

サンプル混合度を考慮した遺伝子発現量のがんバイオマーカー探索

Biomarker discovery from gene expression data of mixed tumor samples

村上 勝彦^{*1}
Katsuhiko Murakami

^{*1} 富士通研究所
FUJITSU LABORATORIES LTD.

To diagnose the state of tissues from cancer patients, utilization and discovery of biomarker with high discrimination accuracy is important. It is difficult to determine biomarkers that can cleanly separate cancer and normal tissues when normal tissues are mixed at the time of collecting a cancer sample. The purpose of the study is to provide method to discover biomarkers or characteristics of tissues with high accuracy for distinguishing between cancer tissues and non-cancer tissues. By utilizing a parameter λ , which indicates purity and status of each sample, highly accurate biomarker discovery became possible.

1. はじめに

細胞の状態を診断するため、判別精度が高いバイオマーカー探索方法は重要である。バイオマーカーとは、一般には病気の変化や治療に対する反応に相関し、指標となるものである。より具体的には、血液や尿などの体液や組織に含まれる、タンパク質や遺伝子などの生体内物質の存在、またはその量がバイオマーカーとなる。

本稿で扱う問題は、がん細胞と非がん細胞を判別するためのバイオマーカー（がん細胞で高発現する遺伝子セット）探索である。良い特徴量を選択できれば、新規の患者候補のサンプルが、癌であるかを精度良く判定できる。高精度の判定ができれば、治療の前後でサンプルを判定することで、治療がすすんでいるか、転移していないかを判定することにも用いられる。

バイオマーカー利用の際には、未知サンプルのバイオマーカー値を計測してサンプルの医学・生物学的な状態を推定する。探索時に構築した予測モデルを用いてバイオマーカー値を算定し、未知サンプルが、がんであるかの判定する。判定結果は、採取したサンプルが癌であるかの判定、癌種類や進行度の特定、臨床における治療効果の判定等に利用される。判別モデルは、回帰分析や LinearSVM [Abeel 2009] など通常の機械学習で学習、判別する。

実際のデータでは、がんサンプルを採取する際に正常細胞が混じったり、がん細胞内で特徴が大きくばらつく場合があり、がんと正常細胞をきれいに分離できるバイオマーカーを決定することが困難になることがある。そこでサンプル内の正常細胞とがん細胞の割合をパラメータとして統計モデルを作り ML (最尤法) で求める方法がある [Shen 2015]。しかしこれらは、「癌」と「正常」の2クラス分類で、学習や判定の際に、クラス内サンプルを等価に扱っている。「癌」も「正常」が均一であると仮定してこれら多様性が考慮されてない。また、癌には同種のがん(例えば「大腸がん」)でも様々なタイプや進行度がある。現在は同じ分類でもサブクラス分類が有効かもしれないが解析には考慮されていない。遺伝子発現量も安定な線形を仮定し、遺伝子の多様性を扱っていない。サンプルや遺伝子によって、外れ値のようにモデルにあわない部分がある。これらをノイズと扱うべき場合でも、等価に学習してしまうため、正しいパラメータを推定できなくなる。

本研究の目的は、提案した手法により混合度がどのくらい推定可能かを評価することである。

2. データと方法

2.1 データセット

バイオマーカー探索の際には、癌サンプルと、(ネガティブ)コントロールとしてがんが発生している組織(例: 肺, 肝臓)の正常サンプルを採取する。組織サンプルごとに遺伝子 j の発現量を計測する。サンプル i に対して、測定値 j (遺伝子 j の発現量) x_{ij} とすると、データは行列形式に書ける。サンプルは大きく2つに分かれる。「がんサンプル」とラベルされたものと、「正常サンプル」である。しかしこれらはすべてが正しいものではない、という前提が本研究で重要なポイントである。解析の過程ではこれらのラベルは訂正される段階を経て、全体が解析される。

評価用のデータとして、NCBI GEO 遺伝子発現データベース (<http://www.ncbi.nlm.nih.gov/geo/>) から、アクセッション番号 GSE16515 として公開されている膵臓癌のデータを用いた。これには正常サンプル 16 個、がんサンプル 36 個、合計 52 サンプルのデータである。この実データをもとにして混合率の異なるデータセットを作成した。それらが以下に述べる手法で同定可能かどうかを検討した。

2.2 混合度の推定とバイオマーカー探索方法

「癌細胞として採取したサンプル内に混在する正常細胞の割合」を、隠れた変数として推定できるように、正常細胞の混在率 (0 から 1) λ (正常度と呼ぶことにする) をモデルに組み込む。 λ は正常細胞とみなせる程度のパラメータとも解釈できる。

また、癌細胞はいくつかのステージで進行するものや分岐するなどの例が知られている。現在同じラベル (例「大腸がん」) といってもサブクラスで状態が異なる可能性がある。異なる状態には異なるバイオマーカーで状態を判定するのが適当である。 λ は細胞の正常度・非進行度とも解釈できる。

この状態をとらえる隠れ変数を導入して、データから細胞の状態を推定できるようにする。これは教師ラベル (癌, 正常) と高い相関があるが異なるものである。

正常度パラメータ λ が1に近いサンプル (第1グループ) は正常細胞とみなせる。一方、正常度パラメータ λ が0に近いサンプル (第2グループ) は癌細胞とみなせる。また λ が 0.5 に近いサンプル (第3グループ) は混合したサンプルか、ずれた正常細胞

連絡先: 村上勝彦, 富士通研究所, murakami.ktk@gmail.com

か、初期の癌細胞とみなせる。第3グループのサンプルは、判別の教師データとしてはのぞましくない。これを除いた(第1と第2グループからなる)学習データを作成し、再度判別モデルを構築すると、精度高い判別モデルが作成できる。同時に癌細胞の特徴をよく捉える特徴量のセット(バイオマーカー)が決定できる。

詳細には、以下の手順で行う(図1)。発現量など特徴量の変数を規格化する。SVDにより次元削減(約54,613次元から50次元程度へ)をする。5万という数はプローブの数であり、遺伝子数の数倍になっている。多くの場合はサンプル数が100個程度なので、それが次元の上限となる。

次にサンプルのクラスタリングを行う。例はガウス混合分布だが、k平均法など他の方法でもよい。

クラスター数 $n(>2)$ を決め、 n の数だけのガウス分布の重なりで表現する「混合ガウス分布(GMM)」のパラメータを求める。

ガウス分布のパラメータは、クラスター l に対して、平均 μ_l 、分散 σ_l 、中心位置ベクトル $\vec{c}_l$ ($l=1,2,\dots,n$) である。

各サンプル (x_i) の生成確率分布は、

$$P(\vec{x}_i) = \sum_{l=1}^n \pi_l C_l e^{-\frac{(\vec{x}_i - \vec{c}_l)^2}{\sigma_l}}$$

とする。ここで π_l はクラスター l の推定比率、 C_l は規格化定数。この全サンプルで和をとり、最大化するように求める。

常細胞サンプルだけから平均ベクトル $\vec{x}_v$ を求めておき、 $\vec{x}_v$ を含むクラスターと含まないクラスターに分ける。

$\vec{x}_v$ に近いクラスター中心を持つクラスターから順に、属するメンバーのクラスターを結合していき、教師データラベルが正常細胞のサンプルのうち、9割以上を含むクラスターが見つかったところで結合を停止する。その時点で結合されたクラスター群 n_n 個を「正常クラスター」、それ以外 $n_a (= n - n_n)$ 個を「異常クラスター」とみなす。よって各サンプル i 毎に

$$\lambda_i = \sum_{l \in N} \pi_{il}$$

が計算できる。正常度 λ_i が両極値を含まない領域(例えば0.2から0.8)のサンプルを、混在サンプルグループ M に属するとみなす。はじめのデータセットのサンプルのうちからサンプル群 M を除き、混合度の低い新規学習データ D_{new} を作成する。 D_{new} を用いてロジスティック回帰などのなんらかの判別器を作成する。結果、判別モデルと、バイオマーカーが得られる。

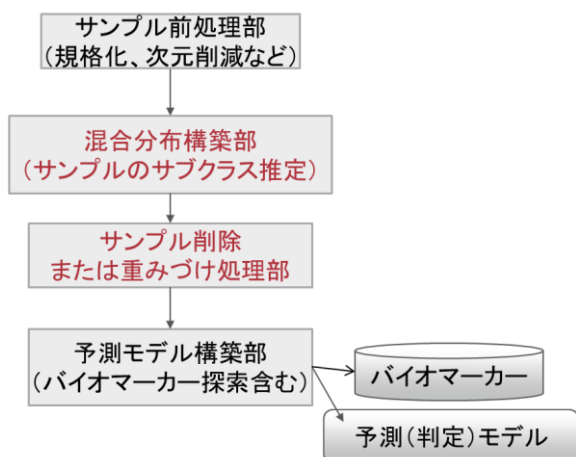


図1 解析方法のフローチャート。赤字が特徴的部分。

3. 結果

混合率の異なるデータセットごとに本手法を適用した結果、人工データにおいて混合サンプルを作成し、混合サンプルの70%を同定でき、それに基づくバイオマーカーを得た。

4. 議論

4.1 パラメータの意義

癌細胞と正常細胞の実サンプルには、それらの混在サンプルがあり得る。サンプルがマッピングされる特徴量空間で、2サンプル群の中間に位置する細胞は混在サンプルであることが想定される。これらのサンプルは、本来離れた位置にあるはずの2種のサンプルが混在した疑似サンプルであり、細胞としては実在しない。これらを、中間的な特徴量をもつサンプルとして学習するのは正しくない。これは生体サンプルでの特別な状況である。本方法ではパラメータ λ を導入して混在状況を把握できる。

混在サンプルが多い場合、通常のモデル構築では、精度の悪い結果を得るだけである。しかし本発明では、もし λ が中間値をとるようなサンプルが多い結果となれば、混合サンプルが多いという情報が得られる。

混合サンプルがあると、正しいモデル構築が非常に困難になる。ここでは混合サンプルをほぼ無視したきれいな学習データによるモデル構築ができる。その結果、高精度で安定的なバイオマーカー探索が行える。

5. おわりに

がん細胞を採取する際、目視で正常細胞と見分けることがしばしば困難であり、採取したサンプルに正常細胞が混在することがある。正常サンプルについても、その外れ値や異常性は考慮できなかった。混合サンプルがあるとそれらがノイズとなり特徴量を正しく学習できない。

本提案では、教師付学習の正解ラベル第1値(正常サンプル)の学習データ数の情報に基づいて、複数クラスを統合した統合クラスを生成することで、正常サンプルのラベル修正を可能とした。その結果、70%の混合サンプルを同定できた。結果、判別に有効なバイオマーカーを得た。今後、モデルを改良したり、さまざまなサンプルについて実験を重ねていく予定である。

参考文献

- [Abeel 2009] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods," *Bioinformatics*, 2009.
- [Emmert 1996] M. R. Emmert-Buck, et al. "Laser Capture Microdissection", *Science*, 1996.
- [Zhao 2018] L. Zhao, V. H. F. Lee, M. K. Ng, H. Yan, and M. F. Bijlsma, "Molecular subtyping of cancer: current status and moving toward clinical applications," *Brief. Bioinform.*, 2018.
- [Ahn 2013] J. Ahn et al., "DeMix: Deconvolution for mixed cancer transcriptomes using raw measured data," *Bioinformatics*, 2013.
- [Shen 2016] Q. Shen, J. Hu, N. Jiang, X. Hu, Z. Luo, and H. Zhang, "contamDE: differential expression analysis of RNA-seq data for contaminated tumor samples," *Bioinformatics*, 2016.