G-013

相互情報量フィルタリングを用いた遺伝的アルゴリズムによる特徴選択

Feature Selection based on Genetic Algorithm with Mutual Information Filtering

小田原 平† 森 康久仁† 松葉 育雄†
Masaru Odawara Yasukuni Mori Ikuo Matsuba

1 はじめに

特徴選択は，観測される特徴の中から有用な特徴を取り出す方法論である．無関係な特徴を取り除くことで計算量を大幅に減らし，冗長な特徴を削減することで識別システム全体の性能の向上が期待できる．このように，特徴選択は識別システムを構築するうえで非常に重要な前処理である．

本研究では非常に多くの特徴を持つデータを対象とする．そのようなデータの識別は難しく，識別に有効な特徴の組を見つけることも難しい．そこで本研究では，識別に有効な特徴を取り出すために，特徴とクラスの相互情報量を用いて特徴にフィルターをかけ，遺伝的アルゴリズム (以下，GA とする) を利用した方法によって，識別に有効な特徴の組を選択する手法を提案する．提案手法の有効性を確認するために，遺伝子の発現量を測定したマイクロアレイデータを使って実験を行い，ガンに有用な遺伝子を選択することを試みる．

2 フィルタリングによる特徴選択

これまでにも特徴をフィルタリングし，その後に特徴選択を行う手法が提案されている [1]．この手法では，クラスの平均値と全体の平均値を用いてクラス内分散とクラス間分散の比で特徴のランク付けを行う．そして，その比が大きい特徴を取り出した後，GA によって特徴選択を行っている．しかしながら，従来手法で用いる基本的な GA では局所解に陥ってしまうことが多いという問題点があった．

そこで本研究では，GA を並列に行う分散 GA により識別に有効な特徴の組を探索する手法を提案する．特徴のフィルタリングは，変数間の関連性を表すのに非常に有効な相互情報量を用いて識別に有効な特徴をランク付け [2] し，ランクが上位の特徴のみを利用する．

2.1 相互情報量

相互情報量はある 2 つの確率変数が共有する情報量の尺度である．一方の変数を知ることでもう一方の変数をどれだけ推測できるかを示す．

本研究では，特徴とクラスの相互情報量を用いて特徴のランク付けを行う．g_i $(i = 1, 2, ...)$ を i 番目の特徴，$C(= \{c_1, c_2, ...\})$ をクラスの集合とする．このとき，特徴とクラスの相互情報量 $I(g_i; c)$ は式 (1) のようになる．

$$I(g_i; c) = \int\int p(g_i, c) \log \frac{p(g_i, c)}{p(g_i)p(c)} dg_i dc. \qquad (1)$$

本研究では，各特徴は相互情報量の全てのクラスについての和の値 $R(g_i)$:

$$R(g_i) = \sum_{c_k \in C} I(g_i; c_k) \qquad (2)$$

によってランク付けする．

2.2 分散 GA

GA は生物の進化の過程を模倣して，ある問題の最適解を求める手法である．GA では解の候補を「個体」と呼ばれる文字列や数字列で表現する．そして，複数の個体からなる集合を作り，その中から解にふさわしい個体を適応度と呼ばれる尺度により抜き出し (選択)，その後，交叉，突然変異等の操作を行って，新たな別の集合を作成する．GA はこのような繰り返しを，ある一定の基準を満たすまで行い，(準) 最適解を求める手法である．GA の基本的なアルゴリズムは以下のようになる．

Step1: N 個の「個体」を作り，0 世代目の個体の集合 G_0 として，$k \leftarrow 0$ とする．

Step2: G_k に含まれる全ての個体の適応度を計算し，もし終了条件を満たせば終了する．

Step3: 選択，交叉，突然変異の操作により N 個の「個体」を新たに作る．それらの「個体」を $k+1$ 世代目の個体の集合 G_{k+1} とする．

Step4: $k \leftarrow k+1$ として Step2 へ．

基本的な GA を用いて特徴選択を行うと，比較的識別に有効な特徴を選択するものの，局所解に陥いり，不適切な特徴を選択してしまうことも多い．そこで，GA の局所解の問題を解消するために考えられたのが分散 GA である [3]．

分散 GA は，全個体を含む母集団をいくつかの準母集団に分割し，各準母集団内で GA を行う手法である．各準母集団内に存在する個体数は少ないため，従来の GA と同様に局所解に陥りやすいが，移住を行うことで多様性を維持している．移住は，ある準母集団内の個体を別の準母集団内に送る操作で，何世代かに 1 度移住を行う．

2.3 提案手法

本研究では，従来手法の GA の局所解の問題を解消するために，相互情報量と分散 GA を用いた特徴選択アルゴリズムを提案する．提案手法の流れは以下の通りである．

Step1: 相互情報量により，特徴のランク付けを行う．

Step2: ランク上位の特徴を取り出す．

Step3: 取り出した特徴を用いて，分散 GA によって識別に有効な特徴を選択する．

†千葉大学大学院融合科学研究科

2.4 個体の評価

本研究で用いる GA では各個体の表現を選択する特徴の番号列とする．そして，個体の適応度を最近傍法の識別率とすることで，識別に有効な特徴の集合を取り出す．

識別率の計算は Leave One Out Cross Validation 法 (以下，LOOCV とする) を用いる．LOOCV は全サンプル m 個から 1 つのサンプルを取り除き，残りの $m-1$ 個のサンプルで構成した識別規則を用いて，取り除いたサンプルのクラスを予測する方法である．ある特徴集合 G の適応度を決める評価関数 $f(G)$ を (3) で定める．

$$f(G) = (m - E_c) \ / \ m \times 100 \qquad (3)$$

ここで，E_c は LOOCV を用いて予測したクラスと実際のクラスを比べ，誤って予測した個数である．

3 実験

提案手法の有効性を確認するために，実データに対して実験を行った．用いたデータはマイクロアレイによる遺伝子発現データである．マイクロアレイは細胞内の遺伝子の状態を数値化することができる技術であり，同時に，観測される特徴数 (遺伝子数) は数千から数万にもおよぶ．一方で，識別対象は数十と少ないため，識別が難しいとされている．実験では，2 種類のマイクロアレイデータを用いて，提案手法と従来手法 [1] を比較する．

3.1 遺伝子発現データ

本研究では 2 種類の遺伝子発現データ (NCI60[1] と GCM[1]) を用いた．これらはさまざまなガン患者のガン組織の遺伝子の発現量を数値化したデータである．NCI60 は特徴数 6117，サンプル数 62，クラス数 9 のデータであり，GCM は特徴数 16063，サンプル数 198，クラス数 15 のデータである．

3.2 識別結果

NCI60，GCM はともに相互情報量に基づいたランキング上位 500 個の特徴 (遺伝子) から分散 GA を用いてガンの識別に有効な特徴の選択を行った．本研究で用いた分散 GA のパラメータは世代数を 100，個体数は 100 とした．5 つの準母集団に分けて，それぞれの個体数を 20 とした．個体を 10 ～50 個の特徴の番号列で表した．分散 GA で用いる遺伝的操作はトーナメント選択，一様交叉と一様突然変異である．トーナメント選択はランダムに 2 つの個体を選択し，適応度の高い個体を残す．これを繰り返し，もとの個体数になるまで繰り返す選択方法である．一様交叉は，選択操作を行った個体集合からランダムに 2 つの個体を取り出し，個体の各構成要素 (特徴の番号) を 1/2 の確率で交換する．一様突然変異は個体集合のすべての個体に対して，ある確率 (本研究では 0.01) で特徴番号を変更する．移民は 5 世代ごとに，10 個の個体に対して行った．

次に，提案手法と従来手法の識別率の比較を行うために，従来手法のパラメータは提案手法と同じとし，遺伝的操作も同じものを用いた．従来手法では，選択操作が SUS 選択を用いているため，SUS 選択による実験も行った．SUS 選択は適応度の高い個体を優先的に残す選択方法である．表 1，表 2 はそれぞれ，NCI60 と GCM についての結果である．表は従来手法と提案手法を 10 回行ったときの選択特徴数ごとの最大適応度 (識別率) を示した．SUS は SUS 選択を表し，TOU はトーナメント選択を表す．識別結果は，NCI60 において個体の長さ (選択特徴数) を 20 のときに従来手法 (TOU) が最も良く，それ以外では提案手法が最も良い結果となった．

表 1: 選択特徴数ごとの最大適応度 (NCI60)

選択特徴数	10	20	30	40	50
従来手法 (SUS)	71.0	79.0	79.0	77.4	77.4
従来手法 (TOU)	77.4	**83.9**	83.9	87.1	87.1
提案手法	**80.6**	82.3	**87.1**	**88.7**	**90.3**

識別率 (%)

表 2: 選択特徴数ごとの最大適応度 (GCM)

選択特徴数	10	20	30	40	50
従来手法 (SUS)	55.6	60.6	61.1	60.1	59.6
従来手法 (TOU)	61.1	66.7	68.2	67.2	67.7
提案手法	**63.6**	**68.9**	**71.7**	**73.2**	**72.7**

識別率 (%)

4 まとめ

本研究では相互情報量と分散 GA を用いて，識別に有効な特徴を取り出す手法を提案した．相互情報量は特徴ごとに識別に適しているかを数値化することができ，それぞれの特徴を評価することができる．また，分散 GA は解を探索するときに局所解に陥りにくい性質を持つ．識別結果から，提案手法は，NCI60，GCM の両方のデータで従来手法よりも良い結果が得られた．つまり，提案手法によって識別に有効な遺伝子が取り出せたといえる．また，実験では 10 回の分散 GA を行ったが，それぞれの分散 GA で取り出された遺伝子を調べると，NCI60，GCM とも一部の遺伝子が選択されやすい傾向があることがわかった．従って，それらの遺伝子はよりガンの識別に有効な情報を持っていると考えられる．

参考文献

[1] Tsun-Chen Lin et al, " Pattern classification in DNA microarray data of multiple tumor types, " Pattern Recognition, 39, pp.2426-2438, 2006.

[2] Haunchuan Peng et al,"Feature selection based on mutual information," IEEE Transactions on pattern analysis and machine inteligence, vol.27, No.8, 2005.

[3] Weilie Yi et al, " Dynamic Distributed Genetic Algorithms, " Evolutionary Computation 2000, vol.2, pp.1132-1136, 2000.