論文内容の要旨

論文題目:

Cancer Class Prediction and Biomarkers Detection Using Microarray Data with Evolutionary Computation (進化論的計算を用いたマイクロアレイデータの分類とガン関与遺伝子群の検出)

氏名： ポール　トポン　クマル

１．背景：

癌治療においては，正確な癌患者サンプルの分類は非常に重要であると考えられている．しかし，腫瘍の形態，発生，微視的な外見及び位置にもとづく診断は非常に困難である．なぜなら，異なる癌の腫瘍であっても同じ外見である場合や，同じ処置を施しても，異なる反応を示す場合があるためである．さらに，癌細胞の採取には外科手術を伴う場合があり，危険性を有している．遺伝子発現データに対してクラス分類手法を用いることで，従来の病理学的な手法と比較して，客観的，明白かつ一貫した癌の分類手法の研究が近年盛んに行われている．本研究は，遺伝子発現量は多くの外的要因によって影響されるという仮説にもとづいている．ここで外的要因とは，温度，光，種々の信号など，ホルモンの分泌に影響を及ぼすものや，特定の遺伝子の発現量に影響を及ぼすような種々の病気を指す.

通常，癌細胞は通常の細胞の DNA が突然変異することによって生じる．そのため，通常の細胞と癌細胞の発現量を比較することで，癌の病状を起こす遺伝子を特定することができると考えられている．本研究の目的は DNA マイクロアレイの遺伝子発現量データから，バイオマーカーを同定し，正確でロバストな癌分類モデルを構築することにある．このような研究においては，種々の機械学習的手法にもとづく方法が提案されている．しかし，データサンプルの数に比べて，冗長な部分などを含む遺伝子の数が非常に多いため，これらの手法は，限定された状況下においてのみ有効である.

２．手法:

本論文では，二つの手法を提案します：random probabilistic model building genetic algorithm (RPMBGA)， majority voting genetic programming classifier (MVGPC). これら二つの手法は，テストデータにおいて，他の手法と比べて非常に高い精度で癌を分類することが可能である．遺伝的アルゴリズムにもとづく RPMBGA は遺伝子の同定のみを行い，クラス分類器を別に必要とするが，MVGPC はそれ自体が分類及び遺伝子同定を行う.

RPMBGA は, 遺伝的アルゴリズムのような従来の手法と比較して高速である. また, RPMBGA には交叉や突然変異はなく，他の手法と比較してコンパクトな遺伝子セットを同定し，高い精度で分類することが可能である．RPMBGA の初期集団は，多くの遺伝子を選択する状態にある個体によって形成される．RPMBGA は，徐々に個体が選択する無関係な遺伝子を減らし，最終的には少数の遺伝子を選択する個体のみを残す．RPMBGA は一度にひとつ以上の遺伝子を選択するような集団を生成することで，遺伝子間の相互作用を考慮することが出来る．このような方法は，一つの遺伝子の分類精度に基づいて，一度に一つの遺伝子を選ぶランクベースの方法より優れている．なぜなら，筆者らは最も高い精度をもたらす遺伝子のセットは，個々の遺伝子ではそれ以上の分類精度をもたらさないということを発見したためである．さらに，個々の遺伝子及び多くの遺伝子を含む遺伝子のセットでは，完全な分類を行うことは出来ない．多くの遺伝子を含む集合では，無関係な遺伝子が含まれることで分類精度を下げてしまう．RPMBGA は以上の点で他手法と比較して優れているものの，RPMBGA においては同定される遺伝子セット及び分類精度は，適合度の算出に用いる分類器に大きく依存しまうという問題点がある．

MVGPC は遺伝的プログラミング（GP）に多数決手法を導入することで，GP より正確に，さらに RPMBGA より高い信頼性で分類することが可能である．MVGPC は異なる GP のルールを統合することで, テストサンプルの種類の推定を確実かつロバストに行うことが出来る. MVGPC においては，独立した GP の進化において得られた複数のルールをひとつずつテストサンプルに適用し，それぞれのルールは同定した癌の種類に対して投票を行う．テストサンプルの種類は，最も支持数の多かった種類に決定される．MVGPC の基本的なアイディアは，GP によって進化した個々のルールでは，サンプルの種類を正確に推定することは困難であるが，ルールが集団で推定した場合は高い信頼性で推定することができるという考えに基づいている．しかし，多数決手法が有効であるかどうかは，多数決に用いるルールに数（Ensemble size）及び，一つ一つのルールの誤判定率に依存する．Ensemble size が小さい場合や，それぞれのルールの誤判定率が 0.5 以上である場合，MVGPC は個々のルールを単独で適用した場合より低い性能しか示すことが出来ない．そこで，本論文では最も高い性能を示す，最適 Ensemble size を調査する．

本論文ではさらに，バイオマーカーの同定には，まず高い精度の分類器を生成し，その後で，分類器に含まれるルールの中で，高い頻度で出現する遺伝子を選出する方法を提案する．選出される遺伝子の定常的な頻度分布を得るには, マイクロアレイデータに対して複数回 MVGPC を適用する必要がある．この手法は，ある特定の遺伝子はどのような遺伝子選択アルゴリズム及び分類器を用いた場合でも，高い頻度で出現するという点にもとづいている．高い頻度で選択される遺伝子は，癌のバイオマーカーである場合と，生物学的には無関係であるが，トレーニング及びテストサンプルと非常に相関のある遺伝子である場合がある．

本論文の主要な提案は以下の点である：

・最適な Ensemble size の決定手法.
・多数決を用いたテストサンプルの種類の同定.
・マイクロアレイデータのバイオマーカーの抽出.

３．結果：

本論文では，Affymetrix の GeneChip ソフトが生成する遺伝子発現データを用い，二分類及び多分類の分類を行った．上記のマイクロアレイデータに対して RPMBGA 及び MVGPC を適用することで，他の手法と比較して高い精度で分類することに成功した．MVGPC は RPMBGA より正確に分類することが可能である．MVGPC におけるテストデータの正確度は，AdaBoost と GP の統合手法を含む他の手法と比較して，非常に高い結果を示す．さらに，MVGPC によって選択された遺伝子のうちいくつかは，本論文で扱った癌と関係があることが知られている．

さらに，MVGPC をマイクロアレイ以外のデータに適用し，MVGPC による正確度は，GP で獲得した単独のルール及び複数のルールで単純に同定を行った場合より高い結果を示した．

４．結論：

MVGPC は遺伝子発現量に基づく癌診断，そして癌のバイオマーカーの同定を行うのに，正確かつロバストな計算手法であると考えられる．AdaBoost は，弱学習器を統合することで，推定精度を改善する手法であるが，遺伝子発現データの分類においては，MVGPC が AdaBoost と GP の統合手法を上回る性能を発揮することが分かった．このような結果となった理由は，AdaBoost によって GP で獲得されたルールは全てのテストサンプルを用いない場合があるのに対して，MVGPC によって獲得されたルールは全てのテストサンプルを用いるためである．

しかし，MVGPC が有効であるかは，個々のルールの性能に依存し，MVGPC によって扱われる遺伝子の数は非常に多いものとなる．さらに，MVGPC の実行時間は，大きな多分類のマイクロアレイデータの場合，他の手法と比較して，非常に長くなるという問題点がある．これらの点は今後の課題であると考えられる．