 考察

本稿では粒子群最適化を用いた特徴選択法に，重回帰分析から生成される新しいBestを加えた手法を提案した．新しいBestによって探索範囲が拡がり，他の特徴選択法より識別率が高くなったと考えられる．また，$\boldsymbol{\beta} > \epsilon$としたことで有効性の低い特徴を除去できたことで，他の選択法より低特徴数である特徴集合を選択することができた．

参考文献

[1] H. Liu and H. Motoda,“Computational Methods of Feature Selection”, Champman & Hall/CRC, 2008

[2] J. Kennedy and R. Eberhart, “Particle swarm optimization”. In IEEE International Conference on Neural Networks,volume 4, pages 1942–1948, 1995.